Catalogue 12-001

Survey Methodology

A Journal of Statistics Canada

June 1992 Volume 18 Number 1

Statistics Canada Statistique Canada

Canada

Statistics Canada
Social Survey Methods Division

Survey Methodology

A Journal of Statistics Canada

June 1992 Volume 18 Number 1

Published by authority of the Minister
responsible for Statistics Canada

June 1992

Price: Canada: $35.00
United States: US$42.00
Other Countries: US$49.00

Catalogue No. 12-001

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is $35 per year in Canada, US $42 in the United States, and US $49 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 18, Number 1, June 1992

CONTENTS

In This Issue 1

Census Undercount Measurement Methods and Issues

D.A. FREEDMAN and W.C. NAVIDI
Should We Have Adjusted the U.S. Census of 1980? 3
Comment: S.E. FIENBERG 25
I.P. FELLEGI 29
N. CRESSIE 32
A.L. SCHIRM and S.H. PRESTON 35
J.A. HARTIGAN 44
T.P. SPEED 51
E.P. ERIKSEN and J.B. KADANE 52
Response from the Authors 59

N. CRESSIE
REML Estimation in Empirical Bayes Smoothing of Census Undercount 75

G.S. DATTA, M. GHOSH, E.T. HUANG, C.T. ISAKI, L.K. SCHULTZ and J.H. TSAY
Hierarchial and Empirical Bayes Methods for Adjustment of Census Undercount: The 1988 Missouri Dress Rehersal Data 95

D. ROYCE
A Comparison of Some Estimators of a Set of Population Totals 109

L. SWAIN, J.D. DREW, B. LAFRANCE and K. LANCE
The Creation of a Residential Address Register for Coverage Improvement in the 1991 Canadian Census 127

S.E. FIENBERG
Bibliography on Capture-Recapture Modelling With Application to Census Undercount Adjustment 143

L.K. ROE, J.F. CARLSON and D.A. SWANSON
A Variation of the Housing Unit Method for Estimating the Population of Small, Rural Areas: A Case Study of the Local Expert Procedure 155

Z. XIA, P.S. LEVY, E.S.H. YU, Z. WANG and M. ZHANG
Single Stage Cluster Sampling in Prevalence-Incidence Surveys: Some Issues Suggested by the Shanghai Survey of Alzheimer's Disease and Dementia 165

Survey Methodology, June 1992
Vol. 18, No. 1, pp. 1–2
Statistics Canada

In This Issue

Despite the best efforts by statistical agencies in counting people in a census, a small undercount always remains. These undercounts are usually not uniformly distributed over various subgroups of the population and therefore they impact differently on various government programs that use census population figures. Consequently, methods of measuring undercounts, adjustment techniques, especially for local areas, and related issues have attracted a great deal of attention from policy makers and statisticians. The six articles included in the special section on **Census Undercount Measurement Methods and Issues** will be a valuable addition to the growing literature on this topic.

The first article in the section is a discussion paper by Freedman and Navidi. It reviews some of the statistical issues and arguments for and against adjusting the United States Census of 1980 as well as discusses statistical evidence presented in a trial against the Department of Commerce and the U.S. Bureau of the Census. The article is a continuation of the discussion between the authors and Ericksen, Kadane and Tukey who are proposing methodology for the adjustment. It also shows how some of the conflicting views were resolved by the trial court. The article is followed by very insightful and lively comments from several statisticians and a reply from the authors.

Cressie presents an empirical Bayes approach to prediction of undercount at subnational levels based on restricted maximum likelihood (REML). The claimed advantage of the REML estimators is that they do not tend to oversmooth the post-enumeration survey data as maximum likelihood estimation does. The REML estimators are compared with the maximum likelihood and method-of-moments estimators by simulation and example.

Prior to the 1990 U.S. Census, a dress rehearsal took place in the state of Missouri. Datta *et al.* use the data from this exercise to study procedures for modelling from census post-enumeration surveys. They consider both hierarchical and empirical Bayes approaches. The results indicate that both approaches lead to improvements on the dual system estimation approach. The authors conclude with an update in light of the adjustment of the actual 1990 U.S. Census.

Four estimators of the base population used as a benchmark in the Population Estimates Program of Statistics Canada are discussed by Royce. These are the unadjusted census counts, adjusted census counts, a preliminary test estimator and a composite estimator. The Weighted Mean Square Error is used as the basis for comparison of these estimators, not only for estimation of population totals but also for estimation of functions of population totals, such as population shares or growth rates *etc.*

Swain *et al.* give an overview of the Address Register that was created at Statistics Canada as a means of reducing undercoverage in the 1991 Census of Canada and represents a frame of the residential addresses for medium and large urban centres. Methodology, post-censal evaluation and future prospects are discussed.

The final article of the special section, by Fienberg, presents a selected annotated bibliography of the literature on capture-recapture estimation of population size. Capture-recapture estimation is the main method used to evaluate the completeness of the census counts and thus the article concentrates on literature related to the estimation of human populations.

Roe, Carlson and Swanson describe a variation of the Housing Unit Method to estimate the population of small rural areas. In this variation, local experts provide data about selected households. The estimates are compared to the census counts for three rural communities.

Xia *et al.* compare the statistical properties and costs of telescopic single stage cluster sampling with that of ordinary single stage cluster sampling. Telescopic single stage cluster sampling is an alternative when sub-sampling of clusters (*i.e.* two-stage cluster sampling) is not possible. The method has been used in the Shangai Survey of Alzheimer's Disease and Dementia, which serves as an illustration of how costs can be reduced without sacrificing precision.

The Editor

Survey Methodology, June 1992
Vol. 18, No. 1, pp. 3–74
Statistics Canada

Should We Have Adjusted the U.S. Census of 1980?

D.A. FREEDMAN and W.C. NAVIDI[1]

ABSTRACT

This paper reviews some of the arguments for and against adjusting the U.S. census of 1980, and the decision of the court.

KEY WORDS: Census; Adjustment; Post Enumeration Survey; Regression; Smoothing.

1. INTRODUCTION

Every ten years, the census gives a statistical portrait of the United States. Geographical detail makes these data unique. However, the counts have more than academic interest: they influence the distribution of power and money. The census is used to apportion Congress as well as local legislatures and to allocate tax money – $40 billion per year in the late 1980s – to 39,000 state and local governments. For these purposes, the geographical distribution of the population matters, rather than counts for the nation as a whole. Indeed, the census is used as a basis for sharing out fixed resources: if one jurisdiction gets more, another must receive less. Adjusting the census is advisable only if the process brings us closer to a true picture of the distribution of the population.

A small undercount is thought to remain in the census, and this undercount is unlikely to be uniform. People who move at census time are hard to count; in rural areas, maps and address lists are incomplete. Central cities have heavy concentrations of poor and minority persons, who may be harder to enumerate. If the undercount can be estimated with good accuracy, especially at the local level, adjustments can – and should – be made to improve the census. Some statisticians argue that the undercount can be estimated well enough, others are skeptical: a bad adjustment may be worse than nothing.

Because of its resource implications, the undercount has attracted considerable attention in the media, the Congress, and the courts. After the 1980 census, New York City joined with other jurisdictions to sue the Department of Commerce, seeking to compel an adjustment based on demographic analysis and capture-recapture techniques. The Commerce Department resisted this pressure. The trial court framed the issue as follows:

> "The plaintiffs contend that a statistical adjustment of the census will improve upon the accuracy of the census, thereby reducing the disproportionate undercount in the City and State [of New York]. The Census Bureau, however, contends that although the census counts are imperfect, a statistical adjustment of the census will inject even greater inaccuracies into the population count, and that therefore, a statistical adjustment of the census is not technically feasible or warranted at this time." (674 F Supp 1091 = volume 674 of the Federal Supplement, page 1091).

[1] D.A. Freedman, Statistics Department, University of California, Berkeley, CA U.S.A. 94720; W.C. Navidi, Mathematics Department, University of Southern California, Los Angeles, CA U.S.A. 90089.

The 1980 case may seem dated, given that the census of 1990 has already been taken. However, among law suits that involve statistical principles, the 1980 census case was one of the most important and closely argued; there is still much to learn from it. This article will review some of the technical issues, and some of the findings of the court.

The balance of this section will sketch the background; for more details, see Cohen and Citro (1985) or Fay *et al.* (1988). There are two methods for evaluating the completeness of the counts in the U.S. Census: demographic analysis and capture-recapture. Demographic analysis uses administrative records (birth certificates, death certificates, immigration visas, *etc.*) to make independent estimates of population totals. The starting point is an accounting identity:

$$\text{Population} = \text{Births} - \text{Deaths} + \text{Immigration} - \text{Emigration}.$$

Demographic analysis provides estimates by age, sex and race but not ethnicity, because of gaps in the records. Data on immigration and emigration are incomplete; birth records are incomplete too, especially prior to 1935. Thus, the data going into the "identity" must be supplemented by a variety of imputations and adjustments. Furthermore, data on internal migration are lacking, so estimates are made primarily at the national level. This completes our sketch of demographic analysis.

Estimates of coverage for small areas (including states and cities) are based on capture-recapture techniques. Capture is in the census; recapture is in a sample survey conducted after the census. In 1980, there were two such surveys, or "*P*-samples:" the April and August CPS (Current Population Survey). Each record from the *P*-samples was matched against the census file to see if the corresponding person was "captured," that is, counted in the census. Records that could not be matched indicated people who were missed by the census - or a failure in the matching process. These data were used to estimate the percentage of persons missed by the census, that is, the rate of omissions.

The census also had a small percentage of erroneous enumerations (for instance, people counted at two different addresses); the number was estimated by taking an "*E*-sample" of census records and trying to check them by field work. In effect, the net undercount was estimated by taking the difference between the omissions and erroneous enumerations. (For details, see Fay *et al.*, Chapter 5.) These undercount estimates were made as part of "PEP," the Post Enumeration Program.

In 1980, there was a fair amount of missing data in the *P*- and *E*-samples: for instance, there was a 4% non-interview rate in the CPS; even after interview, a determination of match status could not be made for another 4% of the subjects. To see the effect of missing data, a variety of imputation schemes were considered, leading to 12 different series of PEP estimates for 66 subareas.

The 66 areas covered the whole U.S. They included cities like New York; states apart from these cities, like upstate New York; and whole states like Wyoming. A PEP "series" consists of 66 estimates, one for each study area; 9 of the 12 series were based on the April CPS, and 3 on the August CPS.

In the 1980 case, expert witnesses for plaintiffs included Gene Ericksen, Jay Kadane, and John Tukey. Their strategy for adjusting the census using PEP data was described in Ericksen and Kadane (1985). Freedman (among other statisticians and demographers) testified for the defendants, and Navidi was a consultant. A critique of the proposed adjustments was summarized in Freedman and Navidi (1986), to be referenced here as FN.

We now indicate some of the technical issues. According to experts from the Bureau of the Census:

(a) There were substantial differences among the 12 PEP series, demonstrating that missing data were a serious problem.
(b) The PEP estimates were subject to large biases, apart from the problems created by missing data.
(c) Each PEP series was subject to unacceptably large sampling error.

Ericksen and Kadane responded that one of the PEP series ("PEP 2-9") was preferred, and that sampling error could be substantially reduced by regression modeling. They proposed a model with two equations. The first equation expresses the idea that y_i, the PEP estimate for study area i, is an unbiased estimate of the true undercount γ_i for that study area. Informally,

PEP estimate for area i = True undercount in area i + Random error.

Formally,

$$y_i = \gamma_i + \delta_i. \tag{1}$$

The second equation expresses a theory about the variation of the undercounts from area to area, in terms of a vector of explanatory variables X_i and a vector of hyper-parameters β. Informally,

True undercount in area i = Linear combination of explanatory variables for area i + Random error.

Formally,

$$\gamma_i = X_i \cdot \beta + \epsilon_i. \tag{2}$$

The assumptions on the error terms can be stated as follows:

$$E(\delta_i) = E(\epsilon_i) = 0. \tag{3}$$

$$\mathrm{var}\delta_i = K_i,\ \mathrm{var}\epsilon_i = \sigma^2. \tag{4}$$

$$\delta_1, \delta_2, \ldots, \delta_{66}, \epsilon_1, \epsilon_2, \ldots, \epsilon_{66} \quad \text{are independent.} \tag{5}$$

$$\delta_i \quad \text{and} \quad \epsilon_i \quad \text{are normally distributed.} \tag{6}$$

In (4), K_i is the split-sample variance for y_i computed by the Bureau; randomness in K_i is ignored; σ^2 does not depend on i and is treated as constant even though it is estimated from the data. The role of assumptions, and departures from them, was examined in FN; also see the discussion papers and rejoinder, as well as sections 6-7 below.

The Ericksen-Kadane model was used in the 1980 case to smooth the PEP estimates, with the objective of reducing sampling error. The main focus of FN was a critique of that model. Ericksen, Kadane and Tukey (1989) – to be referenced here as EKT – replied to FN, and the present paper continues the exchange.

EKT cited a paper by Schirm and Preston (1987), which considers adjusting states and the District of Columbia by the "synthetic method." For instance, demographic analysis (with one set of assumptions on illegal immigration) estimated a national undercount rate of 5.9%

for blacks and 0.7% for whites in 1980. The synthetic method adjusts each state as follows: increase the number of blacks by 5.9% and the number of whites by 0.7%. In short, undercount rates are assumed to depend on race but not geographical area - or anything else.

This completes our summary of the technical background. For an update on the 1990 census, see Freedman (1991); some of the introductory material here was excerpted with minor changes from that paper. For other views, see Hogan and Wolter (1988), Schirm (1991), Wolter (1991), Wolter and Causey (1991), or Ericksen, Estrada, Tukey and Wolter (1991). The balance of the present paper responds to the salient points raised by EKT, and indicates how some of the the conflicting views were resolved by the trial court.

2. DO THE ADJUSTMENTS IMPROVE ON THE CENSUS?

The most important question is whether adjustments improve on the census counts. EKT "... are confident of improving upon the raw census count (p. 943)"; indeed, there are

> "two simple [synthetic] adjustments that improve upon the census ... the question of the Ericksen and Kadane model is not whether it proves that adjustment is feasible, but whether it improves upon the simpler methods (pp. 927-8) ... Study of the method will not "prove" that an adjustment will improve the census. This has already been demonstrated by Schirm and Preston and the results of Tables 5 and 6 (p. 933)."

Thus, EKT's Tables 5 and 6 are the main pieces of empirical evidence to show that adjustment will improve on the census. And Table 6 on erroneous enumerations is redundant, because the PEP estimates in Table 5 include the effect of erroneous enumerations. Table 5 is the critical one, and it is reproduced here for ease of reference. In our opinion, the table says very little about the possibility of improving on the census; to see why, some numerical detail is needed. (Schirm and Preston will be discussed in the next section.)

"Group 1" in the table consists of 16 central cities; "group 2" consists of other study areas that have relatively high minority populations; "group 3" consists of study areas with small minority populations. At best, the table shows that several methods for adjusting these groups are in general agreement. The table does not show that any of the methods improve on the accuracy of the census. It cannot, because there is no external standard against which to measure improvement.

Moreover, we believe the impression of agreement in the table to be largely illusory. There are dramatic differences among EKT's preferred PEP series, or between these series and the synthetic adjustment of Schirm and Preston. Of course, drama depends on scale, and our next task is choosing units. Proponents of adjustment often use "loss functions" to make their argument; squared error is a common choice: see Ericksen, Estrada, Tukey and Wolter (1991, p. 20). EKT view Schirm and Preston as demonstrating census adjustment to be advantageous, so we compute the root mean square difference between the census and the "Synthetic B" line in Table 1, which is based on the Schirm and Preston adjustment. (The mean is weighted by population shares.)

$$\sqrt{.11 \times (.12)^2 + .44 \times (.06)^2 + .45 \times (.18)^2} \approx 0.13 \text{ of } 1\%.$$

In short,

$$\text{rms difference between census and synthetic } B = 0.13 \text{ of } 1\%. \qquad (7)$$

Table 1

EKT's Table 5. Changes in National Population Shares Resulting When Counts are Adjusted by Sample Estimates Pooled Across Areas and Synthetic Estimates. [The entries for the three groups represent changes in shares, or differential undercounts; the entries in the last column represent total undercounts.]

PEP estimate	Group 1	Group 2	Group 3	Estimated national undercount rate
2-20	+.52%	+.09%	−.61%	+1.9%
3-20	+.51%	+.08%	−.59%	+1.7%
2-9	+.50%	+.06%	−.56%	+1.6%
3-9	+.49%	+.04%	−.53%	+1.4%
2-8	+.41%	+.04%	−.45%	+1.1%
3-8	+.39%	+.03%	−.42%	+1.0%
5-9	+.31%	+.25%	−.56%	+2.1%
5-8	+.22%	+.23%	−.45%	+1.7%
14-20	+.21%	+.02%	−.23%	−.2%
10-8	+.19%	+.07%	−.26%	+.3%
14-9	+.19%	−.01%	−.18%	−.5%
14-8	+.10%	−.03%	−.07%	−1.0%
Synthetic A	+.17%	+.14%	−.31%	+1.4%
Synthetic B	+.12%	+.06%	−.18%	+1.4%
Shares of Census Count	10.76%	44.24%	45.00%	

Notes: (i) Group 1 includes 16 central cities. Group 2 includes three state remainders (California, Maryland, and Texas, excluding Group 1 cities) and 17 whole states. All areas are at least 10% Black or Hispanic. Group 3 includes nine state remainders and 21 whole states. All Group 3 areas are less than 10% Black or Hispanic.

(ii) The Synthetic A estimates assume that (a) Blacks have the same undercount rates as Hispanics, 5.9%; (b) the undercount rate of persons neither Black nor Hispanic is 0.3%; (c) the undercount rates for Blacks, Hispanics, and all others are invariant across geographic areas; and (d) there are 3 million undocumented aliens, 9.6% of whom are Black.

(iii) Following Schirm and Preston (1987), the Synthetic B estimates assume that (a) the Black undercount rate is 5.9%; (b) Hispanics and other non-Blacks have an undercount rate of .7%; (c) the undercount rates for Blacks, Hispanics, and all others are invariant across geographic areas; and (d) there are 3 million undocumented aliens, 9.6% of whom are Black.

EKT prefer the first 8 of the PEP series (pp. 933 and 938). We next compute the rms difference between PEP 2-20 and 3-8, which are among EKT's preferred series. (PEP 2-20 and 3-8 were both based on the April CPS; differences between them are due only to procedures for handling missing data.)

$$\text{rms difference between PEP 2-20 and 3-8} = 0.14 \text{ of } 1\%. \tag{8}$$

EKT also recommend averaging as a way of eliminating indeterminacies (pp. 931 and 937). Table 2 compares population shares from the census, the synthetic B estimates, and the average preferred PEP estimates. We take the rms difference between the average preferred PEP and synthetic B:

$$\text{rms difference between average preferred PEP and synthetic B} = 0.25 \text{ of } 1\%. \tag{9}$$

Table 2

Population Shares from the Census, the Synthetic B Estimates, and the Average of EKT's Eight Preferred PEP Series (2-20, 3-20, 2-9, 3-9, 2-8, 3-8, 5-9, 5-8).

	Group 1	Group 2	Group 3	Total
Average Preferred PEP - Synthetic B	.30%	.40%	−.34%	.00%
Census - Synthetic B	−.12%	−.06%	+.18%	.00%
Average Preferred PEP	11.18%	44.34%	44.48%	100.00%
Synthetic B	10.88%	44.30%	44.82%	100.00%
Census	10.76%	44.24%	45.00%	100.00%

A comparison of (7), (8) and (9) reveals three salient points:

(a) the difference between the census and synthetic B is rather small;
(b) the range in the preferred PEP series is larger than the difference between the census and synthetic B;
(c) the difference between the average preferred PEP and synthetic B is twice the difference between the census and synthetic B.

EKT must view a difference of 0.13% as serious: see (7). On this scale, the PEP series do not agree among themselves. Furthermore, the PEP series are very different from the synthetic adjustment. Of course, the reason may be that Schirm and Preston did not go far enough. However, a National Academy of Sciences review panel – with Jay Kadane as a prominent member – reached the tentative conclusion that Schirm and Preston already over-adjusted the census: see Cohen and Citro (1985, p. 287).

The PEP estimates are in better agreement with the "synthetic A" adjustment in Table 1. But this is circular: the undercount rate for hispanics in synthetic A was estimated from PEP, while synthetic B was based on demographic analysis. Differences among the PEP estimates are an awkward reality; and so are differences between the PEP estimates and synthetic adjustments.

We now quote the principal claim made by EKT (p. 927):

> "Our conclusion is that regardless of whether we use one of the simple methods or the composite method and regardless of how we vary the assumptions of the composite method, an adjustment reliably reduces population shares in states with few minorities and increases the shares of large cities."

Giving more money to cities by changing the census counts is a good idea only if the adjustment reliably improves the accuracy of the census. Accuracy is the crucial issue, and we wish EKT would address it more directly. Their Table 5 is almost irrelevant.

3. SCHIRM AND PRESTON

Can synthetic adjustment reliably improve on the accuracy of the census? EKT think so, citing Schirm and Preston (1987) for the evidence. Schirm and Preston present two major arguments, one analytic and one based on simulation. However, both have serious flaws.

Table 3

A Counter-example to the Analytical Argument.
There are Two States and Two Races.

	White		Black		Total	
	Census count	True count	Census count	True count	Census count	True count
State A	90	89	1	2	91	91
State B	910	890	99	119	1,009	1,009
Total	1,000	979	100	121	1,100	1,100

The analytic argument (p. 966):

"Our finding is that synthetic adjustment will always move the estimated ratio of a state's population to the national population closer to the true ratio if:

(a) the state's black undercount is closer to the national black undercount than it is to the national undercount for both races combined and

(b) the state's white undercount is closer to the national white undercount than it is to the national undercount for both races combined."

As a matter of mathematics, this proposition is wrong. A counter-example is given in Table 3: state A, for instance, has by construction 89 whites and a census count of 90.

The counter-example has been set up to make the arithmetic easy; more complicated and realistic examples could undoubtedly be provided. In Table 3, the overall error in the census (white plus black) is 0, for each state and for the nation. Thus, the census gets the state shares right, and any adjustment will make matters worse. Error rates (with the true population as base) are shown in Table 4: Schirm and Preston's conditions are satisfied. Synthetic adjustment moves both states farther from truth, as shown in Table 5; state B is helped, state A is hurt. To compute Table 5 from Table 3, the number of whites in state A is multiplied by:

$$\text{true national total for whites/national census total} = 979/1{,}000. \tag{10}$$

The arithmetic for the other cells is similar.

The counter-example may be informative, as a parable: state A is sparsely populated, with a small minority population; state B is heavily populated, and has a large, hard-to-count minority population. Synthetic adjustment may favor states of type B at the expense of type A. The mathematical error in Schirm and Preston's appendix appears to be in their reasoning from display A.2. Professor Preston informs us (personal communication) that the theorem holds, with a more complicated set of conditions involving weighted averages.

Table 4

Undercounts from Table 3, in Percent.
(Negative undercounts correspond to overcounts.)

	White	Black	Total
State A	−1.1%	50%	0%
State B	−2.2%	17%	0%
Total	−2.1%	17%	0%

Table 5
The Synthetic Adjustment, "Syn".

	White		Black		Total	
	Syn	True count	Syn	True count	Syn	True count
State A	88	89	1	2	89	91
State B	891	890	120	119	1,011	1,009
Total	979	979	121	121	1,100	1,100

This completes our discussion of the analytic reasoning in Schirm and Preston. What about the simulation results? Basically, Schirm and Preston consider 51 areas (the states and D.C.) and two races (black and white). They set up a joint distribution for an assumed "true" population and the census counts; both are taken as stochastic. The census counts can be adjusted by the synthetic method, and the question is whether the raw counts or the adjusted counts are closer to the assumed true counts. Schirm and Preston actually consider several joint distributions, defined by different "scenarios," that is, choices of parameters; the results are quite similar across scenarios. They also consider several loss functions, or measures of closeness.

We focus on Scenario I, and make two brief comments.

(a) The claimed improvement is rather modest. For example, on average, just over half the population lives in states whose shares are made more accurate by adjustment - no matter how small the improvement.

(b) The "true" population was constructed on the basis of the synthetic assumption - no systematic variation in undercount rates within race across geography; random variation was allowed. See equation (2) in Schirm and Preston. Thus, the definition of "truth" favors synthetic adjustment.

On the whole, however, Schirm and Preston have a reasonable argument. If the assumptions of the synthetic method more or less hold, its estimates will be good. There remains the crucial question: do those assumptions hold? what kind of geographical variation is there in undercount rates? On this score, Schirm and Preston offer no evidence. In the 1980 case, the trial court found that "the synthetic method simply ignores geographical variations and assumes that a person is as likely to be missed in the census whether he lives in Alabama or in Alaska. However, as defendants' experts persuasively explained, this assumption that the undercount rates for the various age, race, and sex groups are constant from one subnational area to another has no basis in fact whatsoever ... the synthetic method is simply inadequate as a means of adjusting the census." (674 F Supp 1098, footnotes and citations omitted).

4. ADJUSTING SMALL AREAS

Statistical adjustment of census counts is more likely to be beneficial at fairly high levels of geographical aggregation (for instance, census regions or divisions). However, there are 39,000 state and local governments in the U.S., all claimants for tax money. Many of these jurisdictions are further subdivided, into city council seats, *etc.* If census counts are to be adjusted, they must for legal and policy reasons be adjusted at quite fine levels of geographical detail. Indeed, the proposal for 1990 is to adjust down to the block level. (A "block" is the smallest unit of census geography; there are 6.5 million blocks in the U.S.).

EKT discuss two synthetic methods for adjusting subareas of the 66 study areas, as well as a regression method (p. 941). In the end, however, there is no evidence that adjustment of small areas will improve on the raw census counts. With respect to 1980, EKT say (p. 943):

> "For the 66 areas included in our study, we are confident of improving upon the raw census count, especially in those areas with large undercounts or overcounts where an adjustment is most needed. Our findings do not permit definitive conclusions for suburban areas, for central cities other than the 16 included in our data set, or for other rural or urban parts of individual states. To compute estimates for such areas, we would prefer not to extrapolate from the regression equations presented in this article."

EKT go on to describe alternative designs for capture-recapture sampling, leaving open the question of small-area adjustment for 1990. Much of the dispute in 1980 centered on the feasibility of adjusting small sub-areas of the 66 study areas. To win its case, New York had to show such adjustments would improve on the census. EKT now seem to concede there was little evidence on this score.

5. AVERAGING AND SENSITIVITY ANALYSIS

The 12 PEP series were the results of a sensitivity analysis on missing data. Since the amount of missing data was large relative to the undercount, methods for handling missing data have impact. In response, EKT offer quite a variety of procedures for adjusting the census on the basis of the various PEP series, including: (a) eliminating discrepant series (pp. 937-9); (b) eliminating systematic differences between the series (pp. 937-8); (c) regression on other variables (the "composite" estimator, pp. 933ff); (d) averaging (pp. 931 and 937).

This list makes clear the essential indeterminacy of census adjustment schemes. And in this context, the use of averages to reduce indeterminacy needs discussion. Arbitrary modeling decisions may be defensible if they do not matter - the usual robustness argument. Sensitivity analysis (changing the assumptions to see if the results change) may refute the robustness argument. However, averaging the results from a sensitivity analysis is self-defeating. The different PEP series are not repeated measurements of the undercount. It is the spread in the PEP series that is interesting, not the average–because it is the spread (among, say, the April series) that demonstrates the impact of different modeling assumptions on the same data.

6. ASSUMPTIONS

EKT (p. 937) say the model improves on the PEP estimates and the synthetic method. The model does improve on the PEP estimates, if you grant its assumptions–equations (1) through (6) above. So far, however, these equations still seem quite implausible. Likewise, the model improves on the synthetic estimates only if it uses the additional variables in a sensible way, bringing us right back to assumptions.

At times, EKT seem to argue that the model can be inferred from the data (pp. 933ff). Of course, there is more to a regression model than choice of variables on the left hand side and the right hand side, although that is difficult enough, as will be seen below. There are many questions to answer: Why are effects linear and additive - equations (1) and (2) above? What about the assumptions on the errors - equations (3) through (6)? And so forth. EKT put forward no evidence to justify their assumptions, except by attempting to rebut our rebuttal (p. 931). Do they think a model is right unless it can be proved wrong?

In any case, we stand by our critique. For some data on correlation bias, see Fay *et al.* (1988, esp. sec. 6F); for a critique of Ericksen and Kadane's estimates, see Fellegi (1985, p.118). Other sources of bias in the PEP series include matching errors and errors in census-day address reports.

EKT argue that PEP is "conservative" (p. 931). This seems to be both wrong and irrelevant; wrong because the biases generally increase the apparent undercount: and irrelevant because geographical variation in the biases matters a great deal. Assumption (3) is rather unlikely: the errors probably do not have mean 0. The undercounts estimated by PEP are likely to be biased upward, the size of the bias depending on the area. For a review of the evidence, see Fay *et al.*, chap. 6; also see FN. The trial court in the 1980 case concluded:

> "The evidence at trial established that the PEP was plagued by various errors caused by inadequacies in the PEP methodology. This type of error is referred to as 'bias.' A significant source of bias in the PEP arises because the process of matching people from the CPS to the census ... is an extraordinarily difficult and inexact task. Because of inaccurate, irregular, and incomplete information in both the CPS and the census, the Bureau undoubtedly and inevitably made many errors in determining the match status of individuals enumerated in the CPS, thereby distorting the *P*-sample's undercount estimate. Moreover, the evidence at trial established that most of this matching error occurred because the Bureau erroneously determined many cases to be misses when they were in fact matches. This error, therefore, resulted in the PEP overstating the undercount. The extent of this error and the degree to which it varies from one geographic area to another is unknown." (674 F Supp 1100, footnotes and citations omitted).

We turn now to equations (4) and (5). Take the independence assumption. In 1980, there were 3 processing offices and 12 regional offices. EKT's counter: there were 400 district offices. Granted. There were also several dozen area managers, several hundred thousand census staff and about 1,500 CPS interviewers. The sources of error are numerous, and dependence seems likely. Processing offices, regional offices, managers, census interviewers, and CPS interviewers all must contribute components of error, to say nothing of respondents. Likewise, the constancy of σ^2 in (4) seems unlikely: different parts of the country are undercounted for different reasons, not readily captured in a linear regression equation.

We pointed out that random events like snowstorms might cause correlated errors in several areas; EKT respond that there were no snowstorms. This issue goes to the foundations of statistics: if the weather is good, the errors are independent; but in foul weather, all bets are off. The distributions in the model, and the statistical inferences, are therefore conditional on certain events. Which ones, and why?

Fortunately, we do not need to resolve the problem of conditional *vs.* unconditional inference. There was a major event that disrupted census operations over several states in the Pacific Northwest. Mt. St. Helens erupted in May 1980, while follow-up interviewing was in full swing.

7. OTHER ISSUES

7.1 Does it Matter which Series is Used?

At the level of precision EKT demand of the census, the different PEP series – even among their preferred ones – really do lead to quite different adjustments, as shown by equations (7) through (9). EKT, however, claim that the preferred PEP series all lead to similar adjustments. And to support their position they offer Table 11, which suggests for example that New York City has a differential undercount of 3.27% with an uncertainty of 0.62%.

For many purposes, a uniform undercount would not be material; it is differential undercounts that create inequities. The "area effects" seem to be measures of differential undercount – the policy variable of main interest.

The "area effects" in the table were computed by EKT as follows:

(i) Restrict attention to 8 of the 12 PEP series.
(ii) Smooth each of these using the regression model.
(iii) For each area, take the average of the 8 estimates.
(iv) Subtract the corresponding national estimate of undercount.

Table 6 below compares "area effects" with differences in the PEP estimates, attention being restricted to the preferred series based on the April CPS. Differences among these PEP estimates are due only to differences in the handling of missing data. Taking the range seems fair: reasons for data to be missing can differ from area to area, and so will the appropriate imputation procedure. Adding in the August series would increase the range, but some of the difference would be due to sampling error.

The table shows that for some areas, the effects are large relative to differences between PEP series, suggesting that missing data have little impact on the results. Upstate New York is an example. But for other areas, like Chicago, the reverse holds and imputation procedures matter.

All 66 areas are plotted in Figure 1. The x-axis shows the area effect; the y-axis shows the range in the preferred April PEP series. In root mean square (across the 66 areas), the spread among EKT's preferred PEP series – based on the April CPS – is about 75% of the area effect. In other words, the impact of missing data (never mind other biases in PEP) is similar in magnitude to the effect EKT are trying to measure. Bringing in alternative imputation models would make matters even worse. Nor is averaging the results a good fix, for reasons given earlier.

Table 6

Comparing Area Effects with Differences in the PEP Estimates,
Restricted to Preferred Series Based on the April CPS.
Subareas Match those used in FN.

	Preferred April PEP series			Area effect
	Min.	Max.	Range	
Alabama	–.37	.60	.97	–1.07
Alaska	2.79	3.53	.74	1.63
Los Angeles	4.56	7.72	3.16	3.16
San Diego	–.98	1.45	2.43	.65
San Francisco	4.31	6.25	1.94	2.31
Rest of California	2.84	3.92	1.08	1.03
Chicago	3.57	6.56	2.99	1.77
Rest of Illinois	1.21	1.75	.54	–1.04
New York City	6.04	7.90	1.86	3.27
Rest of New York	–1.61	–1.44	.17	–2.55
Wyoming	3.91	4.04	.13	1.16

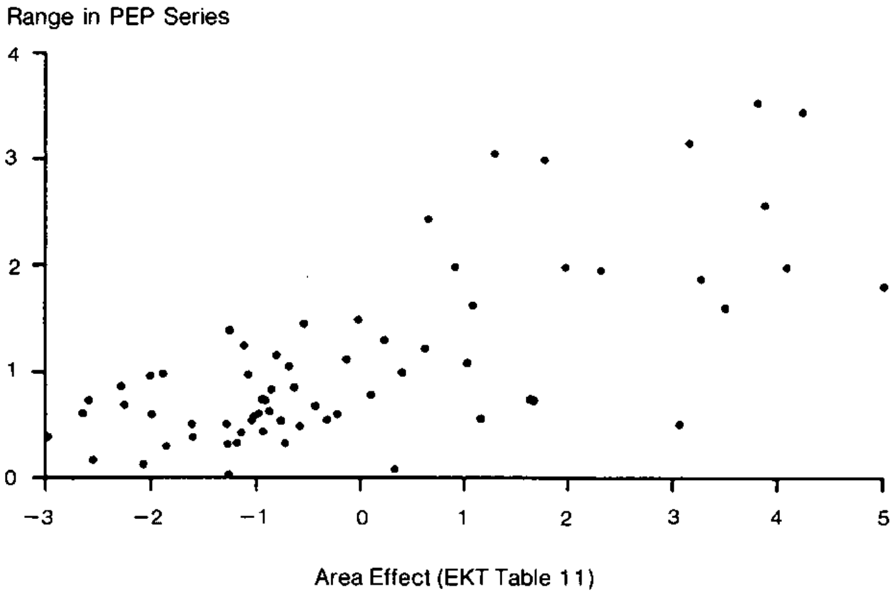

Figure 1. PEP and data quality. For each of the 66 study areas, the horizontal axis shows the EKT "area effect." The vertical axis shows the range in the preferred April PEP series.

The positive association in Figure 1 is quite striking, and so is the change in the joint distribution when the area effect changes from negative to positive. Our explanation: PEP estimates of undercount are indicators of poor data quality – in PEP as well as the census. Large apparent undercounts indicate areas with poor data. In such areas, there is a lot of missing data, so the effect of changing the imputation rules will be large too. Areas that are hard to count are also hard to adjust. See FN p. 9 or Wolter (1986, p. 26, points 8 and 9).

There may be some reasonable way of choosing a compromise version among the PEP series. But why are any of the PEP series, or their averages, an improvement over the census? That is the crucial question, and EKT do not answer it. In our view, adjustment – whether by a synthetic method, or a PEP series, or a regression model, or any convex combination – will in the end be driven mainly by assumptions.

7.2 Which PEP Series is Best, and which Explanatory Variables should be Used?

At trial, and in their discussion of FN, Ericksen and Kadane recommended an adjustment based on PEP 2-9, apparently the most preferred of all 12 series. We chose PEP 10-8 as an alternative for study. EKT defend 2-9, and try to exclude 4 of the series – especially our foil 10-8. The arguments were reviewed in court and in FN (p. 8, the discussion, and the rejoinder p. 36). Our opinion remains the same: there is no rational basis for choosing 2-9 over 10-8.

EKT impute to us the position that "proportion urban" should have been considered as an independent variable (p. 934). This is not quite right. We felt that EKT's choice of independent variables was somewhat arbitrary, and wanted to show that changing variables made a real difference to the results - another sensitivity analysis. The difference was observed mainly for small areas (FN, p. 9). Since EKT no longer advocate adjustment of small areas in 1980, this argument may be moot.

There is one new twist to the reasoning: EKT argue for choosing models by "reliance on statistical criteria (p. 941)." In essence, they recommend choosing variables so as to minimize the rms residual in an OLS fit. However, the rms residual measures association in the data not correctness of underlying theory.

For reasons that remain unclear, EKT restrict attention to models with 2, 3, or 4 variables; and they require coefficients to have t-statistics of 2 or more. Their preferred equation seems to be:

$$\text{PEP 2-9} = \underset{(-4.0)}{-2.23} + \underset{(5.4)}{.079}\,\text{min} + \underset{(3.6)}{.036}\,\text{crime} + \underset{(3.5)}{.028}\,\text{conv} + \text{residual} \tag{11}$$

$$\text{rms residual} = 1.53.$$

The right hand side variables are the percent minority in the study area, the crime rate, and the percent conventionally enumerated; t-statistics are shown in parentheses; the rms residual is computed using the unbiased divisor $n - p$. This equation is used only for variable selection; after the variables are chosen, the model is refitted by GLS: see (1-6) above, and FN for discussion.

The statistical logic is not apparent, and EKT's criteria have to be read quite literally. For example, here is another candidate equation:

$$\text{PEP 2-9} = \underset{(7.6)}{.120}\,\text{min} + \underset{(3.4)}{.026}\,\text{crime} + \underset{(3.8)}{.029}\,\text{conv} - \underset{(-4.4)}{.176}\,\text{pov} + \text{residual} \tag{12}$$

$$\text{rms residual} = 1.49.$$

The additional variable is the percentage of persons in the study area with incomes below the poverty level; the intercept was suppressed because the t-statistic was small. Equation (12) fits a little better than (11) in terms of rms residual, and "shows" that the undercount goes down as the percentage of poor people goes up - other things being equal. EKT reject this equation because the coefficient of "pov" is significantly negative rather than significantly positive.

Preconceptions about the undercount may be incompatible with the data, and best-subsets OLS may not be a suitable analytic technique. We reject neither interpretation, but our main conclusion is this. In the present context there are no objective, statistically defensible criteria for model selection. Much rides on the subjective judgment of the modeler.

With this in mind, we return to the points at issue - choosing a PEP series, and deciding between the crime rate or the percent urban as explanatory variables. As far as we can see, on the criteria chosen by EKT, the difference between crime rate and percent urban is trivial. And PEP 10-8 is clearly better than 2-9. See Table 7.

On pages 935 and 940 of EKT, σ denotes the rms residual. There is some conflict in notation, because we wrote σ^2 for $\text{Var}(\epsilon)$ in equations (2) and (4), following Ericksen and Kadane (1985, p. 105) or FN (p. 5). To avoid conflict, let $\text{SE}(\epsilon)$ be the estimated value for our σ; this is what controls the standard errors of the 66 area undercounts computed by the Ericksen-Kadane model, as shown by equations (8) and (10) in FN. For PEP 10-8, the estimated $\text{SE}(\epsilon)$ is virtually 0, so a model based on 10-8 fits extremely well and the 66 area undercounts are very precisely estimated (Table 8).

Table 7

RMS Residuals from Regression Equations for PEP 2-9 and PEP 10-8.
Explanatory Variables Include Percent Minority,
Percent Conventionally Enumerated,
and Either the Crime Rate or the Percent Urban.

	Crime rate	Percent urban
PEP 2-9	1.53	1.54
PEP 10-8	1.35	1.33

Table 8

SE(ϵ) and the RMS for the 66 Study Areas; PEP 2-9 and PEP 10-8.
The Models Include Percent Minority, Percent Conventionally Enumerated,
and Either the Crime Rate or the Percent Urban.

	Crime rate		Percent urban	
	SE(ϵ)	rms area SE	SE(ϵ)	rms area SE
PEP 2-9	.75	.65	.76	.65
PEP 10-8	.00	.28	.00	.25

Notes: Let K be a 66 $\times$ 66 diagonal matrix, whose (i,i) element is K_i. Let X be the 66 $\times$ 4 matrix of explanatory variables. Let $H = X(X^TX)^{-1}X^T$ and $\Gamma^{-1} = K^{-1} + \mathrm{SE}(\epsilon)^{-2}(I - H)$. The 66 area undercounts are estimated by the Ericksen-Kadane model as $\Gamma K^{-1} y$, where y is the 66 $\times$ 1 vector of PEP estimates. The rms SE for the 66 study areas is $\sqrt{\mathrm{trace}\,\Gamma/66}$. For details, see FN. At trial, Ericksen and Kadane estimated SE(ϵ) from 51 study areas (whole states and DC); we followed suit in FN. Here, we use the 66 study areas, since that seems to be EKT's current recommendation. The difference is noticeable.

On "statistical criteria," contrary to the claims made by EKT, 10-8 is preferred to 2-9 and percent urban is just as good an explanatory variable as the crime rate. Their qualitative critique seems off the mark too. Of course, different urban areas are different, just as EKT say. So are different central cities. Similarly, minority persons living in central cities are likely to be different from those in suburbs. And so forth. All of EKT's variables are "blurred predictors" of undercount, and some are blurrier than the percent urban (p. 934).

With respect to this set of issues, the judge in the 1980 case was harder on Ericksen and Kadane than we are:

> "Moreover, as defendants' experts persuasively explained, no one series of PEP estimates can be reliably shown to be superior to the others, or indeed, to the census itself, because there is insufficient knowledge with respect to which PEP procedures are better suited for measuring census undercount. While two of plaintiffs' experts expressed a preference for the 'series 2-9' PEP estimates based upon the hypothesis that the PEP procedures employed in arriving at those estimates were superior to the procedures used for the other PEP estimates, the plaintiffs' experts offered nothing more than unsupported assumptions in support of that position. On the other hand, the defendants' experts offered equally plausible assumptions which favored different PEP procedures, producing dramatically different PEP estimates." (674 F Supp 1102, footnotes and citations omitted.)

7.3 Simulation Studies

We had a simulation study making three points: (a) you could not infer from the data which variables go into the model, (b) standard errors depend on assumptions about disturbance terms, and (c) the standard errors computed by Ericksen and Kadane were quite optimistic. We had two additional points on this topic: (d) standard errors do not measure the impact of bias; (e) the Ericksen-Kadane smoothing simply passes through any bias in PEP that is well related to the explanatory variables.

Points (a) through (e) are real obstacles to showing that the model improves on the PEP estimates. EKT do not comment on points (b), (d) and (e). They deny (a), but more or less concede point (c). For our part, we concede that in our simulation - which grants half the model - regression does reduce sampling error. We still think (a) is right, as will be argued below. And in other contexts, smoothing may actually increase sampling error (Ylvisaker 1991 p. 7).

EKT (p. 943) criticize our study, because it covered only models with three variables in the equation and did not restrict the t-statistics. So we repeat the simulation here. In essence, we take PEP 10-8 as "truth," and add for each of the 66 study areas i a random error with variance K_i, as in (4). This grants equation (1) and the assumptions on δ_i. We choose variables according to the procedure outlined by EKT (p. 935), and fit the regression model, repeating the whole process 100 times.

Table 9 shows the variables selected in the first 10 runs. As will be seen, there is no consistency - except that the percentage "conventionally enumerated" always comes in. Over the 100 runs - excluding the ones that produced no acceptable model - the nominal rms error was about 30% too small, and improvement of the composite estimator over PEP was exaggerated by a factor of 1.75. Assumptions matter.

Table 9

A Simulation Experiment on Variable Selection; PEP 10-8 is Taken as "Truth."

Run	CC	Min	Crime	Conv	Ed	Pov	Lang	MU
1			x	x	x			
2		x		x				
3		x		x				
4	x		x	x				
5	x			x				
6		x		x				
7			x	x	x			
8		x		x				
9	x			x				
10	There was no model satisfying EKT's criteria							

Notes: CC is an indicator for central cities; Min, the percentage of minorities; Crime, the crime rate; Conv, the percentage who were conventionally enumerated; Ed, the percentage with no high school degree; Pov, the percentage below the poverty line; Lang, the percentage who have difficulty with English; MU, the percentage living in multiple-unit housing.

Table 10

A Simulation Experiment on Variable Selection.
PEP 2-9 is Taken as "Truth"; Percent Uran (Urb) is Permitted as an Explanatory Variable.
The Table Shows The Number of time Each Variable is Entered, and The Average
of its Coeeficient (Over The Time it Enters); 100 Data Sets were Generated.

Variable	No. of times entered	Average coeffient
CC	17	2.954
Min	82	0.071
Crime	53	0.053
Conv	93	0.028
Ed	5	0.085
Pov	1	0.135
Lang	17	0.315
MU	0	*****
Urb	23	0.060

A minor digression on census procedures. "Conventional enumeration" means that respondents were asked to fill out the forms and hold them for collection by an enumerator; this process was used in largely rural areas, particularly in the west. Conv is the percentage of persons living in areas that were conventionally enumerated. (In urban areas, forms were to be mailed back.) The undercount in 1980 was relatively high in rural areas, probably due to incomplete maps and address lists; that may be why conv is such a powerful explanatory variable.

We did an additional simulation with PEP 2-9 taken as truth, allowing percent urban to be selected as an explanatory variable. The results are shown in Table 10. Again, the percent conventionally enumerated comes in as does the percent minority. Otherwise, there is a fair degree of inconsistency. And the much-maligned percent urban is chosen more often than 5 of EKT's variables, including the central-city indicator. The data do not determine the model.

7.4 The Regression Model at Trial

As statisticians, we are intrigued by arguments about regression. However, the court was not impressed:

> "In their rebuttal case, the plaintiffs argued that the application of regression analysis to the undercount estimates derived from the PEP would enable the Bureau to use the PEP to accurately adjust the 1980 census. However, both plaintiffs' and defendants' experts agreed that regression analysis will not in any way alleviate the bias in the PEP and plaintiffs apparently do not contend otherwise. In short, while regression analysis may remove some of the random sampling error in the PEP, regression analysis will not reduce the substantial errors in the PEP caused by erroneous matches, the untested assumptions made with respect to the unresolved cases, and correlation bias. Moreover, the overwhelming weight of the evidence supports the conclusions of defendants' experts that the principal difficulties with the PEP stem from these biases rather than from sampling error." (674 F Supp 1103, footnotes and citations omitted.)

8. SUMMARY AND CONCLUSION

Ericksen, Kadane, and Tukey argue that they can improve on the 1980 census counts by statistical adjustment. They seem now to agree that adjustments would not have been justified for subareas of the 66 PEP study areas. With respect to the 66 areas themselves, disagreement remains. In our opinion, success of any of EKT's proposed adjustments rides on unverified and implausible assumptions–about missing data, undercount mechanisms, bias in PEP, and stochastic errors in regression models. Changing the assumptions changes the results, and taking averages over various sets of assumptions does not, at least in our opinion, make the problem go away. EKT conclude (p. 943).

> "We believe that the Census Bureau creates political difficulties for itself when it ignores the undercount. The bureau will put itself in a better position by making its best effort, using available statistical and demographic methods, to adjust for the undercount. Errors will remain, but they will be smaller and we will no longer know in advance who is losing money and power because of the undercount."

This political analysis has merit, but there are caveats. We think it quite unfair to say that the Bureau has ignored the undercount. Nor are the Bureau's political difficulties entirely of its own creation. Adjustments can indeed be devised to satisfy particular groups or settle individual law suits. However, the census is used to share out fixed resources, so there will always be losers as well as winners. These will have little trouble identifying themselves, after the fact if not before. And up to now, the goal of improving on the accuracy of the census by statistical adjustment has proved illusory.

9. HOW DID THE COURT RULE?

At the time of writing, litigation about the 1990 census goes on. With respect to the 1980 census, however, the court ruled for the defendants on all the issues. We quote from the digest and opinion *Cuomo et al. v. Baldrige et al.* 674 F. Supp. 1089-1108 (SDNY 1987).

> "State, city, and their officials brought action against Secretary of Commerce, Director of the Bureau of the Census, and other officials seeking statistical adjustment of 1980 decennial census. The District Court, Sprizzo, J., held that state and city failed to establish that statistical adjustment of decennial census was technically feasible."
>
> "... it is essential to any such adjustment that a technically feasible adjustment methodology exist which gives a truer picture of the United States population on a state-by-state basis for apportionment purposes, and a sub-state-by-sub-state basis for federal funding purposes ... If it does not, then no adjustment can or should be made ... because ... both congressional seats and revenue sharing funds are fixed quantities, and an increase in the population in one state or sub-state area will adversely affect the shares of other localities ...
>
> "Notwithstanding the complexity of the facts ... this action presents one issue to be resolved by the Court: whether the plaintiffs have sustained their burden of proving that a statistical adjustment of the 1980 census will result in a more accurate picture of the proportional distribution of the population of the United States on state-by-state and sub-state-by-sub-state basis than the unadjusted census. The Court finds as a matter of fact that the plaintiffs have not sustained that burden, and the action must therefore be dismissed ..."

ACKNOWLEDGEMENTS

Freedman is working for the Department of Justice on matters arising from the 1990 census. However, the views expressed in this article may not be shared by the Department. Helpful comments were made by L. Bazel (San Francisco), P. Diaconis (Harvard), S. Klein (RAND) and A. Tversky (Stanford). Research partially supported by NSF Grant DMS 86-01634.

APPENDIX

Synthetic Estimation and Loss Functions

Synthetic Estimation

Section 5 in Wolter and Causey (1991) describes their empirical proof that synthetic adjustment would have brought the 1980 census closer to truth. The evidence is a simulation study: the "census" and "truth" are both defined in terms of an artificial reference population developed by Isaki *et al.* (1987). However, the argument depends rather strongly on the reference population, as shown by Passel (1987). The object here is to sketch a variation on one of Passel's examples. Indeed, if the reference population is defined by using PEP 2-9 to correct the 1980 census, then synthetic adjustment moves the counts farther from truth.

Table 11 shows the data for the four census regions – Northeast, Midwest, South, and West. With squared differences in population shares weighted by size,

$$\text{r.m.s. difference between Synthetic B and PEP 2-9} = 0.21 \text{ of } 1\%. \tag{13}$$

$$\text{r.m.s. difference between the Census and PEP 2-9} = 0.15 \text{ of } 1\%. \tag{14}$$

PEP 2-9 is rather close to the "average preferred PEP" in Table 2. In that table, the census was closer to synthetic B than to the PEP estimates. In Table 11, the census is closer to PEP, and synthetic B is the outlier. The difference between the two tables seems to be the disaggregation. Table 2 disaggregates the U.S. by race and ethnicity; Table 11, according to conventional census geography.

Of course, using another disaggregation or a different synthetic adjustment could reverse the comparisons yet again; so could a change in the loss function. To illustrate the possibilities, consider adjusting the 66 PEP study areas, rather than four regions. Keep PEP 2-9 as 'truth.' Using the loss function (17), the census is preferred to synthetic B, by a little. Using (16), synthetic B shows a much smaller loss than the census.

Table 11

Population Shares from The Census, The Synthetic B Estimates, and PEP 2-9, in Percent; Census Counts, in 1,000s

	Northeast	Midwest	South	West	Total
Synthetic B – PEP 2-9	.08%	.03%	.24%	–.35%	.00%
Census – PEP 2-9	.10%	.06%	.12%	–.28%	.00%
PEP 2-9	21.59%	25.92%	33.15%	19.34%	100.00%
Synthetic B	21.67%	25.95%	33.39%	18.99%	100.00%
Census	21.69%	25.98%	33.27%	19.06%	100.00%
Census count	49,135	58,866	75,372	43,172	226,545

Loss Functions

Proponents of adjusting the 1990 census make analytic arguments based on loss functions: see Wolter and Causey (1991) or Ericksen, Estrada, Tukey and Wolter (1991, p. 20 of the main report; Appendices G and H). The essence of argument can be summarized in the lemma which follows. To set up the notation: the country is divided into n areas, indexed by i; c_i is the census count in area i and t_i is the true count. The "synthetic estimate" for area i is $x_i = \lambda c_i$, where the "adjustment factor" λ is computed from other data.

Lemma. For $i = 1, \ldots, n$ let $c_i > 0$ and $t_i > 0$. Let $0 < \lambda < \infty$ and $x_i = \lambda c_i$. (15)

Then

$$\sum_{i=1}^{n} (x_i - t_i)^2 / c_i \tag{16}$$

is minimized when

$$\lambda = \left[\sum_{i=1}^{n} t_i \right] \Big/ \left[\sum_{i=1}^{n} c_i \right].$$

The proof is omitted as trivial. The "loss function" defined by (16) differs in detail from the one used in (7), (8), (9), (13) and (14), which can be written as

$$\sum_{i=1}^{n} \frac{c_i}{C} \left[\frac{x_i}{X} - \frac{t_i}{T} \right]^2, \tag{17}$$

with

$$C = \sum_{i=1}^{n} c_i, \quad X = \sum_{i=1}^{n} x_i, \quad T = \sum_{i=1}^{n} t_i.$$

The loss function (17) emphasizes shares while (16) emphasizes counts; furthermore, (17) puts more weight on large sub-populations while (16) does the opposite, due to the division by c_i. We are not particularly attached to (17), and see no good way to choose one loss function rather than another.

Lemma (15) is mathematically correct, but it is so far removed from the realities of adjusting the 1990 census that it seems virtually irrelevant. In this connection, there are four points to consider:

(a) The true population total T is unknown; Wolter and Causey attempt to deal with this problem, but the example in Table 11 refutes their argument: synthetic adjustment makes the 1980 census less accurate.
(b) Synthetic estimates do not perform well under aggregation.
(c) At the block level, rounding error may dominate.
(d) Loss functions only capture part of the policy problem, and may obscure more than they reveal.

Points (b), (c) and (d) will be discussed in more detail; but first, a brief review of proposed methods for adjusting the 1990 census. The population is divided into 1,392 "post strata," *e.g.* male hispanic renters age 30-44 in central cities in the Pacific Division. Index these post strata by $j = 1, \ldots, 1{,}392$. For each post stratum j, an adjustment factor λ_j is computed by capture-recapture techniques from data collected in a Post Enumeration Survey (Freedman 1991).

The 1,392 factors are used to adjust all small-area counts as follows. Fix an area, *e.g.* a town. This area will intersect many of the post strata. The census count for each area × post stratum intersection is multiplied by the corresponding λ_j, and the products are summed. In other words, subpopulations are adjusted by the synthetic method, and synthetic estimates are aggregated to obtain totals for small areas.

This completes a sketch of the adjustment process, and we return to points (b), (c) and (d).

(b) Synthetic estimates do not perform well under aggregation. This was already pointed out by Fellegi (1985). See Cohen and Citro (1985, p. 318). For another example, see Tables 3 to 5 above.

(c) At the block level, rounding error may dominate. Census adjustment would in fact be done at the block level. (A "block" is the smallest unit of census geography; there are 6.5 million blocks in the country.) A typical block in an urban area may intersect 25 post strata; each block × post stratum intersection contains only a handful of people. Multiplying by an adjustment factor means adding or subtracting a fractional number of people, and the fractions would be rounded. The next example illustrates how rounding error may offset any advantage from synthetic adjustment.

Suppose there are n "areas" to adjust; these could be viewed as blocks intersected with one fixed post stratum. Suppose each of these areas has the same census count, c. Fix $m < n$. Suppose that in each of m areas, the census has missed one person; in the remaining $n - m$ areas, the census count is exactly right. In all, there is an undercount of m people. These facts are considered as known; but it is not known which blocks have the missing people. According to (16),

$$\text{loss from using the unadjusted census} = m/c. \tag{18}$$

Adjustment would proceed as follows: choose m areas at random, and add one person to each of these areas. Clearly, the expected loss from adjusting is

$$\frac{m}{n} \cdot m \cdot 0 + \left(1 - \frac{m}{n}\right) \cdot m \cdot \frac{1}{c} + \frac{m}{n} \cdot (n - m) \cdot \frac{1}{c} + \left(1 - \frac{m}{n}\right) \cdot (n - m) \cdot 0$$

$$= 2\left(1 - \frac{m}{n}\right) \cdot \frac{m}{c}. \tag{19}$$

Lemma. If $m < n/2$, there is an expected net loss from synthetic adjustment.

Proof. If $m < n/2$, then

$$2\left(1 - \frac{m}{n}\right) \cdot \frac{m}{c} > \frac{m}{c}. \tag{20}$$

Of course, this example is almost as stylized as Lemma (15). In short, the value of census adjustment cannot be established by *a priori* argument.

(d) Loss functions capture only part of the policy problem, and may obscure more than they reveal. To begin with an example, suppose that the census is in error, and the main impact of that error is to transfer a congressional seat from California to Pennsylvania. There is a gain for Pennsylvania, and a loss for California. There may be a net social loss from this misallocation, but attempting to quantify that loss by (16) - or any similar formula - seems quite simplistic.

We now present another example to illustrate point (d). To focus the issue, suppose the census undercount is largely confined to blacks and hispanics in New York, Chicago, Houston and Los Angeles. The census, by assumption, under-estimates the share of the population living in these four cities, and adjustment will partly correct that error.

Due to its reliance on the synthetic method, however, adjustment will change population shares everywhere. Areas which are heavily black and hispanic will have their population shares artificially increased, at the expense of other areas. This will be so even in regions of the country where the census was accurate.

In this example, the distribution of resources between the four cities and other areas may be made fairer by adjustment - at the expense of distortions introduced everywhere else. The loss-function approach slides over this difficulty. Balancing inequities is a political problem, not easily resolved by a statistical formula.

Some observers may consider the example to be extreme. However, the Post Enumeration Survey only samples 5,000 blocks, and there are 39,000 jurisdictions to adjust. Real information about the undercount is necessarily confined to relatively few localities. Adjustments for other areas must therefore be based largely on theory rather than data.

REFERENCES

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

ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year (with discussion). *Journal of the American Statistical Association*, 80, 98-131.

ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1989). Adjusting the 1980 census of population and housing. *Journal of the American Statistical Association*, 84, 927-943.

ERICKSEN, E.P., ESTRADA, L.F., TUKEY, J.W., and WOLTER, K.M. (1991). Report on the 1990 Decennial Census and the Post Enumeration Survey, submitted to the Secretary of the Department of Commerce, June 22, 1991.

FAY, R.E., PASSEL, J.S., ROBINSON, J.G., and COWAN, C.D. (1988). *The Coverage of the Population in the 1980 Census*. Washington, D.C.: U.S. Department of Commerce, Government Printing Office.

FELLEGI, I. (1985). Comment. *Journal of the American Statistical Association*, 80, 116-119.

FREEDMAN, D.A. (1991). Adjusting the 1990 census. *Science*, 252, 1233-1236. Copyright 1991 by the AAAS. Excerpted by permission.

FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census (with discussion). *Statistical Science*, 1, 1-39.

HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a post-enumeration survey. *Survey Methodology*, 14, 99-116.

ISAKI, C., DIFFENDAHL, G., and SCHULTZ, L. (1987). Report on statistical synthetic estimation for small areas. Technical report, Bureau of the Census.

PASSEL, J. (1987). A note about synthetic estimates of undercount. Memorandum, U.S. Bureau of the Census.

SCHIRM, A.L., and PRESTON, J. (1987). Census undercount adjustment and the quality of geographic population distributions (with discussion). *Journal of the American Statistical Association*, 82, 965-990.

SCHIRM, A.L. (1991). The effects of census undercount adjustment on congressional apportionment. *Journal of the American Statistical Association*, 86, 526-541.

WOLTER, K. (1986). Comment. *Statistical Science*, 1, 24-28.

WOLTER, K. (1991). Accounting for America's uncounted and miscounted. *Science*, 253, 12-15.

WOLTER, K., and CAUSEY, B. (1991). Evaluation of procedures for improving population estimates for small areas. *Journal of the American Statistical Association*, 86, 278-284.

YLVISAKER, D. (1991). A look back at TARO. Technical report, Department of Mathematics, UCLA.

COMMENT

STEPHEN E. FIENBERG[1]

Freedman and Navidi give their current thought-provoking retrospective on the issue of undercount in the 1980 U.S. decennial census. Unfortunately they fail to address the question posed in the title of their paper and instead attempt to vindicate their views expressed earlier in Freedman and Navidi (1986) and to rebut commentaries on these views by others. Their theme is a familiar one to those who have read earlier versions of the debate connected with the "1980 lawsuit" over adjustment: The census is very complex and only a small undercount is thought to remain; adjustment utilizes statistical modelling that relies on unverifiable assumptions; a bad adjustment may be worse than nothing.

I disagree with many of the views expressed by the authors and believe that they distort both what should have been at issue with respect to 1980 and what appears to be at issue in litigation currently pending over correction of the 1990 census. In the following, I attempt to explain my differences with the authors and give my perspective on two questions: the one raised in the title and the one implicit in the material introduced regarding the 1990 census. (Note: The aulthor played no part in the litigation over the adjustment of the 1980 census but he is working with the City of New York and other plaintiffs in litigation stemming from the decision by the Department of Commerce not to adjust the results of the 1990 census.)

1. The Title and the Paper Address Two Different Issues

Should we have adjusted the census of 1980? The only sensible way to answer this question in my mind is to ask it in the context of the evidence available at the time, or at least available when the issue was being adjudicated by the courts. As such, the description of the issues identified by the Bureau of Census and presented in the opening section of the paper are important, although they had little to do with the original decision not to adjust in 1980 made by the Director of the Bureau in advance of the availability of coverage information.

The remainder of the paper, however, does not deal with this question. Rather, it addresses the continued attempt by advocates for the two sides to marshal evidence to support their positions from the litigation. In essence, the authors are asking a question about the current evidence in support of a decade old decision. As with all statistical issues, continued data analysis and retrospection can update our judgment on the answer to such a question and thus the authors' effort to revisit the evidence connected with the 1980 census yet again is to be applauded.

We can thus turn to the framing of the question to be answered. For me, the judge's statement of the issues at trial falls short of the mark, as does Freedman and Navidi's description of the undercount issue. They imply that the only real issue is the accuracy of the adjustment process and that there is only a potentially small undercount about which we should be worried. Neither could be further from the truth. At issue is both the accuracy of the census and the adjustment process. And, it is the substantial *differential* undercount, *i.e.* the difference between the undercount for Blacks and the undercount for non-Blacks and between Hispanic and non-Hispanic, that is important when we come to assess census accuracy. This is because census figures are typically used to divide resources among groups in the population, resources such as seats in the U.S. House of Representatives; seats in state legislatures; federal funds; and so on.

[1] Stephen E. Fienberg, York University, North York, Ontario, Canada M3J 1P3.

Using the method of demographic analysis the Census Bureau has documented that, from 1940 through 1980, the difference in the rate of undercount for Blacks and non-Blacks has remained roughly constant, somewhere between 5% and 6% even though the overall undercount declined from 5.6% to 1.4% (see Fay *et al.*, 1988). The 1.4% figure does not mean that the census correctly counted over 98% of the U.S. population in 1980. Rather 1.4% represents the *net* undercount, which can be thought of as the difference between the actual undercount (consisting of missed individuals or omissions) and the overcount (erroneous enumerations and duplications). Even if the errors of overcount and undercount balanced perfectly at the national level, thus producing a 0% national undercount, we might still have a differential undercount problem. For the 1980 census, the Bureau determined that there were 6 million erroneous enumerations in the census, of which as many as 1 million were fabrications, and as many as 2.5 million people were erroneously included twice at the same location. Given the Bureau's report of a net undercount of 1.4% or 3.2 million people in 1980, we have an estimate of 9.2 million omissions (people who were missed) from the 1980 census count. By adding omissions to erroneous enumerations we get a total of 15.2 million errors in counting individuals, which corresponds to almost 7% of the official 1980 census total. To me, this level of error in the census represents a major problem that must be addressed when we talk about the appropriateness of adjustment in 1980. Of course, shortly after the 1980 census was completed the Census Bureau painted a much rosier picture of the accuracy of the raw census counts. Perhaps, in keeping with the literal meaning of the title of this paper, Freedman and Navidi wish us to accept as accurate what we now know to have been a seriously incomplete assessment on the part of the Census Bureau. I hope this is not the case. We now know much more about the level of the error in the raw census counts from 1980. The residual issue is whether we have any better information about the various forms of adjusted counts given the passage of a decade.

2. Facts and Theorems

The present paper is full of statements about the accuracy of the census adjustment procedures. When it comes to stating and proving theorems, I have no doubt that Freedman and Navidi will get them correct. The relevance of such theorems for census adjustment is a different issue.

Freedman and Navidi present a simple and seemingly compelling counterexample to the Schirm-Preston theorem on synthetic adjustment. It is certainly true that the overall totals for state A and B in their example are correct in the census and incorrect in the synthetic adjustment, although barely so. But it is also true that the large shift of the counts of Whites and Blacks in state B is what I understand that an adjustment is designed to accomplish and it does so at the expense of a minor perturbation in State A. Moreover, if the fictional state B is like those in the real U.S., the distributive accuracy of the synthetic data for geographic areas within State B is much improved while that within State A seems not to be seriously affected. Freedman and Navidi also offer their conclusion in the form of a parable to which I respond with one of my own. Small overall undercounts can hide a multiplicity of censal errors, ones that tend to "balance" in the aggregate but exact a heavy toll from states with large hard-to-count minority populations.

I also found the evidence from the Schirm-Preston simulations far more credible than did Freedman and Navidi and wonder whether this may be related to the corrected version of the Schirm-Preston theorem that is referred to as holding under more complicated sets of conditions involving weighted averages. What I am asking is whether the corrected theorem is more relevant to the real problems of undercount in the U.S. than Freedman and Navidi's counterexample.

3. Issues in Dispute with EKT

Freedman and Navidi spend much time rehashing the issue of the multiplicity of PEP series and by stressing the variations amongst them. While there is some merit in the position that there is not a clear and overwhelming choice from amongst the adjustment alternatives, it may still be the case that several choices would be superior to an unadjusted census. The authors focus on the variation amongst the full set of 12 alternatives, some of which to me are implausible given the assumptions that they rely upon. Even though I do find the arguments in support of the use of synthetic adjustments reasonable, I do agree with the authors that there is a clear difference between the synthetic and PEP adjustments.

Where are we left in this debate? I find the conclusion of Ericksen, Kadane, and Tukey compelling even though I agree with Freedman and Navidi that issues remain about the specific choice of techniques favored by EKT. Freedman and Navidi argue that their principal claim is irrelevant to the issue of accuracy. I disagree. Perhaps the authors believe that the millions of uncounted people that virtually all agree were missed in 1980 are still out hiding in the foothills of South Dakota, or in some other state with few minorities.

A familiar theme in various writings by one of the present authors is the problems that arise when assumptions are not satisfied. Here again the authors pursue this theme with respect to the linear equation used for smoothing. They appear to argue that either all assumptions must be perfectly justified or "all bets are off". Nothing could be further from the truth. Surely they don't expect anyone to believe the argument that the eruption of Mt. St. Helens interfered with census taking in a serious way and thereby undercuts the usefulness of the smoothing approach. Similarly, their notion that precise specification of predictor variables is crucial to the accuracy of smoothing is also something with which I take issue. Finally, I read the report by Ylvisaker (1991) who reexamined data from the trial census in Los Angeles in preparation for 1990, but I could not find the evidence Freedman and Navidi state is supportive of their claim that smoothing increases variability.

I do believe with Freedman and Navidi that the census process is enormously complex and that the approach to adjustment that was proposed in connection with the litigation over the 1980 census is far from flawless. Yet I still find their arguments exaggerated and they tend to obscure the old maxim that "the best is the enemy of the good." Of course the assumptions are not satisfied. Of course one could produce a better way to adjust that does not suffer from all of the flaws in the methods advocated by EKT. But this does not mean that adjustment with these flawed methods would not have been an improvement over the badly flawed unadjusted counts.

4. Adjustment in 1990

At various points throughout the paper the authors allude to comparable issues and imponderables in connection with adjustment in the 1990 census. I think that the reader should make a clear distinction between the methods used in connection with analyses presented as part of the 1980 lawsuit and those used as an integral part of the 1990 census. Many of the problems encountered by those who attempted to prepare adjusted figures in 1980 have clearly been overcome and the debate over adjustment in 1990 has become much sharper in its focus. Moreover, unlike in 1980, the key statistical methodologists at the Census Bureau, and the Director herself, found the adjustment methods used in 1990 justifiable and they recommended proceeding with an adjusted census. The statisticians were overruled by the Secretary of Commerce. The matter is now in the hands of the court once again.

Freedman and Navidi do not state their position regarding the use of adjustment techniques for the 1990 census, but Freedman (1991) makes quite clear that his judgment from 1980 has not changed. I disagree with this view. There may well be reason to argue, as the authors do, that the Census Bureau should not have adjusted the census in 1980. But 1990 is another matter. In June, the General Accounting Office, an investigative arm of the U.S. Congress, reported that there were 25.4 million gross errors in 1990 census, or about 10.4% of the resident population. The Bureau estimates that the net undercount was about 5 million people and that the differential undercount was the largest since the Bureau began to estimate it beginning with the 1940 census. Methodology for carrying out an adjustment in 1990 is much improved relative to that at issue in 1980. In my view, the results of the Census Bureau's evaluation studies clearly supported the use of adjustment for the 1990 census results. Perhaps the judge this time will see the issue of adjustment differently than the the way that Freedman and Navidi tend to frame it.

COMMENT

IVAN P. FELLEGI[1]

Freedman and Navidi provide a very thorough and lucid description of the considerations and arguments surrounding the adjustment debate for the 1980 US Census. These arguments focus on the quality of population counts and population distributions for Census day 1980. Furthermore, it is taken as given that whatever decision is made on adjustment, based on this consideration of population counts, will be applied to the complete census database and therefore to all the outputs flowing from it. Rather than commenting in detail on the arguments of the protagonists in this debate (though I am of the view that the correct decision was made for the 1980 Census), I would like to offer, from a Canadian viewpoint, some thoughts that suggest a broader frame of reference for the adjustment debate.

1. The Census is Much More Than a Head Count

Ever since the age of modern census taking began, the objective has always been more than the provision of an accurate count of the population. Yet the increasingly impressive literature dealing with the issue of adjusting the census tries to assess the relative advantages of alternative courses of action solely from the point of view of estimating the total number (and proportion) of persons living in a set of areas. I understand, of course, why this is so: (a) the problems involved are difficult enough as it is, and (b) so much money and political power is associated with the population counts (or estimates).

I will come back to point (b) above. As far as (a) is concerned, I think it would not be a scientifically defensible position to adjust the census by whatever method without taking into account the impact of such an action on the multitude of uses of census data. Indeed, I believe that if the objective of the census was restricted to estimating population totals and distributions, we would most likely (at least in Canada) try to find quite different methodologies to fulfil such a very different role. Given the multiplicity of objectives served by the Census, the fact that this multivariate and rich data base is difficult to model is not an adequate excuse for dealing with the much simpler issue of population counts and then uncritically applying the conclusions to the entire data base.

2. Point-in-time Precision of Population Counts May not Be the Relevant Measure for the Intercensal Distribution of Federal Funds and Power

There seems to be a preoccupation with exquisite precision of population counts and distributions **in the census year**. Of course a periodic stock-taking, providing good and comparable data for small areas and/or small population groups is a main justification for the expense involved in taking a census. But the excessive (it seems to me) preoccupation with the precision of the census count is motivated by equity considerations: a great deal of money and political power is distributed based, in part, on population numbers. Let us examine these two equity issues in turn.

First, dealing with the distribution of funds, indeed substantial sums are distributed in Canada from the federal government to provinces based on formulae that are very sensitive to population numbers and distributions. However, two points are of great significance from the point of view of census adjustments.

[1] Ivan P. Fellegi, Chief Statistician of Canada, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

(1) The formulae use a large array of statistical information (most of it derived from sources other than the census), only one component of which is population. It is well known that several of the other components are subject to significant sampling and non-sampling errors. It is an open question whether any reasonable loss function designed to assess the combined impact of all the errors involved would be materially improved even if the census errors could be entirely eliminated.

(2) Even more important, if the adjusted population numbers result in a smaller loss function, or if more generally they are assessed to be closer to the truth for a significant majority of the areas involved, then these can serve as the basis for improved **population estimates** (and not just in census years) without adjusting the entire multivariate census data base. In Canada (and in the United States) there is a long history of publishing estimates of the census undercount. Serious consideration is, indeed, being given in Canada to taking the next step: publish the census results as taken, and have a set of official population estimates which takes account of the known census undercount. After all, in non-census years the official population estimates incorporate a wide range of estimation techniques - some of them having errors at least as large as the likely errors of undercount estimates (even model-based ones). It may be scientifically quite appropriate to publish the best available population estimates in both census and non-census years - whether or not these estimates coincide with the census counts in census years. It may well be that legislation, or regulations under existing legislation, have to be amended to permit the use, particularly in a census year, of population estimates different from those directly derived from the census. But (a) that has little to do with the scientific arguments involved in the adjustment debate, and (b) it is more honest than relabelling the "adjusted" census counts to be "the" census counts simply because the law might require the latter.

The arguments are different in respect of the distribution of political power based on census counts, although the fixation on point-in-time precision seems to me to be equally misplaced. Indeed, the census population figures are also used to distribute seats in the House of Commons in Canada (and in the House of Representatives in the USA). However, the distribution of seats based on the census is used for ten years. During those years typically massive population shifts occur. Leaving aside the interpretation of laws, it seems to me that the substantive question is whether a suitably defined loss function, designed to capture the average deviation from the objective of "one person one vote" **over a ten year period**, would be materially reduced if the census counts were adjusted for the estimated undercount. I have not made such a calculation. However, it seems to me that the range of population shifts over ten years are substantially larger than the range of estimated undercounts. I would therefore, speculate that even apparently significant potential census year adjustments (and corresponding shifts in the allocation of seats in the legislature) are relatively less significant than the deviations from the "one person one vote" rule occasioned by migration over a ten year period. Since this particular use of the census is mandated by the constitution, changing the law is not an option. But a scientifically informed debate regarding the appropriate interpretation of the constitution is very much in order - taking full account of the two main causes of deviation from equity in political representation during the ten year intercensal period: census errors and population shifts (mostly migration).

3. Conclusion

(a) The census is a multivariate integrated data base. The case for "adjusting" it is far from obvious, even if (a big if) the simplest variable involved - the count - can be improved by doing so.

(b) If a set of population estimates that are judged to be better than the census counts (according to suitably defined criteria) can indeed be generated, these should be produced and used, without necessarily adjusting the entire census data base. The criteria should relate to the set of areas (and other breakdowns) for which estimates are required.

(c) If the law requires "census" derived population counts when in fact substantively the best available population estimates are called for, it would appear to be preferable to try to change the law rather than to adjust (in effect weight) the entire census data base to agree with estimated population numbers - solely in order to be able to refer to the population estimates as "the census".

(d) Equity considerations, both in terms of the distribution of federal funds and political representation, apply to the entire intercensal period, not simply for the year of the census. They should be studied using models that take full account of this fact.

COMMENT

N. CRESSIE[1]

A critical assessment of our past successes and failures makes us better equipped to provide future successes. Missing data and matching problems in the 1980 Post Enumeration Program were major impediments to a successful adjustment of the 1980 U.S. Decennial Census. A court case, Cuomo *et al.* versus Baldridge, was brought by New York State and others to require the Census Bureau to adjust the 1980 Census numbers for undercount. Testimony from Barbara Bailar, then Associate Director of Statistical Standards and Methodology at the Census Bureau, and Kirk Wolter, then Chief of the Statistical Research Division at the Census Bureau, made it clear that 1980 data and methods were inadequate for an accurate adjustment of the whole country.

In 1987, Judge Sprizzo ruled against New York. However, that decision did not make the differential undercount go away; even the judge in his ruling acknowledged its presence. There is little disagreement that, differentially by race, national U.S. Census numbers have been persistently too small. Using demographic methods, the following estimates are available.

1950: Black (and other non whites) demographically estimated undercount was 9.7%. White demographically estimated undercount was 2.5%. (Siegel 1974, Table 3).

1960: Black (only) demographically estimated undercount was 8.0%. White (and other races) demographically estimated undercount was 2.1%. (Siegel 1974,Table 2, set D estimates).

1970: Black (only) demographically estimated undercount was 7.6%. White (and other races) demographically estimated undercount was 1.5%. (Passel, Siegel and Robinson 1982, Table 1).

1980: Black (only) demographically estimated undercount was 5.3%. White (and other races) demographically estimated undercount was −0.2%. (Passel and Robinson 1984, Table 2).

Further, there is little disagreement that racial composition is different within administrative regions (both large and small) across the U.S.A. The consequence of these two virtually undeniable facts is that undercount will be differential across administrative regions, leading to an unrepresentative geographic/racial profile of the nation and an unfair apportioning of political and financial resources. So, Freedman and Navidi state in their introduction " ... If the undercount can be estimated with good accuracy, especially at the local level, adjustments can - and should - be made to improve the census."

Almost everyone agrees there is a problem. The adage, "If it ain't broke don't fix it," does not apply here. It is an uncomfortable defence for a statistics professional to argue that uncontrolled-for biases and errors will not allow an adjustment for an undercount that is known to be there and known to be damaging. During the early 1980s, Bailar and Wolter established the Undercount Research Staff within the Statistical Research Division of the Census Bureau. Staff members have produced high-quality research that demonstrated "that it is technically feasible to correct the 1990 Census for differential undercoverage": (Childers *et al.* 1987).

It is time for Freedman and Navidi to relinquish their role as devil's advocates; it is time for them to put their knowledge and talents into a constructive mode; and it is time for them to say what they mean by "good accuracy," "local level," and various other qualitative affirmations. The adversarial atmosphere of the courts has spilled over into the various articles,

[1] N. Cressie, Department of Statistics, Iowa State University, Ames, IA U.S.A. 50011.

comments, and rejoinders we have seen on census undercount in the last 10 years. To solve a problem as hard as adjustment for undercount, the common goal needs to be recognized. From there, debate should center around differences on how that goal might be reached. If Freedman and Navidi's position is that the goal is impossible to reach (which is what they seem to have implied over the years), then it should be stated.

For the rest of this comment, I shall address a number of important technical matters that were raised by Freedman and Navidi (1986) and now, surprisingly, again in the article under discussion. In 1990, I presented a paper at the Census Bureau's Annual Research Conference (Cressie 1990) that David Freedman was invited to discuss. At the last minute, he was unable to attend the conference but I continued to send him the discussion version and the final version and invited his comments. The paper is rather technical but addresses, successfully I believe, several major criticisms made by Freedman and Navidi (1986) of the statistical modeling approach to undercount adjustment.

First, the paper expresses a preference for the "stratification approach" over the "regression approach". Stratification is a special case of regression where the explanatory variables are restricted to 1 and 0, indicating presence or absence in a particular (demographic) stratum. There is little disagreement that undercount is differential across sex × age × race/ethnicity strata. Because the Census Bureau was committed to a regression approach, the bulk of the paper addressed the more general problem.

Second, if one allows the regression error (see Freedman and Navidi's *e.g.* (2)) to be dependent, such models can absorb bias and misspecification into the error term. The important concept to maintain is that true undercount in regions is unknown and the ignorance is quantified into a probability model. The goal is *not* estimation of the coefficients β but *prediction* of the undercount. With an error term that does not have to be independent and identically distributed, this prediction is insensitive to misspecification (see also Cressie 1991, Chapter 3).

Third, the inconsistency of the model to changes in geographic level is addressed by modeling adjustment factors, not undercounts, and by assuming the variance of the regression error of a particular area is inversely proportional to that area's population. This assumption is justified, from both a Bayesian and frequentist point of view, in Cressie (1989).

Fourth, the effect of estimation of variance-covariance parameters can be taken into account by modifying the results of Prasad and Rao (1990) to a multivariate context. One could also use a parametric boostrap, by generating data from the estimated model, re-estimating all parameters, and repredicting the undercount.

Finally, it is acknowledged that all preceding model-based methods will likely do poorly if the model does not fit. Diagnostic methods are crucial to the success of statistical model-based adjustments for undercount.

There is room for critical assessment of our past successes and failures. It is time to move on and solve this monumentally important problem with cutting-edge technology. A well designed, well implemented, and quality-assured 1990 Post Enumeration Survey with excellent computer matching and precise geography make the 1980 case look very different indeed. It is my opinion that adjustment can now be successfully carried out at the state level. Research and debate on whether that success can be carried down to lower levels of geography deserves our collective resources (*e.g.* Tukey 1983; Cressie 1988; Wolter and Causey 1991). *Expected* losses (or risks) can be used to measure the efficacy of adjustment procedures. Cressie (1988) gives sufficient conditions under which synthetic adjustment improves over census count; those conditions were satisfied in the 1980 Census and PEP 3-8 series.

ADDITIONAL REFERENCES

CHILDERS, D., DIFFENDAL, G., HOGAN, H., SCHENKER, N., and WOLTER, K. (1987). Paper presented in the Session *1990 Census Undercount and Adjustment*, American Statistical Association Annual Meetings, San Francisco, California.

CRESSIE, N. (1988). When are census counts improved by adjustment? *Survey Methodology*, 14, 205-222.

CRESSIE N. (1989). Empirical Bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.

CRESSIE, N. (1990). Weighted smoothing of estimated undercount. *Proceedings of Bureau of the Census 1990 Annual Research Conference*. Bureau of the Census, Washington, D.C., 301-325.

CRESSIE, N. (1991). *Statistics for Spatial Data*. New York: Wiley.

PASSELL, J.S., and ROBINSON, J.G. (1984). Revised estimates of the coverage of the population in the 1980 Census based on demographic analysis: A report on work in progress, in *Proceedings of the Social Statistics Section, American Statistical Association*, 160-165.

PASSELL, J.S., SIEGEL, J.S., and ROBINSON, J.G. (1982). *Coverage of the National Population in the 1980 Census, by Age, Sex and Race: Preliminary Estimates by Demographic Analysis.* Current Population Reports, Special Studies P-23, No. 115. Bureau of the Census, Washington, D.C.

PRASAD, N.G.N., and RAO, J.N.K. (1990). On the estimation of mean square error of small area predictors. *Journal of the American Statistical Association*, 85, 163-171.

SIEGEL, J.S. (1974). Estimates of the coverage of the population by sex, race and age in the 1970 Census. *Demography*, 11, 1-23.

TUKEY, J.W. (1983). Affidavit, presented to District Court, Southern district of New York. *Cuomo et al. versus Baldridge*. 80 Civ. 4550 (JES).

COMMENT

ALLEN L. SCHIRM and SAMUEL H. PRESTON[1]

1. Introduction

We thank the editor for inviting us to comment on this provocative article by Freedman and Navidi (hereafter, "F and N") and continue this important policy debate. Our comment mainly responds to F and N's criticisms of our earlier research (Schirm and Preston 1987; hereafter, S and P). Although we disagree with much of F and N's critique of Ericksen, Kadane and Tukey (1989), we leave to the authors of that article the task of defending their work.

We disagree with many of F and N's specific criticisms of S and P. Before discussing our detailed responses, we want to take a broader perspective and view our article and F and N's criticisms of it in their entirety.

F and N wrongly characterize our article in stating that we "present two major arguments, one analytical and one based on simulation." In fact, we presented three analytical results. F and N criticize only one, and a minor one at that. Our most important analytical result suggests that synthetic adjustment would likely have improved the accuracy of the population distribution in 1980. As for our simulations, they were not intended to support any one argument. Instead, we simulated an extremely wide variety of circumstances to permit us to address several questions about synthetic adjustment and its effects. We found, however, that adjustment would have improved accuracy under all conditions simulated, including highly unfavorable circumstances.

2. Analytical Results

In S and P, we presented three analytical results. All three are mathematically correct. However, the second result - the sole target of F and N's criticism - is, as we stated in our article, potentially "misleading because it ignores influences on overall adjustment success of systematic relationships between variations across states in census coverage for a group and differences between groups in how they are distributed across states." Our third result, which is clearly the focus of our algebraic analysis and which does not depend on the second result, addresses this issue and takes into account the patterns of variations in undercounts across states. Although potentially misleading, we presented the second result to illustrate more forcefully a key implication of our third result, that systematic variations in state undercounts can matter.

Our second analytical result suggests that the effect of adjustment for a given state hinges on how "close" the state's undercounts are to the national undercounts. Contrary to F and N's claim, our second analytical result is mathematically correct. F and N are able to dispute our finding only because they choose to define "close" without regard to our precise definition. Thus, F and N's "counterexample" to our second result does not pertain to that result at all, since their example violates the conditions that we derived and stated precisely in the appendix to our article. To repeat that result, we showed that the estimated proportion of the total national population residing in state i is made more accurate by adjustment if

$$\left| \sum_{j=1}^{J} \frac{N_{ji}^C}{N_{..}^C} \left(\frac{N_{..}^T}{N_{..}^C} - \frac{N_{ji}^T}{N_{ji}^C} \right) \right| > \left| \sum_{j=1}^{J} \frac{N_{ji}^C}{N_{..}^C} \left(\frac{N_{j.}^T}{N_{j.}^C} - \frac{N_{ji}^T}{N_{ji}^C} \right) \right|,$$

[1] Allen L. Schirm, Mathematica Policy Research, Inc., 600 Maryland Avenue, S.W., Suite 550, Washington, D.C. USA. 20024-2512, Samuel H. Preston, Population Studies Center, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA USA 19104-6298.

where J is the number of racial (more generally, demographic) groups, a dot indicates summation over an index, and T and C superscripts designate true and census population counts, respectively. For state A in F and N's "counterexample," this expression implies

$$\left| \frac{90}{1{,}100}\left(\frac{1{,}100}{1{,}100} - \frac{89}{90}\right) + \frac{1}{1{,}100}\left(\frac{1{,}100}{1{,}100} - \frac{2}{1}\right) \right| = 0$$

$$\not> \frac{21}{13{,}750} = \left| \frac{90}{1{,}100}\left(\frac{979}{1{,}000} - \frac{89}{90}\right) + \frac{1}{1{,}100}\left(\frac{121}{100} - \frac{2}{1}\right) \right|.$$

Therefore, the condition for improved accuracy from adjustment is violated for state A. Similarly, for state B, we get

$$\left| \frac{910}{1{,}100}\left(\frac{1{,}100}{1{,}100} - \frac{890}{910}\right) + \frac{99}{1{,}100}\left(\frac{1{,}100}{1{,}100} - \frac{119}{99}\right) \right| = 0$$

$$\not> \frac{21}{13{,}750} = \left| \frac{910}{1{,}100}\left(\frac{979}{1{,}000} - \frac{890}{910}\right) + \frac{99}{1{,}100}\left(\frac{121}{100} - \frac{119}{99}\right) \right|.$$

Again, the condition for improved accuracy from adjustment is violated.

F and N's "counterexample" says nothing about our second analytical result. However, it is useful for numerically illustrating our third and clearly most important result. According to that result, when blacks are most heavily undercounted where they are *least* prevalent and whites are most heavily undercounted where they are *most* prevalent, synthetic adjustment *may* not improve the accuracy of the proportionate distribution. In F and N's example, state A has a higher black undercount than state B (50% versus 17%) but proportionately fewer blacks (2% versus 12%). State A has a higher white undercount (smaller overcount) than state B (−1% versus −2%) and proportionately more whites (98% versus 88%). Therefore, F and N's finding that the adjusted estimates in their example are less accurate overall than the census estimates, although not guaranteed, is not surprising in light of our third analytical result.

F and N's critique of our algebraic analysis of the effects of adjustment is based on a highly selective reading of our article that misrepresents our findings. F and N's criticism of our second analytical result is wrong as is their characterization of that result as central to our article. Our third analytical result is by far more important. It helps to expose those conditions on which adjustment's success or failure depends. Based on available empirical evidence cited below, the conditions of F and N's numerical example did not prevail in 1980, and our result suggests that synthetic adjustment would have improved the accuracy of the geographic distribution.

3. Simulation Results

As noted before, the purpose of our simulations was to answer several questions pertinent to synthetic adjustment and its effects on the accuracy of population estimates. The central questions addressed in our article were:

- How often would synthetic adjustment improve the accuracy of population estimates?
- How much would synthetic adjustment typically improve the accuracy of population estimates?

- Do the effects of synthetic adjustment on accuracy depend on how much census coverage varies from state to state?
- Do the effects of synthetic adjustment on accuracy depend on how well we measure national undercounts?

F and N focus on the second question. For the most part, we agree that the average magnitude of improvement in accuracy from synthetic adjustment is modest if our conservative assumptions about the state of the nation pertain. Under Case 22 of Scenario I, which probably exaggerates interstate variations in census coverage but is presented in S and P as our "moderate" variation case, the average reduction in the weighted sum of squared errors is just 8% while the average reduction in the unweighted sum of absolute errors is only 4%. It is important to understand, however, that larger improvements could be realized, as suggested by our third analytical result presented in S and P. The gains in accuracy would be somewhat greater, for example, if Hispanics had the same national undercount as blacks and were included with blacks instead of whites. In that case, the average reduction in the weighted sum of squared errors would be over 12%. The gains in accuracy would also be greater if black undercounts were higher in states with proportionately more blacks. We will return to this point shortly. Of course, improvements from synthetic adjustment might be smaller if there were substantial errors in measuring undercounts, although as we showed in S and P, the effects of measurement error are generally small.

What is easily forgotten in assessing the average gain in accuracy is the likelihood of realizing some gain, large or small. F and N are guilty of this oversight. Under the assumptions of Case 22, Scenario I, the likelihood of a gain in accuracy, according to the weighted sum of squared errors criterion, is 84%. We are impressed by this finding. Some improvement, perhaps only modest, is highly likely.

This result and the result on the average magnitude of improvement raise critical questions. What is the implication of the average improvement being "only modest"? Does the average improvement have to be overwhelming to justify adjustment? Put differently, should adjusted estimates be held to a higher standard than census estimates? The secretary of commerce imposed a higher standard in making the 1990 adjustment decision. How would the Census Bureau's coverage improvement and imputation procedures fare by an equally high standard? We suspect that some would not fare well, having almost certainly exacerbated rather than ameliorated the differential undercount. Finally, would adjustment be recommended if it did little to improve accuracy but reduced systematic inequity? We will return to this last question in Section 4.

F and N answer these questions - which, by and large, do not have statistical answers - only implicitly, if at all. They suggest, however, that adjustment might be attractive (its estimates "will be good") if the assumptions of our paper hold, the issue to which we now turn.

F and N wrongly characterize both the synthetic method and our simulation model. The underlying assumption of the synthetic method is not that there is no systematic geographic variation in undercounts for a given race but that there is no variation at all. Our simulation model shows how synthetic adjustment performs when this synthetic assumption is violated. We considered cases of extreme, albeit nonsystematic, interstate variation in undercounts by race, as well as cases with more moderate random variation. We did not construct true populations "on the basis of the synthetic assumption," and our "definition of truth" did not "favor synthetic adjustment." As we showed analytically in S and P, synthetic adjustment would have been favored by assuming a positive association between the black undercount and the prevalence of blacks or a negative association between the white undercount and the prevalence of whites. (A precise statement of the result is contained in the appendix to S and P.)

For purely illustrative purposes, we assumed in a new round of simulations that white undercounts are generated according to the assumptions of Scenario I, Case 22 in S and P but that the expected black undercount rises with the state's proportion black such that the black undercount is 2.0% when the proportion black is 11.7% (the national proportion black in 1980 according to the census) and 5.2% when the proportion black is 20.0%. Under those conditions, which we do not claim to be realistic although they preserve the average simulated differential in national undercounts, synthetic adjustment improves the accuracy of the proportionate distribution according to the weighted sum of squared errors criterion in all 1,000 iterations. The average reduction in the weighted sum of squared errors is over 17%, despite extreme variation in state total undercounts.

Do the assumptions of S and P pertaining to interstate variations in undercounts hold? Probably not. Although one of our purposes was to simulate a wide range of circumstances, it is very likely that our assumptions tended to put adjustment at a disadvantage.

Did we "offer no evidence" on the matter of geographic variation, as F and N claim? No, although admittedly there was not a wealth of information available. For judging our assumptions and their implications, there are two relevant empirical issues: whether variation is systematic or random and the extent of variation. We addressed both in S and P.

As we noted in S and P, according to the 1980 PEP blacks are hardest to count where they comprise large proportions of the population. In contrast, there is essentially no relationship between the white undercount and the relative prevalence of whites. (Ericksen and Kadane 1983). These conclusions are based on broad categories measuring racial composition and data for Standard Metropolitan Statistical Areas and state remainders, not state-level data. The only published undercount estimates by state and race are the "Developmental Estimates" for 1970. Although seriously flawed, based on heroic assumptions about internal migration (Wolter 1987), those estimates imply a direct relationship between the black undercount and the prevalence of blacks and a weak inverse relationship between the white undercount and the prevalence of whites. By ignoring either pattern of covariation, the simulations in S and P tend to *understate* the gains in accuracy from synthetic adjustment.

Since writing S and P, we have obtained unpublished state population and undercount estimates by race from the 1980 PEP. Because the raw black undercount estimates are imprecise for several states, it is not clear whether blacks are hardest to count – at the state level – where they are most prevalent. For whites, although there is evidence of a direct, rather than an inverse, association between their prevalence and the undercount, we believe that this is attributable to the inclusion of Hispanics in the white population and, to a much smaller degree, to the relatively heavy reliance on the conventional method of enumeration in a few predominately white states in the western U.S. Indeed, we find that if the true 1980 population followed the pattern of either the Series 2-9 or 10-8 estimates, a synthetic adjustment for the differential between the undercount of blacks and Hispanics and the undercount of all other persons would almost certainly have improved accuracy.

The available empirical evidence generally suggests that geographic variations are, if not random, systematic with a pattern that would enhance the gains in accuracy from synthetic adjustment. It seems unlikely that there is a strong inverse association across states between the black undercount and the prevalence of blacks or a strong direct association between the white undercount and the prevalence of whites. (Even if one or both of these patterns existed, adjusted estimates might still be more accurate, as we showed in S and P.) Thus, our assumption of randomness in S and P was probably conservative, working against synthetic adjustment.

Our assumptions about the extent of interstate variations in undercounts were also probably conservative, as we discussed in S and P with reference to the 1970 Developmental Estimates. Due to substantial sampling errors for many states, the unpublished state undercount estimates by race from the 1980 PEP do not reliably reveal how much black undercounts vary across states. The variance in black undercounts calculated across all 51 states is 0.0128 for Series 2-9, twice the highest assumed value in our simulations. The variance falls to 0.0036, not even midway between the moderate and high variances simulated, when New Hampshire (black undercount equal to −60%) and Vermont (black undercount equal to −24%) are excluded. (For Series 10-8, the interstate variance is nearly equal to the moderate value simulated in S and P, if three states with extreme (and highly unreliable) undercounts - less than −20% or greater than 20% - are excluded.) Raw estimates of state undercounts for whites from the 1980 PEP are far more precise. The interstate variances for Series 2-9 and 10-8 are just slightly below the moderate value simulated. The gains in accuracy under S and P's Scenario I, Case 12 (high variation among black undercounts and moderate variation among white undercounts) differ little in frequency or magnitude from the gains under Case 22 (moderate variation for both black and white undercounts), where improvements are highly likely.

From published PEP estimates for 1980, we can only calculate variances in total state undercounts, not differentiated by race. The largest interstate variance among the 12 published PEP Series is 0.00034, slightly less than the average simulated variance for our moderate variation case (Case 22). For Case 32 (low variation among black undercounts and moderate variation among white undercounts), the average simulated variance is 0.00031, about equal to the interstate variance for PEP Series 2-9, which is favored by Ericksen, Kadane and Tukey (1989) and is the median variance across the 8 PEP Series remaining after excluding 10-8, 14-8, 14-9, and 14-20. Synthetic adjustment reduces the weighted sum of squared errors by about 12% on average under Case 32, compared to 8% for Case 22. Case 23 (moderate variation among black undercounts and low variation among white undercounts) implies an average simulated variance only slightly greater than the variance for PEP Series 10-8, F and N's "favorite." Under the conditions of Case 23, synthetic adjustment reduces the weighted sum of squared errors by 19% on average. Adjusted estimates are more accurate over 92% of the time according to that error criterion. Are such improvements "only modest"?

4. Accuracy and Equity

We have argued before, in S and P and in Schirm (1991), that the foremost concern of statisticians and demographers should be the accuracy of population estimates. Yet, in a single-minded pursuit of statistical accuracy, it is easy to forget considerations of political equity.

A more accurate population distribution is probably more equitable, in general. However, this does not imply that two equally accurate distributions are equally equitable. Although adjustment may do little to improve overall accuracy in a particular year, it may reduce or remove certain systematic errors and systematic inequity, errors and inequity associated with race.

An example, obtained from our simulations, is displayed in Table 1. The implied black and white national undercounts are 5.2% and −1.1%. The adjusted population estimates in Table 1 were obtained using these figures and the synthetic method.

As will become clear, it is hard to draw a sharp distinction between accuracy and equity. For this discussion, we assume that accuracy is narrowly defined in terms of the proportionate geographic distribution.

Table 1

A Numerical Example: Population Counts (1,000s)

State	True		Census		Adjusted	
	Black	White	Black	White	Black	White
Alabama	1,060	2,849	996	2,898	1,051	2,866
Alaska	14	375	14	388	15	384
Arizona	83	2,606	75	2,643	79	2,614
Arkansas	414	1,916	374	1,912	395	1,891
California	1,758	22,105	1,819	21,849	1,919	21,607
Colorado	106	2,719	102	2,788	108	2,757
Connecticut	226	2,875	217	2,891	229	2,859
Delaware	101	507	96	498	101	492
District of Columbia	442	181	449	189	474	187
Florida	1,415	8,290	1,343	8,403	1,417	8,310
Georgia	1,642	3,910	1,465	3,998	1,546	3,954
Hawaii	18	925	17	948	18	938
Idaho	3	936	3	941	3	931
Illinois	2,014	9,677	1,675	9,752	1,768	9,644
Indiana	443	4,775	415	5,075	438	5,019
Iowa	43	2,809	42	2,872	44	2,840
Kansas	135	2,284	126	2,238	133	2,213
Kentucky	254	3,327	259	3,402	273	3,364
Louisiana	1,239	2,856	1,238	2,968	1,306	2,935
Maine	3	1,135	3	1,122	3	1,110
Maryland	1,026	3,159	958	3,259	1,011	3,223
Massachusetts	230	5,223	221	5,516	233	5,455
Michigan	1,272	7,990	1,199	8,063	1,265	7,974
Minnesota	56	4,037	53	4,023	56	3,978
Mississippi	891	1,629	887	1,634	936	1,616
Missouri	535	4,329	514	4,403	542	4,354
Montana	2	761	2	785	2	776
Nebraska	51	1,521	48	1,522	51	1,505
Nevada	52	760	51	749	54	741
New Hampshire	4	911	4	917	4	907
New Jersey	1,071	6,237	925	6,440	976	6,369
New Mexico	27	1,264	24	1,279	25	1,265
New York	2,397	14,891	2,402	15,156	2,535	14,988
North Carolina	1,387	4,443	1,319	4,563	1,392	4,512
North Dakota	3	622	3	650	3	643
Ohio	1,112	9,769	1,077	9,721	1,136	9,613
Oklahoma	208	2,819	205	2,820	216	2,789
Oregon	39	2,602	37	2,596	39	2,567
Pennsylvania	1,177	10,750	1,047	10,817	1,105	10,697
Rhode Island	29	923	28	919	30	909
South Carolina	962	2,182	949	2,173	1,001	2,149
South Dakota	2	678	2	689	2	681
Tennessee	754	3,764	726	3,865	766	3,822
Texas	1,752	12,421	1,710	12,519	1,804	12,380
Utah	10	1,435	9	1,452	9	1,436
Vermont	1	523	1	510	1	504
Virginia	1,116	4,356	1,009	4,338	1,065	4,290
Washington	109	3,867	106	4,026	112	3,981
West Virginia	69	1,877	65	1,885	69	1,864
Wisconsin	198	4,588	183	4,523	193	4,473
Wyoming	3	448	3	467	3	462
Total	27,958	197,836	26,495	200,054	27,956	197,838

Note: "White" includes all nonblacks.

Table 2

A Numerical Example: Congressional Apportionments

State	Number of House Seats		
	True	Census	Adjusted
Alabama	8	7	8
Alaska	1	1	1
Arizona	5	5	5
Arkansas	4	4	4
California	46	45	45
Colorado	5	6	5
Connecticut	6	6	6
Delaware	1	1	1
District of Columbia	0	0	0
Florida	19	19	19
Georgia	11	10	11
Hawaii	2	2	2
Idaho	2	2	2
Illinois	22	22	22
Indiana	10	10	10
Iowa	5	6	6
Kansas	5	5	5
Kentucky	7	7	7
Louisiana	8	8	8
Maine	2	2	2
Maryland	8	8	8
Massachusetts	10	11	11
Michigan	18	18	18
Minnesota	8	8	8
Mississippi	5	5	5
Missouri	9	9	9
Montana	2	2	2
Nebraska	3	3	3
Nevada	2	2	2
New Hampshire	2	2	2
New Jersey	14	14	14
New Mexico	3	3	3
New York	33	34	33
North Carolina	11	11	11
North Dakota	1	1	1
Ohio	21	21	21
Oklahoma	6	6	6
Oregon	5	5	5
Pennsylvania	23	23	23
Rhode Island	2	2	2
South Carolina	6	6	6
South Dakota	1	1	1
Tennessee	9	9	9
Texas	27	27	27
Utah	3	3	3
Vermont	1	1	1
Virginia	11	10	10
Washington	8	8	8
West Virginia	4	4	4
Wisconsin	9	9	9
Wyoming	1	1	1
Total	435	435	435

Although the census and adjusted distributions in Table 1 are equally accurate, according to a weighted sum of squared errors criterion, the adjusted estimates more accurately reflect the racial distribution at the national level and are more equitable. (The census geographic distribution is slightly more accurate according to a sum of absolute errors standard.) The true and adjusted figures imply that 12.4% of the U.S. population is black. According to the census, only 11.7% of the population is black, a serious inequity.

The equity gains from adjustment are made more concrete by the implied congressional apportionments shown in Table 2. Both the census and adjusted estimates allocate one too many seats to Iowa and Massachusetts and one too few to California and Virginia. However, the census estimates also allocate one too many seats to Colorado and New York and one too few to Alabama and Georgia, whereas the adjusted estimates allocate the correct numbers of seats to these states. Based on the census figures, Alabama and Georgia are denied representation because of their high proportions black and the differential undercount to which blacks are subject. Adjustment substantially improves equality of representation. Based on the true population figures and the census apportionment, there are 471,000 persons per representative in Colorado, 508,000 in New York, 555,000 in Georgia, and 558,000 in Alabama. Adjustment narrows the differences, with 565,000 persons per representative in Colorado, 524,000 in New York, 505,000 in Georgia, and 489,000 in Alabama. For the four states combined, there are 519,000 persons for each of the 57 representatives. (For the entire U.S., there are 519,000 persons for each of the 435 representatives.) Adjustment reduces the (unweighted) root mean square deviation from this average by over 21%. (The reduction is between 20% and 21% when deviations are weighted by the number of persons per representative according to the true population figures.) The equity gain from adjustment is also clearly revealed by the weighted average of persons per representative calculated across all 50 states. When weighted by the proportion of the national black population (exclusive of the District of Columbia) living in the state, the true average number of persons per representative is 518,000. If Congress is apportioned according to the census estimates, the average is 524,000. Synthetic adjustment removes most of this racial inequity. The average number of persons per representative is 520,000 when House seats are allocated according to the adjusted estimates. Although there are surely still other ways to measure inequality of representation, it is hard to imagine a reasonable alternative that would not show adjustment reducing the racial inequity attributable to differential census undercounting. The gain in equity in this example is achieved despite no gain in accuracy of the proportionate distribution across states.

The errors in the census are systematic. After adjustment, the remaining errors may not be truly random, and so long as there are errors, there will be inequity. However, the source of those errors would be far less offensive than race.

5. Discussion

In criticizing S and P and Ericksen, Kadane and Tukey (1989), F and N emphasize the role of assumptions underlying adjusted estimates. Their view, however, is extreme, counterproductive, and fundamentally flawed.

Although it is reasonable - and necessary - to ask whether assumptions matter, assumptions do not have to be exactly true as F and N imply. Moreover, proponents of adjustment should have to defend it against only reasonable alternative assumptions. F and N seem to believe that almost any alternative is fair game. As in assessing the magnitude of improvement, they require adjustment to bear a heavy burden of proof with no scientific justification. Nonetheless, as our simulations show, synthetic adjustment improves accuracy even under extremely unfavorable - and probably unreasonable - assumptions.

This raises another important point of disagreement between us and F and N. Not all assumptions have to imply precisely the same estimates if all adjusted estimates based on reasonable assumptions are more accurate than census estimates. Unless equally plausible assumptions have very different implications, we should not reject the better for failure to find the best. We should not settle for census estimates that are less accurate.

F and N offer nothing to suggest that census estimates are more accurate than adjusted estimates. The legal findings on which they rely are no basis for a scientific argument. Moreover, despite Schirm's (1991) finding that the judgmental decisions made in producing census estimates can affect congressional apportionment, F and N fail to scrutinize census procedures and the underlying assumptions. Do they find plausible the census "assumption" that when the final estimates are released, everyone everywhere has been counted in exactly the correct location? If not, do they have any constructive suggestions for improving the accuracy of population estimates? Unfortunately, their critical commentary on those suggestions that have been offered is seriously flawed by misrepresentation and distortion and offers nothing constructive.

"Should we have adjusted the census of 1980?" as F and N ask. Maybe, maybe not. Although it is subject to debate, we may not have known enough about the likely effects of adjustment or been technically and operationally prepared to undertake an adjustment at the time a decision had to be made. Would adjustment have improved accuracy in 1980? We cannot answer with certainty because the true population is inherently unknowable and anomalies cannot be entirely ruled out. With that qualification, the answer is "very likely".

Acknowledgements

The authors thank Gene Ericksen for providing unpublished estimates from the1980 PEP. The views expressed are those of the authors alone.

ADDITIONAL REFERENCES

ERICKSEN, E.P., and KADANE, J.B. (1983). Using the 1980 Census as a population standard. *Proceedings of the Social Statistics Section, American Statistical Association*, 474-479.

WOLTER, K.M. (1987). Comment on Schirm and Preston (1987). *Journal of the American Statistical Association*, 82, 978-980.

COMMENTS

J.A. HARTIGAN[1]

The Adjustment Controversy

Each ten years the United States Census prepares a list, or *enumeration* of names and addresses of persons resident in the United States. The list is subject to error in that persons may be omitted from the list, or erroneously included on it. In the 1990 census, Ericksen *et al.* (1991) estimated there to be 13 million erroneous enumerations and 17 million omissions. Even if these estimates are off by a factor of two, it appears that the enumeration needs some adjustment.

Freedman and Navidi discuss statistical evidence presented in a law suit intended to force the Bureau of the Census to adjust the 1980 enumeration. The origin of the law suit is the *differential undercount* between races. The undercount is perhaps 5% for Blacks and Hispanics, and 1% for Others. Since the undercount is greater for minorities, those localities with larger fractions of minorities press for an adjustment in the census figures that would adjust for the undercount. The undercount has been established by Demographic Analysis (counting births, deaths, emigration, and immigration by race, sex, and age) in censuses since 1940, and by Post Enumeration Surveys, surveys to obtain a more accurate count in a sample of the population, since 1970. The size of the undercount is an important point of dispute, since it makes more sense to estimate and correct for a large differential undercount than for a small one. Freedman and Navidi concede that there may be a differential undercount, but assert, at least for the 1980 census, that the undercount is not sufficiently well estimated in different localities to make adjustments feasible. Freedman and Navidi are concerned mainly to criticize proposed techniques for doing the adjustment; what are their own estimates of the undercount? For example, do they agree that the national undercount is as high as 5% for Blacks and Hispanics compared to 1% for others? I will argue later that if the differential undercount is that high, a synthetic adjustment (each minority person weighted 1.05, each majority person weighted 1.01) will probably improve estimates of state population shares.

In 1980, the Bureau conducted a Post Enumeration Program which it intended to use in adjustment. The bureau decided not to adjust, on the grounds that the PEP estimates were not sufficiently accurate or reliable to give improved counts in small localities. This paper reprises Freedman's testimony in the court case which followed, in which the court's decision supported the Bureau of Census. It may be of interest to report some of the later developments, which show that the issues raised in the present paper are still very much alive. In the 1980's the Bureau planned a more substantial Post Enumeration Survey for the 1990 Census. A dress-rehearsal PES was run in 1988. Some 20 evaluation studies to handle various types of error in the PES were planned and carried out after the 1990 Census. In 1988, the Secretary of Commerce announced that there would be no adjustment of the 1990 census. The government was sued by various localities with high fractions of minorities. The secretary then agreed, on 17 July 1989, to continue planning for the PES, to appoint a committee of 8 experts who would advise the secretary on the feasibility of adjustment, and to publish a set of guidelines under which the Census enumeration would be adjusted or not. The external committee met frequently with Census officials, and advised them on planning , execution and analysis of the PES. The evaluative analyses were carried out by the bureau, and on 21 June, 1991, the steering committee in the census, with some dissent, recommended to the secretary that the census be adjusted.

[1] J.A. Hartigan, Department of Statistics, Yale University, New Haven, CT USA 06520-2179.

The recommendation of the committee was considerably weakened a few days later when an earlier analysis was found to be in error. The external committee divided into two groups of four. The first of these, Ericksen, Estrada, Tukey, and Wolter, with the aid of many consultants, wrote an extensive report that found many defects with the original enumeration, and strongly urged adjustment. The second group of four, Kruskal, McGehee, Tarrance, and Wachter, are as strongly opposed to adjustment. Wachter, with the help of some consultants, offers alternative statistical analysis of the PES that suggest the range of plausible adjustments is so wide as to have quite different effects on reapportionment and other distributive requirements of the Census figures. The secretary decided that the statistical foundation for adjustment was inadequate and recommended against adjustment. The Department of Commerce was sued by the same localities that sued in 1980. The 1980 court case is thus being replayed after the 1990 census.

Synthetic Adjustment

A simple *synthetic* scheme is to multiply each minority person actually enumerated by 1.05 and each majority person actually enumerated by 1.01. I agree with Freedman and Navidi's rejection of Schirm and Preston's (1987) analytic argument.

What about the following analytic argument? Suppose that national undercounts are correctly estimated, but the undercounts differ over states; when does the synthetic adjustment improve the estimate of a state's proportion of the national population? The answer is, if the synthetic adjustment is an overadjustment for a particular state, it is closer to the true proportion than the enumerated proportion if and only if the minority fraction in the state is less than the national minority fraction; conversely, if the adjustment is an underadjustment for a particular state, it is closer to the true proportion than the enumerated proportion if and only if the minority fraction in the state is greater than the national minority fraction. It is plausible to expect the undercount for minorities and non-minorities in high minority states to be higher than in low minority states, which would cause the synthetic adjustment to be an under-adjustment, but nevertheless to be an improvement on the enumerated proportion.

National undercount rates of 5% and 1% are supported by historical evidence from the Bureau, both by demographic analysis and post enumeration surveys, Tables 1 and 2. In a 5-1 adjustment, we multiply non-minority enumerations by 1.01 and minority enumerations by 1.05. Will this improve apportionment of Congressional seats to the various States?

The actual minority and non-minority populations in the different States are unknown. We are comparing the two estimates of the State populations based on the unadjusted and adjusted Census. The census will do best when the States with high minority populations actually have a low differential undercount; the 5-1 adjustment will then overshoot the true proportions for those states. Correspondingly, if the States with low minority populations actually have a high differential undercount, then the 5-1 adjustment will undershoot the true proportions. This tells us how to construct a best case for the census, and a worst case for the adjustment.

I will ignore variations in the non-minority undercount between States, as these should have a minor affect on the overall proportions; I will suppose that all States have a non-minority undercount of 1%. Suppose that the true overall minority undercount is 5%. Suppose this undercount might vary from 3% to 7% in the different States. I assign 3% undercounts to high-minority states, and 7% undercounts to low-minority states, with the division between high and low minority being decided so that the overall minority undercount is 5%. This assumption of true undercounts makes the census look best. The calculation is done for a range of choices of overall undercount and variations across states.

Table 1

Historical Estimates of the Amount and Percent of Net Undercount by Race, as Measured by Demographic Analysis (Report dated 21 June 1991 from the Bureau of Census undercount steering committee)

	1940	1950	1960	1970	1980	1990
Total	5.4	4.1	3.1	2.7	1.2	1.8
Black	8.4	7.5	6.6	6.5	4.5	5.7
Non-black	5.0	3.8	2.7	2.2	0.8	1.3

Table 2

Undercount Estimated by the Post Enumeration Survey and Demographic Analysis in the 1990 Census, by Age, Race and Sex

	Black				Non-black			
	Male		Female		Male		Female	
	PES	DA	PES	DA	PES	DA	PES	DA
	5.4	8.5	4.3	3.0	2.0	2.0	1.4	0.6
0-9	8.0	8.2	7.8	7.8	3.3	2.7	3.4	2.8
10-19	4.0	2.0	4.0	2.2	1.2	−1.0	1.8	−0.5
20-29	6.4	9.4	6.8	3.8	5.0	2.1	3.8	0.9
30-44	5.9	12.4	3.9	2.5	2.2	2.7	1.4	0.1
45-64	3.2	11.7	1.3	0.5	0.4	2.8	−0.5	0.4
65+	1.0	3.0	−0.3	−1.3	−0.9	1.4	−1.1	0.4

The census and adjusted estimates are compared by computing the number of congressional seats that are wrongly apportioned by the two estimates of population. The number of seats allocated to a State with 7.2% of the population is 435 × 7.2 suitably rounded. The rounding makes the actual apportionment a rather poor measuring rod for comparing two methods, because the misapportionment is usually only 1 or 2 seats. Instead, I will use fractional misapportionment, which is half the sum of absolute differences between the estimated and true proportions, multiplied by 435.

It can be seen from Table 3, that the break - even point for census versus adjusted occurs when the true overall minority rate is 3%; we expect this, because then the true differential undercount is 2%, half-way between the 0 rate implied by the census, and the 4% rate assumed by the adjustment. For higher overall minority undercounts, the census does better only when there is a big range of variation across states, and the states with high minority populations happen to have low undercount rates. For example, if the overall rate is 4%, the census achieves 0.8 misapportionment against 1.2 for the adjustment, provided that all the high-minority states have a 2% undercount, and all the low-minority states have a 6% undercount. If the overall rate is 5%, the census achieves 0.8 versus 1.0, only if the high-minority states have a 3% undercount and the low-minority states have a 7% undercount. If the overall rate is greater than 5%, the census is better than the adjusted for no combination of undercount rates in the states having a range of 4% or less.

Table 3

Comparison of Fractional Misapportionment for the Census and a 5-1 Adjustment for a Range of Overall Minority Undercount Rates, with Varying Undercount Rates for the States (The majority undercount rate is fixed at 1%; the minority populations in each state are estimated from the U.S. statistical abstract, 1989)

Overall minority undercount	Minority undercount for low-minority states	Minority undercount for high-minority states	Census fractional misapportionment	5-1 fractional misapportionment
2	2	2	0.2	0.7
2	3	1	0.4	1.1
2	1	3	0.7	0.6
3	3	3	0.5	0.5
3	4	2	0.4	0.9
3	2	4	0.9	0.4
4	4	4	0.7	0.2
4	5	3	0.6	0.7
4	3	5	1.1	0.4
4	6	2	0.8	1.2
4	2	6	1.6	0.9
5	5	5	0.9	0.0
5	6	4	0.8	0.5
5	4	6	1.3	0.5
5	7	3	0.8	1.0
5	3	7	1.8	1.0
6	6	6	1.2	0.2
6	7	5	1.0	0.4
6	5	7	1.6	0.7
6	8	4	1.0	0.9
6	4	8	2.0	1.2
7	7	7	1.4	0.5
7	8	6	1.2	0.4
7	6	8	1.8	0.9
7	9	5	1.2	0.8
7	5	9	2.2	1.4

Table 3 suggests that the overall rates would need to be 3% or less to make this crude 5-1 rule less accurate than the census for apportionment. The 1990 PEP-based 95% 'confidence intervals' for the overall minority rate are 4.3 to 5.7; this range seems overoptimistically narrow, but even if we doubled the quoted margins of error, the interval is 3.6 to 6.4; if the true value lies in this range, then the 5-1 rule will still beat the census.

PEP and PES Based Adjustments

The 1980 PEP survey and the 1990 PES survey are designed to refine a synthetic adjustment by estimating different undercount rates in different localities. Freedman and Navidi are skeptical about the regression used to smooth the estimates, questioning independence, homogeneity of variance, and the reliability of the selection procedures for including variables in the model. I was not persuaded by their examples of how different sets of variables could easily have been selected for inclusion in the model; after all, if the predicting variables are highly correlated, quite different subsets can produce pretty much the same prediction. Thus the fact that different variables were selected does not indicate the smoothed estimates would be very different. Indeed, their table 10 indicates that two variables, percent minority, and percent conventionally enumerated appeared in nearly all equations. I would suspect that the *assumptions* of the regression can not be easily defended, but that the *results* of the regression are reasonable, except perhaps in producing lower standard errors than are justified by the probable lack of independence.

Reduction in sampling variance by regression-based smoothing procedures is not likely to make much difference to estimates in large localities such as States. There, the aggregation of different PEP or PES estimates is already doing as much smoothing as is needed, and the questionable regression assumptions can be avoided. On the other hand, the regression smoothing probably is needed if results are to be projected to small localities.

I agree with Freedman and Navidi that missing data procedures and bias assessment in the post surveys are the key to evaluating the adjusted estimates . Correct handling of missing data, and assessing bias, requires an intimate understanding of survey procedures. Personal judgements by the professionals most closely involved will dominate the conclusions. A healthy skepticism about any resulting 'standard errors' or 'confidence intervals' is justified.

I suggest that the the right loss function to evaluate accuracy in apportionment is not squared error, which is statistically convenient for combining variances and squared biases, nor estimates of the numbers of states or localities that are better estimated by the adjustment. For apportionment, the loss function should be the sum of absolute differences between estimated and true proportions in the different states, because this represents the numbers of people actually misallocated by the estimates, and corresponds at the state level to the number of misapportioned seats.

Although state proportions are of primary interest, let's look at the state populations first. If the true undercount rate in a state is 2%, then the census is better than the estimate just when the estimated undercount rate is less than 0% or greater than 4%. This occurs with probability 50% when the standard error of the estimate is about 3%; thus even a quite inaccurate estimate of undercount is enough to give the adjusted estimate the edge. The census has the same expected difference from truth as the estimate when the standard error is 2.2, and the same expected square difference when the standard error is 2. Out of all this comes the simple rule, that if the true rate is 2%, you do better than the census if you can estimate the true rate with standard error 2%. When estimating population proportions, rather than populations, the relevant computations are on the differential undercounts for the various states, the difference between the undercount for the state and the nationwide undercount, (not the difference between the races); thus adjustments do better than the census in those states where the true difference between the state undercount rate and the nationwide undercount rate exceeds the standard error of the estimated difference.

Under this rule, and accepting the bureau's 1990 estimates of undercount rates and margins of error based on the Post Enumeration Survey, the enumeration is estimated to do better in 24 out of 50 states in estimating proportions. Note however, that the overall estimated loss is quite a bit better for the adjustment than the census, because the states with large (plus or minus) differential undercount rates are estimated better by the adjustment; when the census does better, it does just a little better; when the adjustment does better it often does quite a lot better. Thus the fact that 24 out of 50 states are not estimated to be improved by adjustment should not cause too much excitement; it just means that a lot of states have an estimated undercount rate that is pretty close to the national average, and there is no advantage, and not much difference, in adjusting them. We continue to estimate that substantially more people are correctly allocated by the adjusted figures than the enumeration. Table 4 gives some error estimates when the census PES based figures are incorrect in various ways.

The bureau has produced a number of estimated undercounts, with margins of error, in the various states. I use the 'selected PES method' (called PES from now on) in the report of the Undercount Steering Committee, 21 June 1991. Now there are two popes, the enumeration, and the PES figures. Which is correct? Well, you need a third pope, an infallible one, to decide.

We don't have the third pope. The various follow up evaluations of the PES, the total error model, the loss function analyses, the robustness analyses, are all attempts to feel out what the third pope might decide, but an attitude of skepticism and caution is necessary in believing the decisions of the fictional third pope. In particular, the bureau's 'true population' estimates are all variations on the PES estimates, accepting the basic accuracy and feasibility of the PES, and so most unlikely to find the PES at fault compared to the enumeration. The PES can only be found inferior by some method that is not so closely linked to it. Demographic analysis is by no means in complete agreement with the PES, and provides only national information, but on the whole, it supports the PES rather than the enumeration.

I have done some sensitivity analyses to evaluate how far the PES estimates and margins of error would have to be in error for the census to look competitive. The calculations use the PES state undercount estimates with various multipliers, the PES margins of error with various multipliers, and assume that the true state figures are sampled from normal distributions with the multiplied PES state undercounts and margins of error. I take the different state truths to be independent, which is surely far from correct. The independence will not seriously affect the individual state proportions though, so the average misapportionment of the census and the PES won't be much affected; the variability of the difference will be underestimated.

The results in Table 4 show that the PES estimates have to be in substantial error before the CENSUS starts to be competitive. Accepting the PES rates and margins of error, the CENSUS misapportions 4 seats, the PES 1. If the PES overcount rates are halved, with the margins of error remaining fixed, then the misapportionment rate for the census is 2.5 seats, and for the PES 1.5 seats, and the census will be better in about 40% of the true cases.

However this analysis is in line with the loss function analyses in that it takes the PES as its starting point.

Table 4

Misapportionment of the Enumeration and PES-adjustment, when the True Figures are in Accordance with the PES-undercount Rates and Margins of Error, with Various Multipliers for the Undercounts and Error Margins (Based on 100 simulated true counts.)

Multiplier for pes undercount in each state	Multiplier for margin of error in each state	Census misapportioned seats	Adjusted misapportioned seats	Standard deviation of difference
1	1	3.8	1.1	1.2
1	2	3.7	2.0	1.3
0.5	0.5	2.8	1.2	1.3
0.5	1.0	2.5	1.6	1.4
0.75	0.75	3.3	0.9	1.3
0.75	1.50	3.3	1.6	1.3

I wonder if it might not be useful to make a distinction between the *enumeration*, that lists names and addresses, the *count*, that counts the number of people in various localities according to the list, and *estimates*, using statistical procedures based on various sources of information such as demographics and supplementary lists. This would be perhaps politically divisive, since interested parties would wish to allocate according to the figures most favourable to themselves. There would be the danger that the census professionals, with several estimates available, would be subject to political pressure to choose estimates favourable to one or other group. If we want to estimate populations accurately though, it is a good principle to base the estimates on several mutually supporting surveys rather than a single one. The danger of relying on a single list outweights the dangers and difficulties in combining information from different sources. For one thing, the only way to find out how accurate a survey is to compare it to another survey in one form or another. It should be noted that even in the 'unadjusted' census, population estimates are not simple counts off an enumerated list. Individuals known to be fictitious are included in the count by various kinds of imputation procedures that handle missing data.

Perhaps Freedman and Navidi are right, in asserting that the 1980 PEP figures were too unreliable to permit their use in adjusting the census figures. Perhaps they were right in saying that a nationwide synthetic adjustment is too crude, and has not been demonstrated to improve accuracy. Yet omissions and erroneous enumerations in the tens of millions suggest that some kind of adjustment might improve accuracy. There is plenty of room for improvement. We could misguess the existence or location of a few million people and still be competitive with the raw enumeration.

I have some questions for Freedman and Navidi.

(1) Do they agree with these estimates of 13 million omissions and 17 million erroneous enumerations?
(2) Is the nationwide differential undercount between Blacks and Whites 4%?
(3) Is the PES a useful tool for assessing accuracy of the first census?
(4) Should the PES follow-up sample be used to correct the first census, not only in the specific instances of erroneous enumerations and omissions discovered by comparing the surveys, but also by projecting differential undercounts discovered in the follow up sample to the whole census? If so, how?
(5) If the PES is not good enough, how should the follow-up survey be designed so that it could be used to adjust the census?

COMMENT

T.P. SPEED[1]

Freedman and Navidi ask "Should we have adjusted the census of 1980?" and answer no. I take this as meaning that they have yet to see compelling evidence that it could have been done well, not that they do not see a problem, and not that they think adjustment intrinsically undesirable. My interest in census adjustment was aroused about four years ago, shortly after I came to the U.S. I have read the papers by the main participants in the debate, and have recently had the opportunity to examine some block-level data from the 1990 Post Enumeration Survey. My conclusion is the same as that of Freedman and Navidi: there is simply no evidence to show that adjustment will work at the level proposed.

There are two features of the arguments for adjustments that I find particularly striking. Absolutely no use is ever made of "ground truth" data to demonstrate clearly that adjustments do improve upon the census. And no use is made of available data to justify the key assumptions on which adjustment methods are based.

In 1990 adjustment was to be at the census block level. This would have meant that on the basis of a sample of about 12,000 blocks, each of 6.5 million blocks would have been adjusted. Adjusting a census block count means adding or subtracting people with specific characteristics before further aggregation. This would have been done using procedures based on unverified and implausible assumptions concerning the undercount mechanism. The most important such assumption is that undercount rates are constant within 1,392 demographic subgroups of the population called poststrata, defined by region, race, sex, age and status as a home owner or renter. One such consists of all non-black male Hispanic renters aged 30-44 living in Los Angeles city, or in central cities in the Pacific Census Division (California, Oregon, Washington, Alaska, Hawaii). Another consists of all female owners aged 20-29 living in central cities of 250,000 or more, excluding New York City in the Mid-Atlantic Census Division (New York, New Jersey, Pennsylvania) who are not black, Hispanic, Asian or Pacific Islanders. The parallel with the regression models in the present paper is clear.

Examination of block-level data from the 1990 Post Enumeration Survey from sites in Detroit and Texas showed that the assumption of constancy of the undercount rates within 1,392 poststrata is no better supported than a quite different one: that the undercount is driven by blocks, and is constant across poststrata within blocks. This dual model would have led to different block-level adjustments. The analysis is difficult because the counts of people in the intersections of blocks and poststrata are quite small, heterogeneous, and mostly zero. Details can be found in Hengartner and Speed (1992).

Of course I do not know whether the poststrata-driven or the block-driven undercount model is better; we would need something like "ground-truth" data to answer that. But we can see that certain key assumptions concerning the 1990 undercount model are no better supported by available data than those of a quite different model. In my view, when changing assumptions changes the results, and when we have no way of telling which set of results is closer to the truth, then we have no business adjusting. This is the message I get from the present paper, and it is one I wholeheartedly support.

ADDITIONAL REFERENCE

HENGARTNER, N., and SPEED T.P. (1992). Assessing between-block heterogeneity within poststrata of the 1990 Post-Enumeration Survey. Submitted to *Journal of the American Statistical Association*.

[1] T.P. Speed, Department of Statistics, University of California, Berkeley, CA U.S.A. 94720.

COMMENT

EUGENE P. ERICKSEN and JOSEPH B. KADANE[1]

> "The Court (Judge Sprizzo): I take it your standard error should be a fixed statistical number which you then subtract from your results and you get what is left, basically which is supposed to measure the accuracy of what you are measuring?
>
> The Witness (David Freedman): I hate to argue with you, but it isn't quite like that." (Cuomo *v.* Baldrige: 2629).

We welcome the opportunity to continue the debate with Freedman and Navidi (F and N). Although Judge Sprizzo's decision is now more than 4 years old, the statistical issues are important ones and deserve continued attention. This is especially true because the final scientific judgements are best made by statisticians and demographers, rather than judges and politicians. In this article, Freedman and Navidi review their side of the adjustment controversy, explore some new arguments, and try to use Judge Sprizzo's legal decision to support their scientific position. In this comment, we reexamine certain critical points, restating and clarifying our position where necessary in an effort to demonstrate how adjusting the 1980 Census would have made the data more accurate.

Our disagreements with Freedman and Navidi are fundamental, and we agree that they go to the heart of statistical inference. In their conclusion, F and N write "success of any of EKT's proposed adjustments rides on unverified and implausible assumptions (p. 19)." To the contrary, we believe that our assumptions are realistic and verified by decades of census-taking knowledge, as we will argue below. For their part, F and N's arguments boil down to little more than concern that some assumptions may not be true. To criticize a statistical argument however, it is necessary to do more than that. Assumptions are usually not true exactly - the relevant question is how far they are from being exactly true and what that means for the intended uses of the data. At a minimum, one must show that other assumptions, argued to be just as realistic, or more realistic, lead to substantially different conclusions. F and N do none of this. Moreover, although they concentrate upon the minor differences in various adjustment possibilities, they make no attempt to demonstrate that the adjustments would result in estimates with larger errors than the unadjusted census.

An important part of the disagreement concerns whether or not it is proper to use what we know about the census. F and N give no weight to evidence of greater census-taking problems in some areas than others, and give no credit to the fact that the PEP-measured omission rates and undercounts are higher in those areas with lower mailback rates, higher rates of missing data, and greater problems maintaining the specified long-form sampling rate on the census. Nor do they give any credence to the consistency of the racial differentials in undercount provided by demographic analysis for every census since 1940. This information is not relevant to them, and they are quick to criticize us whenever we rely upon "unverified" assumptions, no matter how realistic or warranted. They also do not explain what "verification" is to them.

At the same time, Freedman and Navidi were not able to make their own argument without reliance upon assertions which are either unverified or are based on the very PEP data which they criticize us for using. Here are some examples:

[1] Eugene P. Ericksen, Temple University, Philadelphia, PA. USA. and Joseph B. Kadane, Carnegie-Mellon University, Pittsburgh, PA USA 15213.

1. A small undercount is thought to remain in the census (p. 3).
2. The census also had a small amount of erroneous enumerations (p. 4).
3. The undercounts estimated by PEP are likely to be biased upward (p. 12).
4. The eruption of Mt. St. Helens caused correlated error between the original enumeration and the PEP (p. 12).
5. Missing data caused a bias in the PEP (p. 13).
6. Minority persons living in central cities are likely to behave differently from those in suburbs (p. 16).
7. The undercount in conventional areas was relatively high (p. 18).

F and N seem to believe that in the absence of substantial direct information about the quality of the PEP data that we should not adjust since different assumptions sometimes lead to different results. This argument, however, ignores the well documented errors in the census enumeration. We have provided extensive documentation, not just in the EKT article, but elsewhere (Ericksen 1983; Ericksen and Kadane 1985) of problems in the census, and others (Citro and Cohen 1985; U.S. Bureau of the Census 1985, 1986 and 1988) have found similar results. To us, the substantial evidence of census-taking problems, the geographic coincidence of census-taking problems with high PEP undercounts, and the consistency of the PEP series with each other and with the results of demographic analysis results provide ample assurance that the additional information derived from the PEP data could have been used to adjust - and improve the accuracy of - the 1980 Decennial Census. This summarizes our general point of view. In the sections that follow, we address some of Freedman and Navidi's specific arguments.

Do the Simple Adjustments Improve Upon the Census?

In their Section 2, F and N criticize our Table 5, in which we claim to show the general agreement of 14 different adjustment schemes. Each of them shifts population share from predominantly White areas outside of cities, where census-taking problems were low, to large central cities with substantial minority populations, where census-taking problems were great. F and N conclude: "The table does not show that any of the methods improve upon the census. It cannot, because there is no external standard against which to measure improvement" (p. 6). If what F and N mean is that the "true" population is unknowable, than their argument, of course, goes too far and no adjustment could ever meet their requirements.

In the EKT paper, we relied upon Schirm and Preston (1987) to show that a simple synthetic method (our Synthetic B) improved upon the census. Since they are also commenting upon Freedman and Navidi's article, we will not repeat their arguments. Given the improvement provided by Synthetic B, we would expect further improvement to be provided by more realistic assumptions, namely that minority populations would be more difficult to count in areas where census-taking problems are greater. These assumptions are consistent not only with PEP results, but with the result of a separate Census Bureau study of New York City which showed that omission rates were strongly and negatively correlated across district offices with mailback rates (Ericksen and Kadane 1986).

Freedman and Navidi base their argument on the apparent differences in the adjusted distributions provided by the different PEP series. We do not believe this evidence to be pertinent, since we know that the eight "preferred PEP's," as well as the more reasonable Synthetic A, will be different not only from Synthetic B, but also from the four less preferred PEP's. Among the six preferred PEP's based on April data, the average rms difference is 0.07%. Differences between these and the two preferred August PEP's are larger, but we explained in our paper why we thought the April and August data were different. More importantly, all the 14 adjustments improve upon the census by shifting population shares from areas where census-taking problems were low to areas where they were high. The fact that some of the adjustments, *e.g.* Synthetic B, make only small adjustments is no argument against making any adjustment.

F and N raise some additional questions, each of which we easily dispose of. First, Freedman and Navidi appear to disagree with our strategy of incorporating information from several sources. However, there was nothing about the sources of information that made combining them inconsistent or unusual. Since we start from the proposition that additional information is generally a useful thing, we do not find any merit in F and N's criticism on this point. Second, finding that the demographic method does not give a decomposition of the undercount that is geographically detailed, they set it aside as if it had no use. For us, the demographic method gives at least two important pieces of information. It gives a reliable estimate of the national undercount, and it also gives a powerful covariate: Blacks are undercounted more than Whites. Neither of these estimates should be taken to be without error, but they certainly give us confidence that each of the preferred PEP series coincide with these observations.

Finally, they found irrelevant our Table 6, which showed that omission rates, relative to rates of erroneous enumeration, were high in those areas with high undercounts. Turning to EKT's Table 5, we find it to be consistent not only with Table 6, but with the results of demographic analysis. This increases our confidence in the utility of the PEP. The argument is called "convergent validity," and is commonly made in the social sciences. It should also be noted that the series we do not take seriously because of the implausibility of their assumptions, Series 10-8, 14-8, 14-9, and 14-20, are less coincident with demographic analysis. We find:

1. Validation of the 8 preferred series, because their national undercount rates and the results in areas in which Blacks are concentrated are consistent with the demographic results, and
2. Evidence that Series 10-8, Freedman and Navidi's foil, is indeed an outlying series.

Can We Expect Improvements in Small Areas?

In our court testimony, we were concerned mainly to show that improvements could be expected for the 66 areas defined as PEP sampling areas. In a separate document, Tukey (1983) showed that if improvement was to be obtained in larger areas, then it could also be expected on average in its smaller components. Since then, both conceptual advances and empirical verifications (Ericksen *et al.* 1991, Appendix H; Wolter and Causey 1991) have been obtained.

Averaging and Sensitivity Analysis

Freedman and Navidi assert that "it is the spread in the PEP series that is interesting, not the average - because it is the spread that demonstrates the impact of applying different modeling assumptions to the same data (p. 11)." We differ from F and N in two ways. First, we believe that **both** the spread and the average are relevant, and we discussed each. Second, and more importantly, we used a different measure of the spread, the root mean squared error (rmse) instead of the range. F and N give little argument to support their choice of statistic. We prefer the rmse because it takes all the data into account, and the squared error feature gives extra weight to large errors. We found that "The root mean squared error among all 792 residuals is 0.59. In contrast, the root mean square of the 66 area effects is 1.60. The area effect is more than double the root mean square residual 47 of 66 times (EKT, p. 938)." We also showed that when we restricted attention to the "preferred eight," that the root mean squared residual was 0.33, and that the area effect was more than double the rmse 59 of 66 times.

We believe that F and N's use of the range in Table 6 and Figure 1 is wrong for another reason. Even among the preferred April estimates there is some difference among the national rates of undercount. If, as F and N say, we are concerned with shifts in shares of population, we should be concerned with deviations from the national average, as in our Tables 10 and 11. For example, in Florida the 2-20 estimate is 2.63% and for 3-8 it is 1.42%, for a range of 1.21%. Subtracting the national rates of 1.9 and 1.0 percent, the respective deviations are 0.73 and 0.42 percent, for a range of only 0.31%. Use of this statistic weakens the correlation displayed by Freedman and Navidi.

Assumptions

Freedman and Navidi argue that some of the assumptions underlying our regression model are "unverified" and "implausible." As we have already argued, both in the EKT paper and elsewhere (Ericksen 1986; Kadane 1986) we believe that they are both realistic and based on a body of knowledge that has been collected for decades. F and N assert that our model improves upon the synthetic estimates only if it uses additional information in a sensible way, bringing us right back to assumptions. We believe, despite F and N's assertions, that our assumptions are surely sensible, and indeed more realistic than the assumptions underlying a decision not to adjust.

At the same time, we believe that it is possible to make too much of the role of modeling in undercount estimation. For small areas, some type of modeling is surely needed. For the 66 areas our article was concerned about, the modeling did not usually make a lot of difference. For example, if we compare the mean residuals in our Table 11, which average residuals from the "preferred eight" estimates, with the corresponding mean residual of the eight sample estimates we find the following. For the 50 states, 46 of the residuals are within one percent of each other, and 48 are within one and one-half percent. The two remaining states, South Carolina and Tennessee, as we explained in EKT, appear to have sample estimates that are wrong, and the use of the regression model seems to provide a clear improvement. Turning to the 16 cities, five of the differences are in fact greater than two percent. For these, the sample sizes were smaller, and the weighted average is much closer to the regression estimate than to the original sample estimate. Although F and N would prefer us not to calculate the weighted average, we prefer to let the sample data play some role, perhaps small, to account for factors not necessarily included in the regression model. Either way, although we hold to our claim of their sensibility, we believe that the argument should be focused more on the quality of the PEP data than on the assumptions of our estimation model.

Does It Matter Which PEP Series is Used?

F and N hold to their position that there is no good reason to choose one PEP series over another. On the contrary, while it may be difficult to select a series from among the "preferred 8," there is good reason not to include Series 10-8 in this group. It is no solution simply to drop the movers from the analysis, as was done for Series 10-8, just because the August CPS had a problem identifying the April address of movers. As F and N themselves recognize, and as we learned from the PEP, movers had higher rates of omission and undercount. The problem with Series 10-8 is indicated in two additional ways. First, its national undercount rate, 0.3%, is well below the 1.4% estimated by demographic analysis. Second, the between-area variability is unrealistically too small, as we show in Table 5. The shift in shares created by Series 10-8 is similar to that of Schirm and Preston's Synthetic B which, while it improved over no adjustment, clearly did not go far enough. As a result, the between-area variability among the 10-8 estimates for our 66 areas is too low. For example, assigning equal weights to each of the 66 areas as Freedman and Navidi appear to have done, the between-area variance for the Series 2-9 estimates is more than twice the corresponding between-area variance for the Series 10-8 estimates. Relative to the national average, the Series 10-8 estimates are too low in the high undercount areas and too high in the low undercount areas. It is little wonder, then, that the residuals from regression are a little bit smaller for Series 10-8 than for 2-9, and F and N's Table 7 has no real meaning.

Which Explanatory Variables Should Be Used

Freedman and Navidi believe that when Series 2-9 was the dependent variable in regression, we misapplied our own rules to select the independent variables. They argue that we should have added the percent living in poverty to our three selections - the minority percentage, the crime rate, and the percent conventional. This is because all four predictors in their equation (11) have coefficients which are more than twice their standard errors, and this equation has a smaller rms residual. They go on to assert that because the coefficient for the poverty variable was negative, we rejected the equation that use of our statistical criteria would otherwise obligate us to select. In other words, they assert that we let our subjective preconceptions overrule our statistical sense.

The problem with F and N's criticism is that they did not replicate our selection procedure correctly. As we explained in the article, and elsewhere (Ericksen and Kadane 1985; 1987, Section 6), "Our estimate of the undercount rate is a matrix-weighted average of a regression estimate and the initial sample estimates (EKT, p. 935)." The observations were weighted by the inverses of the standard errors of the initial sample estimates. This matters, because some states, like South Carolina, had aberrant sample estimates and large variances, and the sample sizes for the 16 cities were also smaller, causing the sample estimates to be less precise. Weighting the data by this procedure, for example, reduced the proportion of total weights assigned to the cities from 24 to 12 percent. When the poverty variable was added to our chosen three in a weighted regression, its coefficient was less than twice its standard error, and it was therefore excluded.

Freedman and Navidi also mistake a statistical decision for substantive motivation. On the contrary, had the poverty variable, with its negative coefficient, satisfied our statistical criteria, it would have added interesting and useful information to our estimates. In general, there are two types of areas with high rates of poverty, central cities with substantial minority populations and rural areas in states like Kentucky and West Virginia with small minority populations. Census errors are more likely to occur in either type of area than elsewhere, but the nature of the errors differ. In the cities, omission rates, as Table 6 in EKT demonstrates, were high, but in the rural areas, the rates of erroneous enumeration were high.

The effects of adding the poverty variable can be seen by subtracting F and N's equation (11) from equation (12), providing the following:

$$\text{difference in 2-9 fit} = 2.23 + .041 \text{ min} - .010 \text{ crime} + .001 \text{ conv} - .176 \text{ pov.}$$

In areas where the percents minority and living in poverty are both high, or both low, the difference may not be great. In areas with many minorities, but perhaps a slightly higher than average rate of poverty, the difference may be positive, but in areas with few minorities, but a high rate of poverty, the difference is negative. Of the 66 areas, the difference obtained from the above equation exceeded one percent only four times, and fell between 0.8 and one percent an additional six times. The ten most extreme areas are:

Area	Equation 12	Equation 11	Difference
		percent	
Maryland R	2.3	1.2	1.1
Houston	5.3	4.2	1.1
Washington, DC	8.1	7.2	0.9
Cleveland	4.3	5.1	−0.8
Arkansas	−0.3	0.5	−0.8
Mississippi	1.0	1.8	−0.8
South Dakota	0.4	1.3	−0.9
Kentucky	−1.3	−0.4	−0.9
Saint Louis	5.5	6.6	−1.1
Boston	3.4	4.9	−1.5

If we simply apply equation (11) to the 66 areas, with no averaging with the initial sample results, the shift in shares is as follows: Group 1, +0.36%; Group 2, +0.20%; Group 3, −0.56%. Substituting equation (12) we get: Group 1, +0.33%; Group 2, +0.21%; Group 3, −0.54%. While the difference between equations (11) and (12) is easily explained, and is consistent with our theory of census error, it really makes little difference to the final results.

Freedman and Navidi also return to the question of whether it was just as reasonable to use the percent urban as the crime rate. As we explained in EKT, use of the crime rate produced a lower rms residual and smaller standard errors than use of the percent urban. In their Tables 7 and 8, F and N appear to get different results. The discrepancy is explained by the same mistake noted above. By using unweighted data, they did not replicate our regression procedure, hence they got different results. Since their strategy gives greater importance to the cities, which had smaller sample sizes and therefore more uncertainty, it is not surprising that the percent urban becomes more important in their criticism.

Perhaps Freedman and Navidi think that our decision to weight the data by their estimated reliability is yet another arbitrary decision. Weighting seems obviously correct to us and is consistent with the strategy the Census Bureau followed in 1990. Where the observation seemed to be more reliable, we gave it greater weight. However, because they did not weight the data, much of F and N's analysis is simply different from ours, and their results in this article are not pertinent to what we did. This applies to their simulation study as well, both in this paper and in Freedman and Navidi (1986). Had they weighted the data, F and N may well have gotten different results. Even so, the fact that the variables selected for regression differ is not the real issue. The real issue is how much the actual estimates obtained from the different regression equations vary. The answer to that, as we have shown above, is that the undercounts do not differ substantially.

Final Comments

Perhaps the main point of the EKT paper is that within the range of reasonable PEP series, for any set of predictor variables that are well correlated with the undercount, results of undercount estimation are similar. In the end, the resulting undercount estimates are rather insensitive to changes either in the predictor variables or the choice of a PEP series. By a similar token, we do not give much weight to F and N's simulation results. The fact that different simulations adding random errors find different "best sets" of predictor variables does not tell us much, unless the distribution of the undercount turns out differently, which it does not.

In the absence of direct evaluation data, we carried out our sensitivity analysis, to see what the effects of various assumptions had on the estimates. In our view, substituting reasonable alternatives for the PEP series and undercount predictors made little difference. Moreover, the results followed a very reasonable pattern in light of the well-documented history of census-taking problems. In those areas where the Census Bureau had greater problems taking the census, the rates of omission, erroneous enumeration, and undercount were higher. In the end, we believe that the substantial and largely unchallenged evidence of serious census-taking errors combined with the consistency of estimates across choices of independent and dependent variables, and the agreement of the pattern of undercount with results of demographic analysis, provides ample reason to adjust.

Freedman and Navidi hold the adjustment data to a higher standard than unadjusted data. They take on faith, and contrary to decades of Census Bureau evidence, that the unadjusted data are accurate, and they do not seem to be concerned with an evident pattern of bias across areas. At the same time, and in the absence of any direct evidence, they assume large biases in the PEP data, when the Census Bureau studies do not demonstrate the existence of such biases (U.S. Bureau of the Census 1988, Section 6F) In other words, they do not seem to place the unadjusted and adjusted data at the same starting point when making their analysis. In doing so, F and N are able to throw out "possible problems" as if they were real ones and to neglect real problems with the unadjusted census as if they did not exist. They reject adjustment on this basis alone.

ADDITIONAL REFERENCES

ERICKSEN, E.P. (1983). Affidavit presented to U.S. District Court, Southern District of New York, in *Cuomo v. Baldrige*, 80 Civ. 4550 (JES).

ERICKSEN, E.P. (1986). Comment on Regression Models for Adjusting the 1980 Census, by D.A. Freedman and W.C. Navidi. *Statistical Science*, 1, 18-21.

ERICKSEN, E.P., and KADANE, J.B. (1986), Using administrative lists to estimate census omissions. *Journal of Official Statistics*, 2, 397-414.

ERICKSEN, E.P., and KADANE, J.B. (1987). Sensitivity analysis of local estimates of undercount in the 1980 U.S. Census. In *Small Area Statistics: An International Symposium*. (Eds. R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh). New York: John Wiley & Sons.

FREEDMAN, D.A. (1984). Testimony given in U.S. District Court, Southern District of New York, in *Cuomo v. Baldrige*, 80 Civ. 4550 (JES).

KADANE, J.B. (1986). Comment on, Regression Models for Adjusting the 1980 Census, by D.A. Freedman and W.C. Navidi, *Statistical Science*, 1, 12-17.

TUKEY, J.W. (1983). Affidavit presented to U.S. District Court, Southern District of New York, in *Cuomo v. Baldrige*, 80 Civ. 4550 (JES).

U.S. BUREAU OF THE CENSUS (1985). The Coverage of Housing in the 1980 Census, 1980 Census of Population and Housing Evaluation and Research Report PHC-80-E1. Washington, DC: U.S. Government Printing Office.

U.S. BUREAU OF THE CENSUS (1987). Programs to Improve Coverage in the 1980 Census, 1980 Census of Population and Housing Evaluation and Research Report PHC-80-E3. Washington, DC: U.S. Government Printing Office.

U.S. BUREAU OF THE CENSUS (1988). The Coverage of Population in the 1980 Census, 1980 Census of Population and Housing Evaluation and Research Report PHC-80-E4. Washington, DC: U.S. Government Printing Office.

RESPONSE FROM THE AUTHORS

1. Introduction

After some general remarks, we respond to each of the discussants' main points. There is some overlap among their arguments; we try to deal with each point only once. Like the other participants, we have learned something over the years – and from the present exchange – but have not changed our opinions on the central questions. One issue cannot be in dispute: Editor M.P. Singh deserves thanks from all sides.

2. A brief Outline of Adjustment

There is a proposal to adjust the census using capture-recapture techniques. A person is "captured" if they are counted in the census; "recapture" is in a special sample survey done after the census. In 1980, this survey was called PEP, or Post Enumeration Program. In 1990, the terminology shifted to PES, or Post Enumeration Survey.

These surveys measure the rate at which people are missed from the census ("gross omissions"), as well as the rate at which people are counted in error ("erroneous enumerations"). Erroneous enumerations include babies born just after census day, people counted at the wrong address, *etc.* To a first approximation, the net undercount is estimated as the difference:

$$\text{gross omissions} - \text{erroneous enumerations.}$$

There is a significant additional complication. In 1980, sampling error was a large-enough problem (according to many observers) so that estimates from the survey could not be used directly. Instead, in EKT's terminology, "sample estimates" from the PEP had to be run through a smoothing model to get "composite estimates." In 1990, the terminology is different: "raw adjustment factors" from the PES are modeled to get "smoothed adjustment factors." But the problem of sampling error is even more salient. For more details, see Freedman (1991), U.S. Department of Commerce (1991a, pp. 4.2-4.18), or Wolter (1991).

3. The Census is Bad so the Alternatives must be Better

Many discussants make an argument which, baldly summarized, comes down to this: the census is bad; the PES must be better; therefore, we should adjust. This is a confusion: it treats the census and the PES as alternatives. However, you cannot choose the survey instead of the census; at most, you can try to use the PES to correct flaws in the census. The question, then, is not whether the survey is better, but whether it is good enough for its intended use.

The Secretary of Commerce framed the issue as follows:

> "I concede the census' imperfections, but the critical inquiry ... is not how flawed the census is, but whether the PES can fix it .... [W]hile identifying flaws in the census is important for planning the next one, it simply begs the question .... Is there convincing evidence showing that the adjustment is more accurate than the enumeration? [U.S. Department of Commerce 1991a, p. 2.13]."

4. Fienberg

Fienberg defines the central issue as follows:

> "At issue is both the accuracy of the census and the adjustment process. And, it is the substantial differential undercount, *i.e.*, the difference between the undercount for Blacks and the undercount for non-Blacks and between Hispanic and non-Hispanic, that is important when we come to assess census accuracy. This is because census figures are typically used to divide resources among groups in the population, resources such as seats in the U.S. House of Representatives; seats in state legislatures; federal funds; and so on. [p. 25, emphasis omitted]."

We think this is misleading. The argument is about shares: more specifically, the accuracy of shares computed from adjusted figures and from the census. But the shares that matter are for geographical areas - states, cities, counties, and so forth. The total share of blacks or hispanics in the U.S. population, at the national level, matters much less. Seats in Congress are allocated to states, and within states to geographical areas. They are not distributed to national racial or ethnic groups. Similarly, tax moneys go to some 39,000 local governments, defined by area not race or ethnicity. The crucial issue is whether adjustment improves the accuracy of population shares for geographical areas rather than groups.

Fienberg is misleading at other points as well. We give two examples.

(i) Fienberg (p. 27). "[Freedman and Navidi] focus on the variation amongst the full set of 12 alternatives, some of which to me are implausible given the assumptions that they rely upon." But we did study variation among EKT's preferred series rather than the full set: see pp. 7-8 and 13-14. We did this not because we agreed with EKT's choices, but to make irrelevant Fienberg's kind of argument. That didn't stop him.

(ii) Fienberg (p. 27). "I read the report by Ylvisaker (1991) who reexamined data from a trial census in Los Angeles in preparation for 1990, but I could not find the evidence that Freedman and Navidi state is supportive of their claim that smoothing increases variability." Ylvisaker did a bootstrap experiment using data from Los Angeles, where there was a test census and a test post enumeration survey in 1986. At the tract level, bootstrap SEs for the smoothed estimates are generally larger than the SEs for the raw estimates. (See Ylvisaker's Table 3; smoothing reduced the SEs in 19/61 tracts, increased the SEs in 26/61, and the remaining 16/61 were ties; at the block level the effects go the other way but are small in either case.) For the whole site, the comparison is as follows (Ylvisaker p. 7):

$$\text{SE for smoothed estimate} = 0.75.$$
$$\text{SE for raw estimate} = 0.68.$$

As we said (p. 17), "smoothing may actually increase sampling error."

Nothing is Perfect, and don't Let the Best be the Enemy of the Good

Fienberg says (p. 27),

> "A familiar theme in various writings by one of the present authors is the problems that arise when assumptions are not satisfied. Here again the authors pursue this theme with respect to the linear equation used for smoothing. They appear to argue that either all assumptions are perfectly justified or 'all bets are off.' Nothing could be further from the truth."

Alas, our position is more complicated than that. We think the census is imperfect, but good. We think the smoothing models are quite questionable, and the arguments to defend them are bad. Proponents of adjustment have an obligation to state their assumptions and produce data to validate them. The models don't have to hold perfectly, but departures from assumptions and their impacts need to be studied. Otherwise, the algorithms have no justification except familiarity.

5. The Burden of Proof

As the exchange with Fienberg indicates, modelers are reluctant to accept the burden of proof. Once they make an assumption, it is taken as truth unless it can be disproved. Even then, they may view the assumption as useful until it can be replaced by some other assumption.

Language is used in a specialized way. An assumption is "reasonable" if the modelers think it is reasonable. If questioned, they introspect again. The introspection confirms the original conclusion; after all, the assumptions are by now familiar parts of the technical literature. The modelers become indignant at those who do not share the faith. If all "reasonable" options favor adjustment, arguments on the other side must be "unreasonable."

So far as they are reported, the modelers' thought experiments do not seem especially rigorous; and the pro-adjustment argument can be peculiarly non-empirical. Illustrations follow. One axiom in the Ericksen-Kadane smoothing model is independence. See equations (1-6) in our paper. Independence drives the variance calculations, because small correlations can have a big cumulative impact. Variances determine whether smoothing is a help or hindrance. The independence assumption matters.

As far as we can see, the adjusters' main arguments for independence are the following:

(i) The errors are not perfectly correlated. (We adapt to present context an argument by Madansky 1986, p. 29.).

(ii) (a) "The 1980 census was administered by more than 400 district offices, an average of eight per state. (b) To our knowledge no one has suggested that there actually was an April snowstorm or any other event that affected the census in neighboring states. (c) When we correlated PEP estimates for cities with the corresponding estimates in their states (*e.g.*, Detroit with the remainder of Michigan), we found no evidence of a correlation." [EKT, p. 931; we responded to (b) by noting the eruption of Mt. St. Helens.]

(iii) "Surely they don't expect anyone to believe the argument that the eruption of Mt. St. Helens interfered with census taking in a serious way ...." [Fienberg p. 27.]

In fact, what we expected from the modelers was serious argument about the validity of assumptions, rather than intuitions about possible sources of dependence like snowstorms and volcanoes. Over time, the force of that expectation has dwindled. Real empirical evidence is hard to get, on both sides. Their mainstay is the rhetoric: Nothing is perfect, so anything goes. That is the adjusters' standard for the models. On the other hand, the census is required to be right to within a few percentage points - where "right" is defined by the models.

6. Fellegi

We agree with many of Fellegi's points. In the U.S., for instance, small-area income data help determine funding allocations. These data have weaknesses of their own, not addressed by census adjustment. Likewise, there are substantial shifts in the population between censuses. Better income data, or a mid-decade census, might be more useful than any adjustments to the decennial census.

There is one point we would like Fellegi to consider. A decision to adjust the census, whether in the U.S. or in Canada, has major organizational costs: it encourages the replacement of data collection by modeling.

> "In sum, real data (with real flaws) would be replaced by complicated and poorly tested mathematical models of data. We do not see that as progress." [Beran *et al.* 1988.]

7. The PEP Series

In 1980, there was a substantial amount of missing data in the surveys used to assess census error. Different ways of filling in the missing data lead to different estimated undercount rates. In the end, the Bureau had a dozen different PEP series: each provides estimated undercount rates for 66 geographic areas (central cities, states apart from their central cities, whole states). A series is identified by a pair of numbers, *e.g.*, PEP 2-9 or PEP 10-8. For more details, and arguments about the merits of the various series, see FN p. 4, EKT p. 929, Fay *et al.* 1988, p. 63.

8. Cressie

Cressie agrees (p. 32) that in 1980, "data and methods were inadequate for an accurate adjustment of the whole country." Of course, many of the arguments are relevant to the 1990 decision, and on those, Cressie's opinion may differ from ours. This is not the place for an extended discussion of 1990, but we can respond to some of his points, at least in outline.

Demographic Analysis

Cressie – like other discussants – relies on estimates from demographic analysis, a technique that uses administrative records (birth certificates, death certificates, *etc.*) to make an independent estimate of the total population. For details, see Fay *et al.* (1988).

How good is demographic analysis? It may be surprising to some, but government statistical agencies keep changing their minds about the past. The estimated GNP for a year in the past – 1985, for example – depends on the year in which the estimate is made. The numbers keep on changing, and the revisions give some clues about the reliability of the initial data.

Table A gives a brief history of revisions to demographic analysis for the 1980 census. As will be seen, the numbers are far from stable. The difference between estimates made in 1984 and 1988 may reflect new understanding about the role of illegal immigration. The change from 1988 to 1991 may reflect the impact of adjustments to earlier adjustments intended to correct for under-registration of births in the period 1935-1960. Apparently, these were over-adjustments, which may now have been fixed.

Table A

A short history of revisions to demographic analysis of the 1980 census: Estimated undercounts by date of estimate

	1984	1988	1991
All races	0.5	1.4	1.2
Blacks	5.3	5.9	4.5
Non-Blacks	−0.2	0.7	0.8
Differential	5.5	5.2	3.7

Source: Col. 1. Cressie, citing Passel and Robinson (1984); the figure for all races is derived.
Col. 2. Fay *et al.* (1988, p. 95, series DA-2).
Col. 3. U.S. Department of Commerce (1991c, Table 3).

Demographers can use data from administrative records to estimate the population of the U.S., and they seem to get it right to within a percentage point or two – a remarkable achievement. However, it seems unlikely that the errors are much less than a percentage point. If so, demographic analysis may not be reliable enough for adjusting the census.

Modeling

Cressie says (p. 33),

> "The important concept to maintain is that true undercount in regions is unknown and the ignorance is quantified into a probability model. The goal is not estimation of the coefficients β but prediction of the undercount. With an error term that does not have to be independent and identically distributed, this prediction is insensitive to misspecification ... [emphasis omitted]."

We disagree. A model for one investigator's ignorance is no basis for public policy. And results must depend strongly on specifications. To illustrate, we note some of the assumptions in the model developed by Cressie (1988). Equations (2.7) and (2.10) in that paper effectively rule out nonsampling error in the PES, as well as systematic variation in undercount rates across geographical areas; and no correlations appear. Why? Equation (2.10) specifies a sampling variance which avoids the internal inconsistencies in the Ericksen-Kadane model (Cressie 1988, p. 193). However, logical consistency does not imply empirical truth. Where does the real sampling design come in? Finally, why should we use Cressie's loss function (2.15)? Until Cressie answers these questions, and others like them, his model outputs have no claim to be taken seriously.

When Cressie gets down to cases, he is computing estimated risks (expected losses). See his equations (2.28-2.31). That means he has to compute variances. Variances are extremely sensitive to assumptions, as Cressie knows:

> Needless to say, these results rely on the correctness of the assumed model. [p. 193].

An elementary illustration may help. Suppose $\epsilon_1, \ldots, \epsilon_{66}$ are exchangeable, with mean 0, variance σ^2, and pairwise correlation ρ. Now

$$\binom{66}{2} = 2145.$$

Therefore,

$$\text{var}[\epsilon_1 + \ldots + \epsilon_{66}] = (66 + 2145\rho)\sigma^2.$$

In this game, a correlation of, say, 0.05 makes a huge difference. And correlations that small would be quite hard to detect empirically. Cressie doesn't try. (With 16 data points, even a correlation of 0.5 might be hard to estimate, so EKT's test #3 on p. 931 cannot have much power.)

The example may seem artificial. However, sampling error was a major obstacle to adjusting the 1990 census on the basis of the PES, even at the state level. Indeed, published data show that for a clear majority of states, the population shares from adjustment would be within two standard errors of the census shares (U.S. Department of Commerce 1991b). Such adjustments could result entirely from sampling error in the PES. ("Loss function analysis" might be the adjusters' response, and we discuss that briefly when answering Hartigan.)

The standard errors, like the estimated adjustments, are outputs from a smoothing model akin to the EK model. Bootstrap experiments reported in Fay (1992) show that these standard errors are too small by a factor of 2 or so. (Fay gives the range 1.4 to 2.2, with a preferred multiplier of 1.7.) When it comes to computing variances, assumptions make all the difference.

PEP 3-8

On p. 32, Cressie agrees that the 1980 adjustment data were not strong enough to use. By p. 33, he wants to adjust using his model and PEP 3-8 (see Section 7 above). He seems to have assumed away all the problems created by non-sampling error, missing correlations, and so forth. If so, his calculations are unrelated to the policy questions.

The Quality of the PES

Cressie says (p. 33) that the PES was "well designed, well implemented, and quality assured." So it was, relative to a typical market research survey, or perhaps even relative to other Census Bureau surveys. However, to fix a small error in the census, you need a sample survey which makes much smaller errors. And we do not believe the PES meets that standard. For example, the PES estimated a national undercount rate of 2.1%. Between 1/3 and 2/3 of that 2.1% can be attributed to non-sampling error in the PES. See Mulry (1991, Table 15) and Bryant (1992). The PES seems to be fatally flawed. We return to this topic in answering Hartigan, below.

Conclusion

Cressie's main point seems to be this (p. 33):

> "To solve a problem as hard as adjustment for undercount, the common goal needs to be recognized. From there, debate should center around differences on how that goal might be reached. If Freedman and Navidi's position is that the goal is impossible to reach (which is what they seem to have implied over the years), then it should be stated."

Let us be clear. In our opinion, PEP could not solve the problem in 1980, and the PES cannot solve the problem in 1990. Nor are we optimistic about the year 2000, whatever acronym may be in use then. If you can't count them, you shouldn't make them up afterwards by running capture-recapture data through smoothing models.

9. Passel (1987)

Many of the discussants defend synthetic adjustment, some very strongly. Few of them are much taken with our counter-example (Table 3). However, Passel (1987) used 1980 census data to show that synthetic adjustment was unlikely to improve accuracy. His work was summarized in the Appendix to our paper. No discussant responds to his argument.

10. Schirm and Preston

The Counter-example

SP (1987, p. 966) make a claim about synthetic adjustment:

> "Our finding is that synthetic adjustment will always move the estimated ratio of a state's population to the national population closer to the true ratio if (a) the state's black undercount is closer to the national black undercount than it is to the national undercount for both races combined and (b) the state's white undercount is closer to the national white undercount than it is to the national undercount for both races combined."

Our counter-example (Table 3) showed this result to be wrong. They should concede the point.

Under some conditions, and by some criteria, synthetic adjustment is doubtless a good thing to do; see, *e.g.*, (15) in our paper. The result in SP's (1987) appendix is correct but not illuminating: the inequality they assume in equation (A.2) on p. 976 is exactly the inequality on absolute error they seek to prove, up to multiplication by a scale factor.

The Simulations

S and P prefer a strict definition of the synthetic assumption – "there is no variation at all" in undercount rates within race across geography. They say (p. 37) that they did not construct true populations on the basis of the synthetic assumption, and their definition of truth did not favor synthetic adjustment.

We adopt their terminology for a moment. They constructed the true populations from the synthetic assumption plus random error. Indeed, the simulations hold the census counts fixed, and randomize the true population. The true population of racial group j in state i is assumed to equal the corresponding census count, multiplied by a random adjustment factor u_{ij}. See SP (1987) equation (1) on p. 967. This adjustment factor is drawn at random from a distribution which depends – by assumption – on the racial group but not the state. See SP (1987) equation (2) on p. 967.

The simulations assume away systematic variation in undercount rates within race across geography. On the other hand, synthetic adjustment assumes that the structure of undercounts is determined by race not geography. That was our point on p. 10, and it is right.

Indeed, S and P concede (p. 37).

> "We considered cases of extreme, albeit nonsystematic, interstate variation in undercounts by race ...."

The "albeit nonsystematic" is their concession; the "extreme" must be the defense.

The *a fortiori* Argument

S and P say their simulations were conservative; the real pattern of variation in undercount rates across areas would favor synthetic adjustment even more strongly than the assumptions they made. (See *e.g.* p. 38). SP (1987) had *a priori* arguments to that effect. Passel (1987) shows, among other things, that such arguments do not prove much about 1980; see the Appendix to our paper. SP (1987, p. 977) make some empirical arguments, using data that are "seriously flawed, based on heroic assumptions"; S and P's language (p. 38), not ours. Further discussion seems unnecessary.

On p. 38, S and P introduce new analysis based on PEP to justify the parameters in the simulations. In present context, that is quite a move: EKT want us to believe the PEP series because they are like the synthetics, while S and P want us to believe the synthetics because simulations are like PEP.

Before we accept either, we want some evidence. Circular reasoning is not persuasive.

11. Hartigan

Synthetic Adjustment

Hartigan rejects Schirm and Preston, but argues strongly in favor of synthetic adjustment (p. 45). He says,

> "What about the following analytic argument? Suppose the national undercounts are correctly estimated, but the undercounts differ over states .... National undercount rates of 5% and 1% are supported by historical data from the Bureau, both by demographic analysis and post enumeration surveys ... I will ignore variations in the non-minority undercount between States ...."

Unless we are much mistaken, these analytic arguments are too far from the facts to be relevant. Hartigan's basic assumption is that the national undercount rates are known. That assumption is wrong: we doubt that the rates can be reliably estimated, either by demographic analysis or the PES, to within a factor of 2. See our discussion of Cressie, above. Furthermore, Hartigan ignores variations in non-minority undercount rates across states. Such variation has to matter: for example, a 1% undercount among 9 million people in a state has almost twice the impact of a 5% undercount among 1 million.

Modeling

Hartigan says

> "I would suspect that the assumptions of the regression can not be easily defended, but that the results of the regression are reasonable, except perhaps in producing lower standard errors that are justified by the probable lack of independence .... Reduction in sampling variance by regression-based smoothing procedures is not likely to make much difference to estimates in large localities such as States .... A healthy skepticism about any resulting 'standard errors' or 'confidence intervals' is justified. [p. 48 emphasis omitted]."

For 1980, the choice of variables makes a lot of difference to the adjustments for small areas. See FN p. 9. For 1990, the "raw" adjustment factors (computed directly from the sample without regression) have such large sampling errors as to be unusable, even at the state level. So the adjusters need to smooth. But the choice of smoothing models makes quite a difference to the results. See the Secretary's Decision (U.S. Department of Commerce 1991a, pp. 2.46-2.55) and consider the numbers in the Press Release (U.S. Department of Commerce 1991b).

Furthermore, the argument for adjusting rides on a "loss function analysis," which uses variances computed from the smoothing model to make unbiased estimates of risk. The model is known to be too optimistic about its variances, perhaps by a factor of 5; see FN p. 10, Ylvisaker (1991), our main paper Section 7.3, and Fay (1992). If "healthy skepticism" is applied to the loss functions, we see no arguments left on the table for the efficacy of proposed adjustments.

We expect to discuss the Bureau's loss function analysis in another paper. Hartigan does his own calculations on p. 49; again, they are too far removed from the data to carry much weight. In any event, readers can look at the Bureau's analysis (Mulry 1991; Woltman *et al.* 1991) before buying any conclusions.

The Third Pope

> "The bureau has produced a number of estimated undercounts, with margins of error, in the various states. I use the 'selected PES method' (called PES from now on) .... Now there are two popes, the enumeration and the PES figures. Which is correct? Well, you need a third pope, an infallible one .... [p. 49]."

Hartigan is on to something important here. The Bureau's "third pope" consists of the loss function analysis discussed above, and a "total error model" (Mulry 1991). These seem highly fallible: the loss function analysis because it depends on variances computed from the smoothing model, and the total error model because it depends on results from the Evaluation Followup to measure non-sampling error. (Furthermore, the two models interact in crucial ways, but that is a topic for another day.)

The adjusters are trying to fix an undercount of maybe 2%. To do that, they need to control non-sampling error in the PES to well below 1%. They say they did it, on the basis of data from yet another sample survey - the Evaluation Followup. If they are measuring non-sampling errors in the PES to within a fraction of 1%, the errors in the Evaluation Followup have got to be an order of magnitude smaller. They must be kidding.

The Five Questions

Hartigan concludes with five questions, and we will answer two. (The first is edited slightly, for clarity.)

(i) "Do Freedman and Navidi agree with these estimates of 17 million omissions and 13 million erroneous enumerations?" We accept the numbers as rough estimates, subject to large and unknown biases as well as large and unknown standard errors. The difference of $17 - 13 = 4$ million may be off by a factor of 2 or more. Estimating a small number by taking the difference of two large numbers is a time-honored recipe for trouble.

Furthermore, a crucial issue is where to put the 4 ± 2 million people. Fienberg doesn't like the foothills of South Dakota. That narrows the options to 6.5 million blocks spread over 39,000 jurisdictions. The PES gave us data on 0.2 of 1% of the blocks, and perhaps 10% of the jurisdictions. Great theater compels the audience to suspend disbelief. Adjustment does not reach that level.

(ii) "If the PES is not good enough, how should the follow-up survey be designed so that it could be used to adjust the census?"The answer is a question of our own: What on earth makes him think it can be done at all?

12. Speed

Adjustment depends on models and assumptions for which there is no empirical proof. That is Speed's message, and we agree.

To adjust the 1990 census, the population is divided into 1,392 "post strata," or demographic groups. One example is post stratum 90302112, male hispanic renters age 10-19 in cities in the Pacific Division. The adjustment depends on the "homogeneity assumption," that undercount rates are more or less constant with each post stratum across geographical areas. See Freedman (1991) or (U.S. Department of Commerce 1991a, pp. 2.37-2.45, pp. 4.16-4.18).

This assumption is hardly an obvious truth. The Bureau did some work to test it (Kim 1991). However, that study seems to have been quite poorly designed, and in any case gives rather mixed results. The theory of adjustment is particularly shaky when it comes to small areas.

13. Ericksen and Kadane

The Role of Assumptions

We say that "success of any of EKT's proposed adjustments rides on unverified and implausible assumptions." EK answer (p. 52) that their "assumptions are realistic and verified by decades of census-taking knowledge, as we will argue below."

The argument they have in mind seems to be on p. 55:

> "Freedman and Navidi argue that some of the assumptions underlying our regression model are 'unverified' and 'implausible.' As we have already argued, both in the EKT paper and elsewhere (Ericksen 1986; Kadane 1986) we believe that they are both realistic and based on a body of knowledge that has been collected for decades."

We reviewed the EKT paper, as well as (Ericksen 1986) and (Kadane 1986). We found no empirical evidence to substantiate the assumptions, or to quantify failures (*e.g.*, to determine the real sizes of the correlations assumed to be 0), or to determine the impact of failures on model output. Instead, EK rely on arguments from convenience (a good model is "simple and tractable" and "permits smoothing," Kadane 1986 p. 13). They also have their own variation on nothing-is-perfect rhetoric:

> ". . . . in applications, only a very naive user would believe in the literal truth of the assumptions. Thus in my view, when I state and use an assumption, I mean that I think something like this is true, but surely I do not mean that exactly this is true . . . . (Kadane 1986, p. 14)."

What makes EK think that "something like" their model is true? Convenience and nothing-is-perfect, even in combination, do not validate assumptions or quantify the impact of failures.

Opening another front, EK tax us with having our own unverified assumptions. Our guilt on this score would hardly imply their innocence; but we deny the charges, or at least most of them. Three examples give the flavor of our "unverified assumptions" (p. 53).

(i) A small undercount is thought to remain in the census.

(ii) Minority persons living in central cities are likely to behave differently from those in suburbs.

(iii) The undercount in conventional areas was relatively high.

Point (i) still seems to be right. If EK will concede that it is wrong, we can all save a lot of courtroom time and journal pages. Point (ii) is obvious to anyone who has spent a few days in a big city in the U.S., but if data are needed, see Freedman *et al.* (1991), which also reviews some literature on this topic. On point (iii), we should have said "estimated undercount." Touché.

Evidence for Assumptions

One example is enough. EK say (p. 52) there is

> "evidence of greater census-taking problems in some areas than others, and . . . PEP-measured omission rates and undercounts are higher in those areas with lower mailback rates, higher rates of missing data, and greater problems maintaining the specified long-form sampling rate on the census."

(For a brief review of the PEP series, see Sections 2 and 7 above.) At most, EK are proving that the PEP data have some relationship to undercount rates, and that we never denied. However, not all relationships can be summarized in a regression model. To get the model going, EK (p. 54) say only "there was nothing about the sources of information that made combining them inconsistent or unusual." This is astonishingly weak, because the tests are only the following: (i) the model should have no internal contradictions; (ii) somebody else should already have done something similar.

Adjusting Small Areas

Earlier, EKT seemed to concede that they could not adjust small areas (p. 943, also see Section 4 of our main article). EK now withdraw the concession (p. 54), citing work by Tukey and Wolter and Causey. That work was reviewed in the Appendix to our paper. We do not find it convincing, and explained why. EK do not respond to our arguments.

EKT's Table 5

EK say that our argument "goes too far." However, their Table 5 is supposed to show that their preferred PEP series are in general agreement with synthetic estimates. Such agreement would demonstrate the value of PEP only if synthetic estimates were known to be accurate. That premise is doubtful, as discussed before.

Furthermore, on the scale EKT chose, we found remarkable disagreement among their preferred PEP series. EK's response: they previously restricted attention to 8 of the 12 PEP series, but now want to eliminate two more (from August). That is not good: among other reasons, the extreme difference noted in our equation (8) occurs with April series making the final cut. Next, EK average across their most preferred series. Averaging results from a sensitivity analysis to reduce variation is a peculiar idea, as discussed in Section 5 of our paper. We return to the point, below.

EK go on to say (pp. 53-54):

> "More importantly, all the 14 adjustments [the 12 PEP series and the two synthetic adjustments] improve upon the census by shifting population shares from areas where census-taking problems were low to areas where they were high."

This is euphemistic. As Table 5 in EKT makes clear (and see EKT p. 927, EK p. 53), the areas where census-taking problems were high are the areas with a high concentration of minority persons. However, as we explained in responding to Fienberg, legislative seats and tax moneys are allocated to geographical areas, not racial or ethnic groups. The key issue is whether adjustment would improve the accuracy of population shares for small geographical areas - states, cities, counties. EKT's Table 5 is about broad groupings of cities and states. Such aggregates seem artificial.

Which PEP Series?

EK try yet one more time (p. 54) to justify their preference for 8 out of the 12 contending PEP series; they particularly seek to eliminate our dreaded foil, series 10-8. The main argument is concordance with demographic analysis at the national level. EK also claim that "the results in areas in which Blacks are concentrated are consistent with the demographic results." This must be a slip in the prose, since demographic analysis gives no results below the national level.

EK indicate (p. 54) that concordance matters, because demographic analysis "gives a reliable estimate of the national undercount." Demographic analysis probably is more reliable than PEP. But it has real problems of its own: see our discussion of Cressie, above. Concordance is a weak argument.

Furthermore, any agreement between PEP and demographic analysis at the aggregate level masks substantial differences in detail, as Jeff Passel showed in court. The arguments have been reviewed before, but we try again. PEP 2-9 is the most preferred of EKT's preferred PEP series; Table B compares PEP 2-9 to demographic analysis. PEP 2-9 is a bit low on black males, 100% too high on black females, 33% too low on white males, and too high on white females (0.5 of 1% *vs.* 0). The agreement has evaporated.

Table B

Comparing two ways of estimating undercount rates in the 1980 census: Demographic analysis (DA) and PEP 2-9.

	DA	PEP 2-9
Black Males	8.8	8.1
Black Females	3.1	6.4
White Males	1.5	1.0
White Females	0.0	0.5

Source: Fay *et al.*, 1988, Appendix D.
Note: Demographic analysis is based on the series DA-2; "white" includes "other races."

Moreover, EK's defense is not totally consistent. For instance, compare p. 54 with p. 55. On p. 54, national undercount rates are important, on p. 55, variation in national undercount rates is unimportant. And their position of the moment – averaged across pages – is inconsistent with that taken by the National Academy of Science Panel, where prominent participants were Steve Fienberg and Jay Kadane:

> "There are a number of reasons, both *a priori* and *a posteriori*, supporting the various individual [PEP series] from this list of 12 .... For example, estimate 10-8 reduces the problem for movers when using the August P-sample .... These points among others are detailed in Bailar['s affidavit in Cuomo *v*. Baldrige]."

> "The use of these 12 estimates produced very different estimates of undercoverage for national demographic groups .... Some analysts have suggested that the number of acceptable estimates should be narrowed considerably. For example, Ericksen ... would discard all but the 2-8, 2-9, 3-8, and 3-9 estimates as either based on August data, which had a higher rate of cases with unresolved match status, or as making use of extreme assumptions in the adjustments for missing data. However, even within this restricted set, the national undercount rate ranges from 0.8 to 1.4 percent. [Cohen and Citro 1985, pp. 147-148.]"

In short, even among EK's most preferred series, different imputation models give different results. Nor is there good reason to discriminate against our foil 10-8.

Likewise, EK say (p. 55) that "Schirm and Preston's Synthetic B ..., while it improved over no adjustment, clearly did not go far enough." This contradicts previous positions taken by the panel, albeit tentatively; see (Cohen and Citro 1985, p. 287; our paper, p. 8).

Averaging and Sensitivity Analysis

EK invite us (p. 54) to replace the various PEP series by the average, and to consider rms deviations from average. However, the point of the 12 different imputation schemes was to measure the impact of modeling. For that purpose, the range is the right statistic: two randomly selected imputation models may give similar results, yet a third may be quite different. In the end, EK want to do a sensitivity analysis, but downplay any model that is different from the other ones.

EK propose again to subtract each series' estimated national undercount rate from its estimates for the 66 study areas. EK are tacitly assuming – with no basis – that one imputation model holds for the whole country. Our analysis takes the view that data may be missing for different reasons in different parts of the country (Section 7.1).

EK give new reasons (p. 55) for rejecting PEP 10-8. The argument comes down to this: if their preferred series are right, our foil 10-8 is wrong. Just so. Conversely, if 10-8 is right, their series are wrong. In other words, it matters which PEP series is used. When you are estimating small undercount rates, 8 percentage points of missing data make a difference. No statistical manipulation can change that awkward fact.

Replication

EK say,

> "The problem with FN's criticism is that they did not replicate our selection procedure correctly. As we explained in the article, and elsewhere (Ericksen and Kadane 1985; 1987, Section 6),
>
> (i) The observations were weighted by the inverses of the standard errors of the initial sample estimates.
>
> (ii) 'Our estimate of the undercount rate is a matrix weighted average of a regression estimate and the initial sample estimates.' [p. 56, order of points interchanged from original.]"

Regrettably, EK are confounding two issues: (i) how you select variables, and (ii) what you do after selecting them. With respect to point (ii), EK are using the Lindley-Smith hierarchical Bayesian regression model. There is only one wrinkle: the parameter σ^2 is unknown and must be estimated; see FN pp. 5 and 11.

Once the variables are selected and σ^2 is estimated, there is no ambiguity about EK's estimator. See Ericksen-Kadane (1985) equation (3), FN equation (9), or the notes to Table 8 in our main paper. Indeed, we were able to replicate their numbers in court; the judge even complimented us on the accuracy of the Berkeley computers. We illustrate the point again, with data in EKT. Their composite estimator based on PEP 2-9 can be extracted from Table 10. The lead example is St Louis, and their value for the estimator is

$$1.24 + 0.66 + 4.16 + 1.09 = 7.13.$$

Our value is 7.12. (The largest discrepancy we found was for Dallas: 6.22 *vs.* 6.18.) Given the variables, we can do the rest.

Most of the discussion in FN, and in the present paper, depends on what you do after selecting the variables, and is immune from EK's criticism of incorrect replication. In particular, despite EK having singled it out by number, our Table 8 is fine. It has nothing to do with the algorithm for variable selection, and we stand by it.

The situation is otherwise with our equations (11-12), Table 7, and Tables 9-10, corresponding to Tables 5 and 6 in FN. Those calculations really do depend on the variable selection algorithm, and we discuss the implications after a few remarks to provide context.

In 1986, EK criticized our simulations, but not on present grounds: the simulations started from the infamous series 10-8. The issue of OLS *vs.* GLS was not mentioned. In 1989, EKT criticized the simulations again, for yet another set of reasons: (i) we restricted attention to models with 3 variables, and (ii) we did not require the coefficients to be significant.

They raise the issue of OLS *vs.* GLS now, for the first time. In response, we redo our calculations once more, using GLS with observations "weighted by the inverses of the standard errors of the initial sample estimates"; coefficients must be significant, but negative values are permitted. We report first on equations (11) and (12); t-statistics are shown in parentheses.

$$\text{(OLS 11)}\quad \text{PEP 2-9} = \underset{(-4.0)}{-2.23} + \underset{(5.4)}{.079}\,\text{min} + \underset{(3.6)}{.036}\,\text{crime} + \underset{(3.5)}{.028}\,\text{conv} + \text{residual}$$

rms residual = 1.53.

$$\text{(OLS 12)}\quad \text{PEP 2-9} = \underset{(7.6)}{.120}\,\text{min} + \underset{(3.4)}{.026}\,\text{crime} + \underset{(3.8)}{.029}\,\text{conv} - \underset{(-4.4)}{.176}\,\text{pov} + \text{residual}$$

rms residual = 1.49.

$$\text{(GLS 11)}\quad \text{PEP 2-9} = \underset{(-6.0)}{-3.37} + \underset{(3.6)}{.054}\,\text{min} + \underset{(5.4)}{.061}\,\text{crime} + \underset{(5.0)}{.026}\,\text{conv} + \text{residual}$$

rms residual = 1.60.

$$\text{(GLS 12)}\quad \text{PEP 2-9} = \underset{(7.3)}{.118}\,\text{min} + \underset{(4.1)}{.030}\,\text{crime} + \underset{(5.2)}{.031}\,\text{conv} - \underset{(-5.4)}{.217}\,\text{pov} + \text{residual}$$

rms residual = 1.53.

Min is the percentage of minorities; crime, the crime rate; conv, the percentage who were conventionally enumerated; pov, the percentage below the poverty line.

As will be seen, the weights make little qualitative difference (although the difference in t-statistics is noticeable). Under either regime, pov is quite significant. And the equation involving pov is superior, for it has smaller residuals.

The poorer an area is, the *less* its undercount will be. That is what equation (12) "shows"; other variables (*i.e.*, racial makeup, crime rate, method of census enumeration) controlled for by the regression. This is in some conflict with EK's theory of the undercount, despite their ingenious argument on p. 56.

The best equation satisfying EK's current criteria is, in fact, equation (GLS 12). It does not have an intercept. If an intercept is required, the best equation is

$$\text{PEP 2-9} = \underset{(2.1)}{1.260} + \underset{(2.9)}{2.609}\,\text{CC} + \underset{(5.1)}{.109}\,\text{min} + \underset{(4.1)}{.0262}\,\text{conv} - \underset{(-3.1)}{.190}\,\text{pov} + \text{residual}$$

rms residual = 1.56.

(CC is an indicator for central cities.) Thus, EK cannot have selected their variables quite the way they say they did.

Again, pov comes in with a significant negative coefficient. Within a central city, there are only two variables: min and pov. The equation says that among minority neighborhoods, the poorer they are, the easier they are to count.

(Equation (2) in EKT is a different GLS regression, with covariance matrix $s^2\mathrm{I} + \mathrm{K}$ rather than K; s^2 is the estimated value of σ^2, and K is the sample-based covariance matrix of the raw undercounts. See equations (1-6) in our paper.)

Our point (p. 15) was that EK could not infer the model from the data; the switch to GLS does not really help them. EK say (p. 57),

> "The real issue is how much the actual estimates obtained from the different regression equations vary. The answer to that, as we have shown above, is that the undercounts do not differ substantially."

That identifies one real issue, out of many. (Another is the impact of variable selection on nominal variances; see Fay 1992). However, if EK are returning to their position of 1986, that they can adjust subareas, then variable selection will matter:

Table C

RMS residuals from regression equations for PEP 2-9 and PEP 10-8.
Explanatory variables include percent minority, percent conventionally enumerated, and either the crime rate or the percent urban.

	Ordinary Least Squares		Generalized Least Squares	
	Crime Rate	Percent Urban	Crime Rate	Percent Urban
PEP 2-9	1.53	1.54	1.60	1.57
PEP 10-8	1.35	1.33	1.39	1.35

"For the 66 areas in the study, the choice of variables has some impact on the adjustments, but not a major one since both sets of variables span essentially the same column space. On the other hand, when extrapolating to subareas, the choice of variables matters a lot. [F and N, p. 9].

We turn next to Table 7, and recompute it using GLS. As Table C shows, for GLS as well as OLS, percent urban is a better variable than the crime rate; and PEP 10-8 is better than PEP 2-9. EK's reasons for excluding 10-8 do not survive inspection.

The simulation in Table 9 comes out very much the same way, whether you select the variables by OLS or GLS. Table D repeats the simulation in Table 10, fitting by GLS. EK are right: urb comes in a little less often, CC noticeably more often. Still, urb beats three of EK's variables (if by a whisker, in the case of MU). Furthermore, negative signs are hardly uncommon in the GLS runs, with paradoxical consequences noted above.

Table D

A simulation experiment on variable selection.
PEP 2-9 is taken as "truth"; percent urban (Urb) is permitted as an explanatory variable.
The table shows the number of times each variable is entered, and the average of its coefficient (over the times it enters); 100 data sets were generated. In both regimes, coefficients must be significant; with GLS, negative values are permitted.

	Ordinary Least Squares		Generalized Least Squares	
Variable	No. of Times Entered	Average Coefficient	No. of Times Entered	Average Coefficient
CC	17	2.954	34	2.922
Min	82	0.071	92	0.084
Crime	53	0.053	40	0.055
Conv	93	0.028	94	0.028
Ed	5	0.085	11	−0.099
Pov	1	0.135	25	−0.212
Lang	17	0.315	5	0.417
MU	0	*****	18	−0.048
Urb	23	0.060	19	0.053

Notes: CC is an indicator for central cities; Min, the percentage of minorities; Crime, the crime rate; Conv, the percentage who were conventionally enumerated; Ed, the percentage with no high school degree; Pov, the percentage below the poverty line; Lang, the percentage who have difficulty with English; MU, the percentage living in multiple-unit housing.

Final Comments

EK say (p. 58),

> "Freedman and Navidi hold the adjustment data to a higher standard than the unadjusted data. They take on faith, and contrary to decades of Census Bureau evidence, that the unadjusted data are accurate, and they do not seem to be concerned with an evident pattern of bias across areas."

That is wrong on all counts. Our article begins with a discussion of errors in the census, their variation across areas, and the resource implications. However, we think that the censuses of 1980 and 1990, with overall accuracy estimated in the range 98% to 99%, were considerable achievements. Management skills have been learned from two centuries of experience, and there was dedicated work by hundreds of thousands of ordinary citizens. These censuses were not perfect, but they were very good of their kind.

Ericksen and Kadane have a novel statistical method which, they say, will improve on the census. Our response is this. Show us. Show us not by the standards of physics on the one hand or ESP research on the other, but by the standards of rational argument. Two court cases and countless journal articles later, we find that Ericksen and Kadane cannot make the argument. But readers will judge for themselves.

ADDITIONAL REFERENCES

BERAN, R., and 12 other statisticians (1988). Statement on census adjustment. U.S. House of Representatives, Subcommittee on Census and Population, hearing of March 3.

BRYANT, B. (1992). Memoranda to Michael Darby and Mark Plant. Reprinted in *Dividing the Dollars*, a report of the U.S. Senate Committee on Government Affairs, S Prt 102-83, Washington, DC, 78-86.

CRESSIE, N. (1987). Comment. *Journal of the American Statistical Association*, 82, 980-983.

FAY, R.E. (1992). Inferences for small domain estimates from the 1990 Post Enumeration Survey. Technical report, Bureau of the Census, Washington, DC.

FREEDMAN, D.A., KLEIN, S.P., SACKS, J., SMYTH C.A., and EVERETT, C.G. (1991). Ecological regression and voting rights. *Evaluation Review* 15, 673-711.

KIM, J. (1991). 1990 PES Evaluation Project P12: Evaluation of Synthetic Assumption. Technical Report, Bureau of the Census, Washington, DC.

MADANSKY, A. (1986). Comment. *Statistical Science*, 1, 28-30.

MULRY, M. (1991). 1990 Post Enumeration Survey Evaluation Project P16. Total Error in PES Estimates for Evaluation Post Strata. Technical Report, U.S. Bureau of the Census, Washington, DC.

U.S. DEPARTMENT OF COMMERCE (1991a). Office of the Secretary. Decision on Whether or Not a Statistical Adjustment of the 1990 Decennial Census of Population Should Be Made for Coverage Deficiencies Resulting in an Overcount or Undercount of the Population; Explanation. Report dated July 15, Washington, DC.

U.S. DEPARTMENT OF COMMERCE (1991b). Press Release CB91-221, dated 6/13/91.

U.S. DEPARTMENT OF COMMERCE (1991c). Press Release CB91-222, dated 6/13/91.

WOLTMAN, H.F., *et al.* (1991). Loss function evaluation. Technical report on project P16, U.S. Bureau of the Census, Washington, DC.

Survey Methodology, June 1992
Vol. 18, No. 1, pp. 75–94
Statistics Canada

REML Estimation in Empirical Bayes Smoothing of Census Undercount

NOEL CRESSIE[1]

ABSTRACT

One way to assess the undercount at subnational levels (*e.g.* the state level) is to obtain sample data from a post-enumeration survey, and then smooth those data based on a linear model of explanatory variables. The relative importance of sampling-error variances to corresponding model-error variances determines the amount of smoothing. Maximum likelihood estimation can lead to oversmoothing, so making the assessment of undercount over-reliant on the linear model. Restricted maximum likelihood (REML) estimators do not suffer from this drawback. Empirical Bayes prediction of undercount based on REML will be presented in this article, and will be compared to maximum likelihood and a method of moments by both simulation and example. Large-sample distributional properties of the REML estimators allow accurate mean squared prediction errors of the REML-based smoothers to be computed.

KEY WORDS: Linear model; Maximum likelihood; Restricted maximum likelihood; Variance components.

1. INTRODUCTION

Although a census attempts to carry out a complete enumeration of the population, for various reasons the final tallies are inaccurate. Census personnel, from its director down to the thousands of temporary enumerators, are part of a mammoth task whose accuracy relies on everyone doing their jobs to perfection.

Moreover, events that are beyond human control (*e.g.* weather, natural disaster) must stay within expected limits. Clearly, in a country the size of the U.S.A. (in terms of both population and geography), many opportunities arise to give an imperfect census count. But size is not the only problem; **heterogeneity** of both population and geography gives a **differentially** imperfect count.

The inaccuracies are typically expressed in terms of undercount, so that a negative value implies an overcount. Suppose the U.S.A. is divided into $i = 1, \ldots, n$ areas (*e.g.* states, including Washington DC). In the i-th area, let T_i be the true (unknown) count and C_i be the census count. Then the undercount, expressed as a percentage of the true count, is defined as,

$$U_i \equiv \{(T_i - C_i)/T_i\}100. \tag{1.1}$$

The problem of **differential** undercount is a serious one when census counts are used to apportion political power and revenue to areas and subareas. (Further discussion of these issues can be found in Ericksen and Kadane 1985, Freedman and Navidi 1986 and Cressie 1988). States like California, Texas, and New York would gain much from adjusting for undercount, *i.e.* from replacing C_i with F_iC_i, where F_i is an **adjustment factor**.

The correct adjustment to use is,

$$F_i = T_i/C_i, \tag{1.2}$$

[1] Noel Cressie, Department of Statistics, Iowa State University, Ames, IA, U.S.A. 50011.

which is related to undercount by,

$$F_i = \{1 - U_i/100\}^{-1}.$$

As it stands, (1.2) is not helpful for adjustment, since the true count T_i is unknown. To obtain extra information that will allow F_i to be estimated, the U. S. Census Bureau conducts a post-enumeration survey (PES) that determines whether people in the PES were or were not counted in the census (*e.g.* Wolter 1986). The survey consists of several hundred thousand households, yielding "raw" adjustment factors $\{Y_i : i = 1, \ldots, n\}$ that are in need of smoothing.

Assume that, given F_i,

$$Y_i \sim \text{Gau}(F_i, \delta_i^2), \tag{1.3}$$

i.e. Y_i has, conditional on F_i, a Gaussian distribution with mean F_i and variance δ_i^2. Adding the further assumption of independence, one obtains,

$$\underset{\sim}{Y} \sim \text{Gau}(\underset{\sim}{F}, \Delta), \tag{1.4}$$

where $\underset{\sim}{Y} \equiv (Y_1, \ldots, Y_n)'$, $\underset{\sim}{F} \equiv (F_1, \ldots, F_n)'$, and Δ is the $n \times n$ diagonal matrix $\text{diag}\{\delta_1^2, \ldots, \delta_n^2\}$.

Now assume that,

$$\underset{\sim}{F} \sim \text{Gau}(X\underset{\sim}{\beta}, \Gamma(\tau^2)), \tag{1.5}$$

where X is an $n \times p$ matrix of explanatory variables, $\underset{\sim}{\beta}$ is a $p \times 1$ vector of (unknown) coefficients of the linear model, $\Gamma(\tau^2)$ is an $n \times n$ diagonal matrix:

$$\Gamma(\tau^2) \equiv \tau^2 D \tag{1.6}$$

and $D \equiv \text{diag}\{1/C_1, \ldots, 1/C_n\}$. The heteroskedastic model (1.5) and (1.6) is discussed at considerable length in Cressie (1990). It is intuitively sensible that the adjustment factor, for an area whose population is large, has a smaller variance; Cressie (1989) provides both a Bayesian and a frequentist justification for this intuition.

Another way to write the model (1.4) and (1.5) is:

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\nu} + \underset{\sim}{\epsilon}, \tag{1.7}$$

where the $n \times 1$ vectors $\underset{\sim}{\nu}$ and $\underset{\sim}{\epsilon}$ are statistically independent, $\underset{\sim}{\nu} \sim \text{Gau}(\underset{\sim}{0}, \Gamma(\tau^2))$, and $\underset{\sim}{\epsilon} \sim \text{Gau}(\underset{\sim}{0}, \Delta)$. Now, assuming that $\delta_1^2, \ldots, \delta_n^2$ are calculated using sampling-variance formulas appropriate for the PES sampling frame, the only parameters left to estimate are $\underset{\sim}{\beta}$ and τ^2. Thus, the two variance components Δ and $\Gamma(\tau^2)$ only contribute one unknown parameter, namely τ^2. It is worth noting that the methods developed in this article can be easily generalized beyond this simple variance-components problem. The general linear model is considered in Section 3.

In Section 2, the Bayes predictor and the empirical Bayes predictor of $\underset{\sim}{F}$ will be given. Estimation of $\underset{\sim}{\beta}$ is straightforward, but there are several possible ways τ^2 could be estimated. Section 3 presents maximum likelihood (m.l.), method-of-moments, and restricted maximum likelihood (REML) approaches. The effect of estimation of τ^2, on mean squared prediction errors, is investigated in Section 4. Section 5 compares the approaches by simulation and by example, and Section 6 presents conclusions and a discussion.

2. EMPIRICAL BAYES PREDICTION

In this article, the true population of any small area is considered to be unknown. After observing the corresponding census population, the uncertainties about the true population are updated. Therefore, statistical models for undercount are **conditional** on the observed census counts. The model (1.4), (1.5), and (1.6) has been introduced in Section 1, and will be assumed throughout Sections 2, 3, and 4.

Using a matrix analogue of squared-error loss, the optimal predictor is $E(\underset{\sim}{F} \mid \underset{\sim}{Y})$ (Cressie 1990), which is,

$$\underset{\sim}{p}^*(\underset{\sim}{Y}) \equiv \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\underset{\sim}{Y} + \{I - \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\}X\underset{\sim}{\beta} \tag{2.1}$$

and the mean-squared-prediction-error matrix is,

$$E\{(\underset{\sim}{F} - \underset{\sim}{p}^*(\underset{\sim}{Y}))(\underset{\sim}{F} - \underset{\sim}{p}^*(\underset{\sim}{Y}))'\} = \{I - \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\}\Gamma(\tau^2). \tag{2.2}$$

For the loss matrix, $L(\underset{\sim}{F},\underset{\sim}{p}) \equiv (\underset{\sim}{F} - \underset{\sim}{p})(\underset{\sim}{F} - \underset{\sim}{p})'$, (2.1) is easily seen to be a **Bayes** predictor of $\underset{\sim}{F}$. In reality, $\underset{\sim}{\beta}$ and τ^2 are unknown and so (2.1) is not a statistic (*i.e.* is not a function only of the data). The proper Bayesian approach would be to put further priors and hyperpriors on all unknown parameters. (This solution to the conundrum of unknown parameters is sometimes called hierarchical Bayes, and demands a prior knowledge of process variability that many scientists do not feel they have. Nevertheless, noninformative priors and hyperpriors, particularly, often yield sensible estimators.) Often the posterior distributions are analytically intractable. Should the model and prior be specified according to their conditional distributions, the Gibbs sampler could be used to obtain, numerically, all required marginal and joint distributions (*e.g.* Gelfand and Smith 1990).

An alternative approach, the one taken in this article, is to treat all parameters, except $\underset{\sim}{F}$, as fixed but unknown, and to use the data $\underset{\sim}{Y}$ to estimate them. This approach is called **empirical Bayes**. Although a parametric (conjugate) prior is assumed in this article, one could also work with a nonparametric prior (*e.g.* Laird and Louis 1987).

Suppose now that $\underset{\sim}{\beta}$ is unknown, but that τ^2 in (1.6) is (for the moment) known. Again, using the matrix analogue of squared-error loss, the optimal linear unbiased predictor is obtained by substituting the generalized-least-squares estimator:

$$\hat{\underset{\sim}{\beta}} \equiv \{X'(\Delta + \Gamma(\tau^2))^{-1}X\}^{-1}X'(\Delta + \Gamma(\tau^2))^{-1}\underset{\sim}{Y}$$

into (2.1), yielding

$$\begin{aligned}\hat{\underset{\sim}{p}}(\underset{\sim}{Y};\tau^2) &= \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\underset{\sim}{Y} + \{I - \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\} \\ &\quad X\{X'(\Delta + \Gamma(\tau^2))^{-1}X\}^{-1}X'(\Delta + \Gamma(\tau^2))^{-1}\underset{\sim}{Y} \equiv \Lambda(\tau^2)\underset{\sim}{Y}\end{aligned} \tag{2.3}$$

(Cressie 1990). The mean-squared-prediction-error matrix is,

$$\begin{aligned}M_1(\tau^2) &\equiv E\{(\underset{\sim}{F} - \hat{\underset{\sim}{p}}(\underset{\sim}{Y};\tau^2))(\underset{\sim}{F} - \hat{\underset{\sim}{p}}(\underset{\sim}{Y};\tau^2))'\} \\ &= \Lambda(\tau^2)\Delta\Lambda(\tau^2)' + (\Lambda(\tau^2) - I)\Gamma(\tau^2)(\Lambda(\tau^2) - I)'.\end{aligned} \tag{2.4}$$

More realistically, τ^2 is also unknown. An **empirical Bayes predictor** is obtained by substituting an estimator $\hat{\tau}^2$ into $\Lambda(\tau^2)$ to yield,

$$\hat{p}(Y; \hat{\tau}^2) = \Lambda(\hat{\tau}^2)Y. \tag{2.5}$$

It is easy to see that when $\hat{\tau}^2$ is the maximum likelihood estimator of τ^2, then (2.5) is the maximum likelihood estimator of the Bayes predictor.

The predictor (2.5) was suggested by Ericksen and Kadane (1985) (and criticized by Freedman and Navidi 1986). Incidentally, the form of their predictors may look different to (2.1), (2.3), and (2.5), but they are in fact identical upon using the identity: $A(A + B)^{-1}B = (A^{-1} + B^{-1})^{-1}$, where A and B are square matrices such that A, B, and $A + B$ have inverses.

By substituting $\hat{\tau}^2$ into (2.4), an estimator of the mean-squared-prediction-error matrix:

$$M_1(\hat{\tau}^2) \equiv \Lambda(\hat{\tau}^2)\Delta\Lambda(\hat{\tau}^2)' + (\Lambda(\hat{\tau}^2) - I)\Gamma(\hat{\tau}^2)(\Lambda(\hat{\tau}^2) - I)' \tag{2.6}$$

is obtained. Since (2.6) does not take into account the estimation of τ^2 in $\hat{p}(Y; \hat{\tau}^2)$, it is likely to be a biased estimator of $E\{(F - \hat{p}(Y; \hat{\tau}^2))(F - \hat{p}(Y; \hat{\tau}^2))'\}$. Further discussion of this important issue is given in Section 4.

Having obtained $\hat{\beta}$ and $\hat{\tau}^2$, model diagnostics can be computed to check the fit of the estimated model. For example, a quantile-quantile plot, of the standardized residuals $(\Delta + \Gamma(\hat{\tau}^2))^{-1/2}(Y - X\hat{\beta})$ against expected order statistics from a unit Gaussian distribution, was used to show no obvious lack of fit of the model used in Section 5. A more complete discussion of model diagnostics is given in Section 6.

3. ESTIMATION OF VARIANCE-MATRIX PARAMETERS

In this section, the general linear model,

$$Y \sim \mathrm{Gau}(X\beta, \Sigma(\gamma)), \tag{3.1}$$

will be assumed, where γ is a $k \times 1$ vector of variance-matrix parameters. In particular, the model given by (1.4), (1.5), and (1.6) yields,

$$\Sigma(\gamma) = \Delta + \Gamma(\tau^2), \tag{3.2}$$

where γ consists of only one parameter, τ^2.

For γ known, estimation of β is straightforward:

$$\hat{\beta}(\gamma) \equiv (X'\Sigma(\gamma)^{-1}X)^{-1}X'\Sigma(\gamma)^{-1}Y. \tag{3.3}$$

More realistically, γ is unknown and has to be estimated; substitution of that estimator into (3.3) then yields an estimated generalized least squares estimator of β. In the rest of this section, three different methods of estimating γ will be considered.

3.1 Maximum Likelihood Estimation

The negative log likelihood of β and γ is:

$$L(\beta, \gamma) = (n/2)\log(2\pi) + (1/2)\log(|\Sigma(\gamma)|) + (1/2)(Y - X\beta)'\Sigma(\gamma)^{-1}(Y - X\beta). \tag{3.4}$$

Minimization of this function yields maximum likelihood (m.l.) estimates $\hat{\beta}_{m\ell}$ and $\hat{\gamma}_{m\ell}$. The difficult part of this minimization involves finding $\hat{\gamma}_{m\ell}$. The Gauss-Newton (scoring) algorithm is given *inter alia* by Harville (1977) and Mardia and Marshall (1984) and is repeated here for notational completeness.

Define,

$$\Sigma_i(\gamma) \equiv \partial\Sigma(\gamma)/\partial\gamma_i;\ i = 1, \ldots, k,$$
$$\Sigma^i(\gamma) \equiv \partial\Sigma^{-1}(\gamma)/\partial\gamma_i = -\Sigma(\gamma)^{-1}\Sigma_i(\gamma)\Sigma(\gamma)^{-1};\ i = 1, \ldots, k, \tag{3.5}$$

the $k \times 1$ vector L_γ to have i-th element:

$$(L_\gamma)_i \equiv (½)\mathrm{tr}(\Sigma(\gamma)^{-1}\Sigma_i(\gamma)) + (½)(Y - X\beta)'\Sigma^i(\gamma)(Y - X\beta), \tag{3.6}$$

and the $k \times k$ matrix J_γ to have (i,j)-th element:

$$(J_\gamma)_{ij} \equiv (½)\mathrm{tr}(\Sigma(\gamma)^{-1}\Sigma_i(\gamma)\Sigma(\gamma)^{-1}\Sigma_j(\gamma)). \tag{3.7}$$

Then,

$$\gamma^{(\ell+1)} = \gamma^{(\ell)} - (J_\gamma^{(\ell)})^{-1}L_\gamma^{(\ell)}, \tag{3.8}$$

where $J_\gamma^{(\ell)}$ and $L_\gamma^{(\ell)}$ denotes J_γ and L_γ, respectively, evaluated at $\gamma = \gamma^{(\ell)}$ and $\beta = \hat{\beta}(\gamma^{(\ell)})$.

When γ consists of only τ^2 in (1.6), the algorithm (3.8) is particularly straightforward. In the simulations and example given in Section 5, the starting value

$$(\tau^2)^{(0)} \equiv \{1/(n-p)\}(Y - X(X'D^{-1}X)^{-1}X'D^{-1}Y)'D^{-1}$$
$$(Y - X(X'D^{-1}X)^{-1}X'D^{-1}Y), \tag{3.9}$$

was used. Then (3.8) is,

$$(\tau^2)^{(\ell+1)} = (\tau^2)^{(\ell)} - \{(½)\sum_{i=1}^{n} 1/(C_i\delta_i^2 + (\tau^2)^{(\ell)})^2\}^{-1}L_\tau^{(\ell)};\ \ell = 0, 1, \ldots, \tag{3.10}$$

where

$$L_\tau^{(\ell)} = (½)\sum_{i=1}^{n} 1/(C_i\delta_i^2 + (\tau^2)^{(\ell)})$$
$$- (½)\{Y - X\hat{\beta}((\tau^2)^{(\ell)})\}'\mathrm{diag}[C_i/(C_i\delta_i^2 + (\tau^2)^{(\ell)})^2]\{Y - X\hat{\beta}((\tau^2)^{(\ell)})\}. \tag{3.11}$$

Iterating (3.8) to convergence yields the m.l. estimator $\hat{\gamma}_{m\ell}$, which upon substitution into (3.3) yields the m.l. estimator $\hat{\beta}(\hat{\gamma}_{m\ell})$. Under appropriate regularity conditions (*e.g.* Mardia and Marshall 1984) $(\hat{\beta}(\hat{\gamma}_{m\ell})', \hat{\gamma}_{m\ell}')'$ is approximately multivariate Gaussian, with mean $(\beta', \gamma')'$ and asymptotic variance matrix,

$$\begin{bmatrix} (X'\Sigma(\underset{\sim}{\gamma})^{-1}X)^{-1} & 0 \\ 0 & J_{\gamma}^{-1} \end{bmatrix}; \tag{3.12}$$

when $\underset{\sim}{\gamma}$ consists of only τ^2 in (1.6), the matrix (3.12) becomes,

$$\begin{bmatrix} (X'\Sigma(\tau^2)^{-1}X)^{-1} & 0 \\ 0 & \left\{(\tfrac{1}{2})\sum_{i=1}^{n} 1/(C_i\delta_i^2+\tau^2)^2\right\}^{-1} \end{bmatrix}. \tag{3.13}$$

In practice, estimated variances and covariances are obtained by evaluating (3.12) at the m.l. estimate $\hat{\underset{\sim}{\gamma}}_{m\ell}$.

3.2 Method-of-Moments Estimation

There is no single method-of-moments estimator of $\underset{\sim}{\gamma}$, but the general idea is to match low-order moments of data with corresponding empirical moments. If only first- and second-order moments are used, it is clear that the Gaussian assumption in (3.1) is not needed.

Let U be a positive-definite symmetric matrix. Consider the weighted regression estimator, $\hat{\underset{\sim}{\beta}}_U \equiv (X'U^{-1}X)^{-1}X'U^{-1}\underset{\sim}{Y}$, and the weighted residuals,

$$\underset{\sim}{e}_U \equiv U^{-1/2}(I - X(X'U^{-1}X)^{-1}X'U^{-1})\underset{\sim}{Y}. \tag{3.14}$$

Then, straightforward matrix algebra shows that,

$$E(\underset{\sim}{e}_U'\underset{\sim}{e}_U) = \mathrm{tr}(\Sigma(\underset{\sim}{\gamma})\Pi_U), \tag{3.15}$$

where $\Pi_U \equiv U^{-1} - U^{-1}X(X'U^{-1}X)^{-1}X'U^{-1}$. Assuming that $\Sigma(\underset{\sim}{\gamma}) = \Delta + \gamma_1\Gamma_1 + \ldots + \gamma_k\Gamma_k$, where Γ_i's are known, one obtains,

$$\sum_{i=1}^{k} \gamma_i \mathrm{tr}(\Gamma_i\Pi_U) = E(\underset{\sim}{e}_U'\underset{\sim}{e}_U) - \mathrm{tr}(\Delta\Pi_U).$$

Choice of k different U_j; $j = 1, \ldots, k$ (*e.g.* $U_1, U_1^2, \ldots, U_1^k$) yields k equations in k unknowns:

$$\sum_{i=1}^{k} \gamma_i \mathrm{tr}(\Gamma_i\Pi_{U_j}) = \underset{\sim}{e}_{U_j}'\underset{\sim}{e}_{U_j} - \mathrm{tr}(\Delta\Pi_{U_j});\ j = 1, \ldots, k, \tag{3.16}$$

which can be solved for $\hat{\gamma}_1, \ldots, \hat{\gamma}_k$. It is important to check that the solution $\hat{\underset{\sim}{\gamma}}$ is in the parameter space $\{\underset{\sim}{\gamma}: \sum_{i=1}^{k} \gamma_i\Gamma_i$ is positive-definite$\}$.

When $\underset{\sim}{\gamma}$ consists of only τ^2 in (1.6), only one matrix U in (3.16) is needed. Previous undercount predictors have based their estimate of τ^2 on $U = I$ (Ericksen and Kadane 1985;

Freedman and Navidi 1986; Ericksen, Kadane and Tukey 1989), but a small sensitivity study for the heteroskedastic model (1.6) suggested a better estimator.

Choose $U_\alpha = \Delta + \Gamma(\alpha)$ in (3.15) to mimic the model (1.7). Then, when $\alpha = \tau^2$ (the true value), Fay and Herriot (1979) show that

$$E(\underset{\sim}{e}'_{U_\alpha}\underset{\sim}{e}_{U_\alpha}) = n - p, \tag{3.17}$$

where n is the number of areas, p is the number of regressors in the matrix X (*e.g.* $p = 3$ for the selected model in Section 5), and $\underset{\sim}{e}_U$ is the standardized residual defined by (3.14). Thus, the proposed method-of-moments estimator of τ^2 is the value of α for which

$$\underset{\sim}{e}'_{U_\alpha}\underset{\sim}{e}_{U_\alpha} = n - p, \tag{3.18}$$

which can be solved using a Newton-Raphson iterative method or a simple bisection method; call the resulting estimator $\hat{\tau}^2_{mm}$.

Fay and Herriot (1979) note that the difference between $\hat{\tau}^2_{mm}$ and $\hat{\tau}^2_{m\ell}$ is manifest in how an area with small δ_i^2 is weighted in the estimation procedure; $\hat{\tau}^2_{m\ell}$ gives relatively more weight to the squared residuals for such an area than does $\hat{\tau}^2_{mm}$. Based on this weighting property, and a small simulation study of bias, Cressie (1990) expressed a preference for $\hat{\tau}^2_{mm}$ over $\hat{\tau}^2_{m\ell}$. However, asymptotically, $\hat{\tau}^2_{m\ell}$ is fully efficient and has an accessible distribution theory. Lack of any (asymptotic) distributional results for $\hat{\tau}^2_{mm}$ causes its own set of problems, such as how to make inference on τ^2, and how to carry out mean-squared- prediction-error corrections in Section 4. A more satisfactory estimator, with better bias properties than the m.l. estimator, is developed below.

3.3 Restricted Maximum Likelihood Estimation

The problem is to find a suitable estimator of the variance-matrix parameters $\underset{\sim}{\gamma}$ in (3.1). The method of restricted maximum likelihood (REML), developed originally by Patterson and Thompson (1971, 1974), applies maximum likelihood to error contrasts rather than to the data themselves. (Rao (1979) calls this method MML, marginal maximum likelihood, in the context of estimation of variance components. Recently, some authors have also called it residual maximum likelihood, although they have retained the abbreviation REML.) A linear combination $\underset{\sim}{a}'\underset{\sim}{Y}$ is called an error contrast if $E(\underset{\sim}{a}'\underset{\sim}{Y}) = 0$, for all $\underset{\sim}{\beta}$ and $\underset{\sim}{\gamma}$; thus, $\underset{\sim}{a}'\underset{\sim}{Y}$ is an error contrast if and only if $\underset{\sim}{a}'X = \underset{\sim}{0}'$.

Let $\underset{\sim}{W} = A'\underset{\sim}{Y}$ represent a vector of $(n - p)$ linearly independent error contrasts; *i.e.* the $(n - p)$ columns of A are linearly independent and $A'X = 0$. Under the Gaussian assumption (3.1), $\underset{\sim}{W} \sim \mathrm{Gau}(\underset{\sim}{0}, A'\Sigma(\underset{\sim}{\gamma})A)$, which does not depend on $\underset{\sim}{\beta}$. Thus, the negative log likelihood function is,

$$L_W(\underset{\sim}{\gamma}) = ((n-p)/2)\log(2\pi) + (½)\log(\,|A'\Sigma(\underset{\sim}{\gamma})A|\,) + (½)\underset{\sim}{W}'(A'\Sigma(\underset{\sim}{\gamma})A)^{-1}\underset{\sim}{W}.$$

If another set of $(n - p)$ linearly independent contrasts were used to define $\underset{\sim}{W}$, the new negative log likelihood function would differ from $L_W(\underset{\sim}{\gamma})$ only by an additive constant (Harville 1974). Indeed, for the A that satisfies $AA' = I - X(X'X)^{-1}X'$ (and $A'A = I$),

$$L_W(\underset{\sim}{\gamma}) = ((n-p)/2)\log(2\pi) - (½)\log(\,|X'X|\,) + (½)\log(\,|\Sigma(\underset{\sim}{\gamma})|\,) + (½)\log(\,|X'\Sigma(\underset{\sim}{\gamma})^{-1}X|\,) + (½)\underset{\sim}{Y}'\Pi(\underset{\sim}{\gamma})\underset{\sim}{Y}, \tag{3.19}$$

where $\Pi(\underset{\sim}{\gamma}) \equiv \Sigma(\underset{\sim}{\gamma})^{-1} - \Sigma(\underset{\sim}{\gamma})^{-1}X(X'\Sigma(\underset{\sim}{\gamma})^{-1}X)^{-1}X'\Sigma(\underset{\sim}{\gamma})^{-1}$; see Harville (1974). A REML estimate of $\underset{\sim}{\gamma}$, denoted $\hat{\underset{\sim}{\gamma}}_{r\ell}$, is obtained by minimizing (3.19) with respect to $\underset{\sim}{\gamma}$. The distinction between REML and m.l. estimation becomes important when p is large relative to n. The REML method was originally proposed to estimate variance-component parameters: Numerical algorithms (Harville 1977), robust adaptations (Fellner 1986), and distribution theory (Cressie and Lahiri 1991) have been developed in this context. Kitanidis (1983) and Zimmerman (1989) give computational details for producing an iterative minimization of (3.19).

Harville (1974) provides a Bayesian justification for REML by assuming a noninformative prior for $\underset{\sim}{\beta}$, which is statistically independent of $\underset{\sim}{\gamma}$, and showing that the marginal posterior density of $\underset{\sim}{\gamma}$ is proportional to (3.19) multiplied by the prior for $\underset{\sim}{\gamma}$. When that prior is noninformative, REML estimates correspond to marginal MAP (maximum *a posteriori*) estimates. Thus, in the situation where noninformative prior distributions for $\underset{\sim}{\beta}$ and $\underset{\sim}{\gamma}$ are independent, REML can be seen as a compromise between m.l. and Bayes estimation with squared error loss. In the case of model (1.4), (1.5) and (1.6), the latter would yield a Bayes estimate, $\int_0^\infty \tau^2 \exp\{-L_W(\tau^2)\}d\tau^2$, which can be obtained equivalently by **averaging** τ^2, weighted by the *full* likelihood, $\exp\{-L(\underset{\sim}{\beta}, \tau^2)\}$. On the other hand, m.l. yields as an estimate of τ^2 the value $\hat{\tau}^2_{m\ell}$ obtained by **maximizing** the **full** likelihood. REML averages the full likelihood over $\underset{\sim}{\beta}$ but maximizes the resulting (restricted) likelihood over τ^2.

Maximum likelihood estimation of τ^2 tends to be biased towards zero because the likelihood, as a function of τ^2, is skewed to the right. When normalized to integrate to one, the mean of such a function is generally larger than its mode (*e.g.* Groeneveld and Meeden 1977). The m.l. estimate is based on the profile of the likelihood surface of $\underset{\sim}{\beta}$ and τ^2, and this favors smaller values of τ^2. (In contrast, REML is obtained by first integrating the likelihood over $\underset{\sim}{\beta}$ and then maximizing the result over τ^2. Notice that Bayesians might advocate further integration over τ^2.)

Although the Bayesian interpretation of REML helps to explain its properties, $\hat{\underset{\sim}{\gamma}}_{r\ell}$ also has the obvious frequentist interpretation of being an estimator based on restricted information.

Minimization of (3.19) with respect to $\underset{\sim}{\gamma}$ can proceed by any of the gradient algorithms. Recall,

$$\underset{\sim}{W} = A'\underset{\sim}{Y} \tag{3.20}$$

and suppose A satisfies:

$$AA' = I - X(X'X)^{-1}X', \text{ and } A'A = I.$$

For the moment, focus all attention on the $(n - p)$ "data" $\underset{\sim}{W}$; their joint distribution depends only on $\underset{\sim}{\gamma}$, and the associated negative log (restricted) likelihood is $L_W(\underset{\sim}{\gamma})$ given by (3.19).

Define the $k \times 1$ vector $\underset{\sim}{M}_\gamma$ to have i-th element:

$$(\underset{\sim}{M}_\gamma)_i \equiv \partial L_W(\underset{\sim}{\gamma})/\partial\gamma_i = (½)\text{tr}\{\Pi(\underset{\sim}{\gamma})\Sigma_i(\underset{\sim}{\gamma})\} - (½)\underset{\sim}{Y}'\Pi(\underset{\sim}{\gamma})\Sigma_i(\underset{\sim}{\gamma})\Pi(\underset{\sim}{\gamma})\underset{\sim}{Y}, \tag{3.21}$$

and the $k \times k$ matrix G_γ to have (i, j)-th element:

$$(G_\gamma)_{ij} \equiv E(\partial^2 L_W(\underset{\sim}{\gamma})/\partial\gamma_i\partial\gamma_j) = (½)\text{tr}\{\Pi(\underset{\sim}{\gamma})\Sigma_i(\underset{\sim}{\gamma})\Pi(\underset{\sim}{\gamma})\Sigma_j(\underset{\sim}{\gamma})\}, \tag{3.22}$$

where $\Pi(\underset{\sim}{\gamma})$ is given below (3.19) and $\Sigma_i(\underset{\sim}{\gamma})$ is defined by (3.5). (The expressions (3.21) and (3.22) were obtained by Harville 1977.) Then, the Gauss-Newton (scoring) algorithm to find $\hat{\underset{\sim}{\gamma}}_{r\ell}$ is:

$$\underset{\sim}{\gamma}^{(\ell+1)} = \underset{\sim}{\gamma}^{(\ell)} - (G_\gamma^{(\ell)})^{-1}\underset{\sim}{M}_\gamma^{(\ell)}, \tag{3.23}$$

where $G_\gamma^{(\ell)}$ and $\underset{\sim}{M}_\gamma^{(\ell)}$ denote G_γ and $\underset{\sim}{M}_\gamma$, respectively, evaluated at $\underset{\sim}{\gamma} = \underset{\sim}{\gamma}^{(\ell)}$.

When $\underset{\sim}{\gamma}$ consists of only τ^2 in (1.6), the algorithm (3.23) is particularly straightforward. In the simulations and example given in Section 5, the starting value (3.9) was used. Then (3.23) is,

$$(\tau^2)^{(\ell+1)} = (\tau^2)^{(\ell)} - (G_\tau^{(\ell)})^{-1}M_\tau^{(\ell)}, \tag{3.24}$$

where

$$M_\tau = (½)\mathrm{tr}\{\Pi(\tau^2)D\} - (½)\underset{\sim}{Y}'\Pi(\tau^2)D\Pi(\tau^2)\underset{\sim}{Y}, \tag{3.25}$$

$$G_\tau = (½)\mathrm{tr}\{\Pi(\tau^2)D\Pi(\tau^2)D\}, \tag{3.26}$$

$$\Pi(\tau^2) = \Sigma(\tau^2)^{-1} - \Sigma(\tau^2)^{-1}X(X'\Sigma(\tau^2)^{-1}X)^{-1}X'\Sigma(\tau^2)^{-1}, \tag{3.27}$$

are evaluated at $\tau^2 = (\tau^2)^{(\ell)}$. Also, recall that $\Sigma(\tau^2) = \Delta + \tau^2 D$ and $D = \mathrm{diag}\{1/C_1, \ldots, 1/C_n\}$.

Iterating (3.23) to convergence yields the REML estimator $\hat{\underset{\sim}{\gamma}}_{r\ell}$. It has been proved by Cressie and Lahiri (1991) that $\hat{\underset{\sim}{\gamma}}_{r\ell}$ is approximately multivariate Gaussian, with mean $\underset{\sim}{\gamma}$ and asymptotic variance matrix,

$$G_\gamma^{-1}. \tag{3.28}$$

When $\underset{\sim}{\gamma}$ consists of only τ^2 in (1.6), the matrix (3.28) becomes a scalar,

$$[(½)\mathrm{tr}\{\Pi(\tau^2)D\Pi(\tau^2)D\}]^{-1}. \tag{3.29}$$

In practice, estimated variances and covariances are obtained by evaluating (3.28) at $\underset{\sim}{\gamma} = \hat{\underset{\sim}{\gamma}}_{r\ell}$. Furthermore, the normalized (estimated) generalized least squares estimator, $\hat{\underset{\sim}{\beta}}(\hat{\underset{\sim}{\gamma}}_{r\ell})$ should be approximately Gaussian with asymptotic variance matrix, $(X'\Sigma(\underset{\sim}{\gamma})X)^{-1}$.

4. IMPROVED ESTIMATION OF MEAN SQUARED PREDICTION ERRORS

In what is to follow, I shall be concerned with the effect, on prediction, of estimation of $\underset{\sim}{\gamma}$ in $\Sigma(\underset{\sim}{\gamma})$ given by (3.1). Generalizing (1.5) to,

$$\underset{\sim}{F} \sim \mathrm{Gau}(X\underset{\sim}{\beta}, \Gamma(\underset{\sim}{\gamma})), \tag{4.1}$$

it is clear that

$$\Sigma(\underset{\sim}{\gamma}) = \Delta + \Gamma(\underset{\sim}{\gamma}). \tag{4.2}$$

In principle, Δ could also depend on unknown parameters (in, *e.g.* a model for sampling variances) and the results of this section are equally applicable. The optimal linear unbiased predictor is,

$$\underset{\sim}{\hat{p}}(\underset{\sim}{Y};\underset{\sim}{\gamma}) = \Gamma(\underset{\sim}{\gamma})(\Delta + \Gamma(\underset{\sim}{\gamma}))^{-1}\underset{\sim}{Y} + \{I - \Gamma(\underset{\sim}{\gamma})(\Delta + \Gamma(\underset{\sim}{\gamma}))^{-1}\}$$

$$X\{X'(\Delta + \Gamma(\underset{\sim}{\gamma}))^{-1}X\}^{-1}X'(\Delta + \Gamma(\underset{\sim}{\gamma}))^{-1}\underset{\sim}{Y} \equiv \Lambda(\underset{\sim}{\gamma})\underset{\sim}{Y}. \tag{4.3}$$

Then, the mean-squared-prediction-error matrix of $\hat{p}(\underset{\sim}{Y};\underset{\sim}{\gamma})$, denoted $M_1(\underset{\sim}{\gamma})$, is given by,

$$M_1(\underset{\sim}{\gamma}) = \Lambda(\underset{\sim}{\gamma})\Delta\Lambda(\underset{\sim}{\gamma})' + (\Lambda(\underset{\sim}{\gamma}) - I)\Gamma(\underset{\sim}{\gamma})(\Lambda(\underset{\sim}{\gamma}) - I)'. \tag{4.4}$$

In reality, $\underset{\sim}{\gamma}$ is unknown and has to be estimated by $\hat{\underset{\sim}{\gamma}}$, say. The empirical Bayes predictor of $\underset{\sim}{F}$ is then $\underset{\sim}{\hat{p}}(\underset{\sim}{Y};\hat{\underset{\sim}{\gamma}})$, given by (4.3) with $\underset{\sim}{\gamma} = \hat{\underset{\sim}{\gamma}}$. In this case, $M_1(\underset{\sim}{\gamma})$ is an inappropriate measure of the predictor's precision; one should use instead,

$$M_2(\underset{\sim}{\gamma}) = E\{(\underset{\sim}{F} - \underset{\sim}{\hat{p}}(\underset{\sim}{Y};\hat{\underset{\sim}{\gamma}}))(\underset{\sim}{F} - \underset{\sim}{\hat{p}}(\underset{\sim}{Y};\hat{\underset{\sim}{\gamma}}))'\}. \tag{4.5}$$

It is the risk matrix (4.5), or an estimate of it, that should be given, along with the predictor $\underset{\sim}{\hat{p}}(\underset{\sim}{Y};\hat{\underset{\sim}{\gamma}})$. However, $M_1(\hat{\underset{\sim}{\gamma}})$ is typically reported; hence, one should ask what inaccuracies result from using $M_1(\hat{\underset{\sim}{\gamma}})$ and whether a more appropriate estimator of $M_2(\underset{\sim}{\gamma})$ is available.

Now, under the assumptions (4.1) and (4.2) (Gaussianity is important here) and provided $\hat{\underset{\sim}{\gamma}}$ is an even and translation invariant function of the data, the results of Harville (1985) can be used to establish that $M_2(\underset{\sim}{\gamma}) - M_1(\underset{\sim}{\gamma})$ is non-negative-definite. (An estimator is even if $\hat{\underset{\sim}{\gamma}}(\underset{\sim}{Y}) = \hat{\underset{\sim}{\gamma}}(-\underset{\sim}{Y})$ and is translation invariant if $\hat{\underset{\sim}{\gamma}}(\underset{\sim}{Y} + X\underset{\sim}{\lambda}) = \hat{\underset{\sim}{\gamma}}(\underset{\sim}{Y})$, for any $p \times 1$ vector $\underset{\sim}{\lambda}$.) When $\underset{\sim}{\gamma}$ consists of only τ^2 in (1.6), the estimators $\hat{\tau}^2_{ml}$, $\hat{\tau}^2_{mm}$ and $\hat{\tau}^2_{rl}$ are all even and translation invariant. Intuitively, estimation of the unknown parameters $\underset{\sim}{\gamma}$ leads to larger mean squared prediction errors; the result above quantifies this intuition.

But, there is another potential source of bias due to the fact that $M_1(\hat{\underset{\sim}{\gamma}})$, not $M_1(\underset{\sim}{\gamma})$, is used to estimate the risk matrix. Suppose that $\hat{\underset{\sim}{\gamma}}$ is chosen to yield an unbiased estimator of the variance matrix of $(\underset{\sim}{Y}', \underset{\sim}{F}')'$, which most would agree is a desirable property. Then the results of Eaton (1985) and Zimmerman and Cressie (1991) can be used to establish that $M_1(\underset{\sim}{\gamma}) - E(M_1(\hat{\underset{\sim}{\gamma}}))$ is non-negative-definite. (The proof relies on a multivariate version of Jensen's inequality and on the fact that $\underset{\sim}{\hat{p}}(\underset{\sim}{Y};\underset{\sim}{\gamma})$, which can be written as $\Lambda(\underset{\sim}{\gamma})\underset{\sim}{Y}$, minimizes the risk matrix over all linear unbiased predictors.)

Upon writing,

$$M_2(\underset{\sim}{\gamma}) - M_1(\hat{\underset{\sim}{\gamma}}) = \{M_2(\underset{\sim}{\gamma}) - M_1(\underset{\sim}{\gamma})\} + \{M_1(\underset{\sim}{\gamma}) - E(M_1(\hat{\underset{\sim}{\gamma}}))\} +$$

$$\{E(M_1(\hat{\underset{\sim}{\gamma}})) - M_1(\hat{\underset{\sim}{\gamma}})\}, \tag{4.6}$$

the results above establish that underestimation of $M_2(\underset{\sim}{\gamma})$ comes from two sources. Even if an expression for $M_2(\underset{\sim}{\gamma})$ were known, it is likely that $M_2(\hat{\underset{\sim}{\gamma}})$ would be biased for $M_2(\underset{\sim}{\gamma})$, further illustrating the inherent difficulty in estimating mean squared prediction errors.

A remedy has been suggested by Prasad and Rao (1990), based on asymptotic expansions of $M_2(\underset{\sim}{\gamma})$. Consider prediction of undercount in the i-th area, and let $[M_2(\underset{\sim}{\gamma})]_{ii}$ and $[M_1(\hat{\underset{\sim}{\gamma}})]_{ii}$ denote the (i,i)-th elements of the risk matrices $M_2(\underset{\sim}{\gamma})$ and $M_1(\hat{\underset{\sim}{\gamma}})$, respectively. Then formal application of Prasad and Rao's proposal yields the estimator of $[M_2(\underset{\sim}{\gamma})]_{ii}$,

$$[M_2(\underset{\sim}{\gamma})]^*_{ii} \equiv [M_1(\hat{\underset{\sim}{\gamma}})]_{ii} + 2\mathrm{tr}\{A_{ii}(\hat{\underset{\sim}{\gamma}})B(\hat{\underset{\sim}{\gamma}})\};\ i = 1, \ldots, n. \tag{4.7}$$

In (4.7), $A_{ii}(\underset{\sim}{\gamma})$ is a $k \times k$ matrix given by,

$$A_{ii}(\underset{\sim}{\gamma}) = \mathrm{var}\{\partial \hat{p}_i(\underset{\sim}{Y}; \underset{\sim}{\gamma})/\partial \underset{\sim}{\gamma}\} \tag{4.8}$$

and $B(\underset{\sim}{\gamma})$ is a matrix that equals or approximates the $k \times k$ matrix,

$$E\{(\hat{\underset{\sim}{\gamma}} - \underset{\sim}{\gamma})(\hat{\underset{\sim}{\gamma}} - \underset{\sim}{\gamma})'\}. \tag{4.9}$$

For m.l. estimation,

$$B(\underset{\sim}{\gamma}) = J_\gamma^{-1}, \tag{4.10}$$

where J_γ is given by (3.7), and for REML estimation,

$$B(\underset{\sim}{\gamma}) = G_\gamma^{-1}, \tag{4.11}$$

where G_γ is given by (3.22).

Kass and Steffey (1989) give approximations (to the conditional variance) that are similar in spirit to (4.7), for probability distributions that are not necessarily Gaussian. However, their approach requires independent replications, which is not a feature of the distributions specified by (3.1).

Should small areas be aggregated, it is important to have an approximately unbiased estimator of all elements of $M_2(\underset{\sim}{\gamma})$. It is not difficult to generalize (4.7) to,

$$[M_2(\underset{\sim}{\gamma})]^*_{ij} = [M_1(\hat{\underset{\sim}{\gamma}})]_{ij} + 2\mathrm{tr}\{A_{ij}(\hat{\underset{\sim}{\gamma}})B(\hat{\underset{\sim}{\gamma}})\};\ i, j = 1, \ldots, n,$$

where $A_{ij}(\underset{\sim}{\gamma}) \equiv \mathrm{cov}\{\partial\hat{p}_i(\underset{\sim}{Y}; \underset{\sim}{\gamma})/\partial\underset{\sim}{\gamma}, \partial\hat{p}_j(\underset{\sim}{Y}; \underset{\sim}{\gamma})/\partial\underset{\sim}{\gamma}\}$. Prasad and Rao (1990) show that, to the same order of magnitude, $A_{ij}(\underset{\sim}{\gamma})$ can be replaced by $\mathrm{cov}\{\partial p_i^*(\underset{\sim}{Y})/\partial\underset{\sim}{\gamma}, \partial p_j^*(\underset{\sim}{Y})/\partial\underset{\sim}{\gamma}\}$, where $\underset{\sim}{p}^*(\underset{\sim}{Y})$ is given by (2.1); these latter derivatives can be simpler to calculate.

When $\underset{\sim}{\gamma}$ consists of only τ^2 in (1.6), calculation of $B(\underset{\sim}{\gamma})$ is straightforward; see (3.13) and (3.26). Now, consider

$$\mathrm{var}(\partial\underset{\sim}{\hat{p}}(\underset{\sim}{Y};\tau^2)/\partial\tau^2) = (\partial\Lambda(\tau^2)/\partial\tau^2)\Sigma(\tau^2)(\partial\Lambda(\tau^2)/\partial\tau^2)', \tag{4.12}$$

where $\Lambda(\tau^2)$ is given by (2.3). In terms of $\Pi(\tau^2)$ defined by (3.27), and Δ defined by (1.4),

$$\Lambda(\tau^2) = I - \Delta\Pi(\tau^2). \tag{4.13}$$

Thus, (4.12) can be calculated from (4.13) using the relationships (3.4) and (3.5). Then, $A_{ii}(\tau^2)$ given by (4.8) is the (i,i)-th element of,

$$\Delta(\partial\Pi(\tau^2)/\partial\tau^2)\Sigma(\tau^2)(\partial\Pi(\tau^2)/\partial\tau^2)'\Delta', \tag{4.14}$$

where

$$\begin{aligned}\partial\Pi(\tau^2)/\partial\tau^2 = &- \Sigma(\tau^2)D\Sigma(\tau^2)\{I - X(X'\Sigma(\tau^2)^{-1}X)^{-1}X'\Sigma(\tau^2)^{-1}\} - \\ &\Sigma(\tau^2)^{-1}X(X'\Sigma(\tau^2)^{-1}X)^{-1}\{X'\Sigma(\tau^2)^{-1}D\Sigma(\tau^2)^{-1}X\} \\ &(X'\Sigma(\tau^2)^{-1}X)^{-1}X'\Sigma(\tau^2)^{-1} + \Sigma(\tau^2)^{-1}X(X'\Sigma(\tau^2)^{-1}X)^{-1} \\ &X'\Sigma(\tau^2)^{-1}D\Sigma(\tau^2)^{-1};\end{aligned} \tag{4.15}$$

recall that $\Sigma(\tau^2) = \Delta + \tau^2 D$, and $D = \mathrm{diag}\{1/C_1, \ldots, 1/C_n\}$.

The estimator of mean squared prediction error, $[M_2(\tau^2)]^*_{ii}$, is conjectured to be approximately unbiased (Prasad and Rao's 1990, results were obtained for a more specific model than is considered here). It is obtained by bringing together the relations (4.7), (4.14) and (4.10) for m.l. estimation, or (4.7), (4.14) and (4.11) for REML estimation. This estimator will be compared to the often-reported estimator $[M_1(\hat{\tau}^2)]_{ii}$, in Section 5, using 1980 U. S. Census and Post Enumeration Survey data.

5. A COMPARISON OF ESTIMATORS BY EXAMPLE AND BY SIMULATION

5.1 Example

The PEP 3-8 data from the 1980 Post Enumeration Survey, for the $n = 51$ states of the USA (including Washington, DC) are used to illustrate the empirical Bayes approach. These data are presented in Cressie (1989, Table 1, "Total" columns) and the variances $\delta_1^2, \ldots, \delta_{51}^2$ in (1.3) are obtained from Cressie's "Total" column labeled MSE$^{1/2}$ (whose squared entries will be denoted $\text{MSE}_1, \ldots, \text{MSE}_{51}$). Using the relation $F_i = [1 - U_i/100]^{-1}$ and the δ-method, $\delta_i^2 \simeq (Y_i)^4(\text{MSE}_i)/10^4$. Eight explanatory variables, given by Ericksen, Kadane and Tukey (1989), were collapsed to the 51 states (from 66 small areas that included cities, rest of states and states). The explanatory variables are:

1. Minority percentage.
2. Crime rate.
3. Poverty percentage.
4. Percentage with language difficulty.
5. Education.
6. Housing.
7. Proportion of population in any of 16 prespecified central cities.
8. Percentage conventionally counted in the census.

To find a subset of these variables that provides a good model for undercount, I used the selection method of Ericksen, Kadane and Tukey (1989), but weighted the data proportionally to the square roots of the small areas' census counts. The variables selected were 1 (minority) and 5 (education), as well as the constant term. Henceforth, in this paper, these three variables will be the only ones considered in the linear model; *i.e.* only regression coefficients β_0, β_1 and β_5 will be fit.

Under the model (1.4), (1.5) and (1.6), the unknown parameters are $\underset{\sim}{\beta}$ and τ^2. From the scoring algorithm (3.8), the m.l. estimate of τ^2 is:

$$\hat{\tau}^2_{m\ell} = 47.32,$$

while from the scoring algorithm (3.23), the REML estimate of τ^2 is:

$$\hat{\tau}^2_{r\ell} = 58.53.$$

This illustrates a phenomenon observed from the realizations of a simulation presented below, namely, that $\hat{\tau}^2_{m\ell} < \hat{\tau}^2_{r\ell}$; an intuitive explanation is given in Section 3.3. (Parenthetically, Cressie (1990), obtained $\hat{\tau}^2_{mm} = 94.96$, but no general inequality between it, m.l., and REML is apparent.)

From the formulas in Section 3, the following estimates (with estimated standard errors in parentheses) were obtained:

m.l.	REML
$\hat{\beta}_0 = 1.03227\ (0.00708)$	$\hat{\beta}_0 = 1.03246\ (0.00724)$
$\hat{\beta}_1 = 0.0006878\ (0.0001402)$	$\hat{\beta}_1 = 0.0006941\ (0.0001436)$
$\hat{\beta}_5 = -0.001070\ (0.000231)$	$\hat{\beta}_5 = -0.001078\ (0.000236)$
$\hat{\tau}^2 = 47.32\ (32.87)$	$\hat{\tau}^2 = 58.53\ (38.1)$.

Notice that there is very little difference between the two sets of estimates, except for that of τ^2. Upon using the m.l. and REML estimates in $\hat{p}_i(\underset{\sim}{Y};\hat{\tau}^2)$ given by (2.5), $[M_1(\hat{\tau}^2)]_{ii}$ given by (2.6), and $[M_2(\tau^2)]^*_{ii}$ given by (4.7); $i = 1, \ldots, n$, small-area predictors and estimated root mean squared prediction errors are obtained. Table 1 shows the results for the $n = 51$ states; also shown in the table are the raw undercount data Y_i, the fitted linear model $(X\hat{\underset{\sim}{\beta}})_i$, and the weight,

$$w_i \equiv \hat{\tau}^2/(C_i\delta_i^2 + \hat{\tau}^2), \tag{5.1}$$

such that

$$\hat{p}_i(\underset{\sim}{Y};\hat{\tau}^2) = w_iY_i + (1 - w_i)(X\hat{\underset{\sim}{\beta}})_i;\ i = 1, \ldots, 51. \tag{5.2}$$

Notice that w_i for REML is consistently larger than w_i for m.l., which is intuitively sensible since $\hat{\tau}^2_{m\ell}$ has a notoriously large, negative bias. Thus, REML estimation of τ^2 results in less weight on the model term $(X\hat{\underset{\sim}{\beta}})_i$, but in a way so that the effect of estimation of τ^2 can be incorporated.

It is interesting to notice that one pays a price for using REML; its root mean squared prediction errors are consistently larger. This is not surprising, since we know that (asymptotically) m.l. is 100% efficient. Further, notice that the improved root mean squared prediction error, $\surd[M_2(\tau^2)]^*_{ii}$, is between 1% and 9% larger than $\surd[M_1(\hat{\tau}^2)]_{ii}$.

With regard to prediction, one can assess the importance of m.l. versus REML estimation of τ^2 by computing the weighted sum of squares,

$$\sum_{i=1}^{51} \{\hat{p}_i(\underset{\sim}{Y};\hat{\tau}^2_{m\ell}) - \hat{p}_i(\underset{\sim}{Y};\hat{\tau}^2_{r\ell})\}^2C_i = 15.$$

When compared to,

$$\sum_{i=1}^{51} (Y_i - 1)^2C_i = 70{,}421$$

and

$$\sum_{i=1}^{51} \{Y_i - \hat{p}_i(\underset{\sim}{Y};\hat{\tau}^2_{m\ell})\}^2 = 26{,}033,$$

Table 1: Columns, from left to right, show the 51 states according to a three-letter identifier, their raw undercounts $\{Y_i\}$, their model fits $\{(X\hat{\underset{\sim}{\beta}})_i\}$, their weights $\{w_i\}$ given by (5.1), their predictors (5.2) (headed F12), their root mean squared prediction errors $\{\surd[M_1(\hat{\tau}^2)]_{ii}\}$ (headed RMPE1), and their improved root mean squared prediction errors $\{\surd[M_2(\tau^2)]^*_{ii}\}$ (headed RMPE2). Table is given over the page.

Table 1

STATE	Y	REML				
		MDLFT	WGHT	F12	RMPE1	RMPE2
ala	0.9965	1.0037	0.1431	1.0026	0.00439	0.00453
aka	1.0288	1.0175	0.4767	1.0229	0.00896	0.00976
arz	1.0204	1.0158	0.0742	1.0162	0.00487	0.00500
ark	0.9895	0.9962	0.1398	0.9953	0.00541	0.00562
cal	1.0307	1.0225	0.0682	1.0231	0.00322	0.00327
col	1.0033	1.0199	0.1926	1.0167	0.00473	0.00495
con	0.9886	1.0079	0.1029	1.0059	0.00435	0.00451
del	0.9938	1.0107	0.4571	1.0030	0.00739	0.00811
fla	1.0144	1.0120	0.0785	1.0122	0.00289	0.00295
gga	0.9955	1.0046	0.1639	1.0031	0.00391	0.00403
hai	1.0111	1.0105	0.2785	1.0107	0.00678	0.00730
idh	1.0125	1.0070	0.5627	1.0101	0.00531	0.00579
ill	1.0211	1.0103	0.1170	1.0116	0.00257	0.00265
ind	0.9936	1.0026	0.1413	1.0013	0.00334	0.00349
iow	0.9932	1.0033	0.1478	1.0018	0.00452	0.00475
kan	1.0056	1.0092	0.2215	1.0084	0.00466	0.00496
kty	0.9845	0.9872	0.1519	0.9868	0.00507	0.00524
lou	1.0234	1.0086	0.0263	1.0090	0.00476	0.00480
mne	1.0201	0.9992	0.3703	1.0069	0.00593	0.00645
mld	1.0242	1.0140	0.0712	1.0147	0.00406	0.00415
mas	0.9882	1.0068	0.1945	1.0032	0.00323	0.00341
mch	1.0079	1.0081	0.1601	1.0081	0.00259	0.00271
min	1.0111	1.0049	0.2793	1.0066	0.00359	0.00383
mis	1.0097	1.0086	0.1279	1.0087	0.00557	0.00575
mou	1.0080	1.0010	0.1681	1.0022	0.00350	0.00367
mon	1.0144	1.0059	0.3785	1.0091	0.00699	0.00761
neb	1.0008	1.0071	0.5117	1.0039	0.00441	0.00480
nev	1.0265	1.0151	0.2852	1.0183	0.00744	0.00802
nwh	0.9842	1.0033	0.3080	0.9974	0.00684	0.00740
nwj	1.0130	1.0105	0.0895	1.0107	0.00305	0.00314
nwm	1.0236	1.0256	0.3276	1.0249	0.00611	0.00648
nwy	1.0166	1.0119	0.0807	1.0123	0.00243	0.00247
noc	1.0118	0.9998	0.0748	1.0007	0.00421	0.00430
nod	1.0005	0.9969	0.8931	1.0001	0.00313	0.00324
oho	1.0108	1.0044	0.1273	1.0052	0.00253	0.00263
okl	0.9977	1.0018	0.1625	1.0011	0.00429	0.00451
ore	1.0027	1.0089	0.2833	1.0071	0.00434	0.00464
pen	0.9972	1.0013	0.1475	1.0007	0.00253	0.00263
rhi	1.0089	0.9939	0.4167	1.0001	0.00625	0.00678
soc	1.0632	1.0040	0.0216	1.0053	0.00555	0.00559
sod	1.0008	0.9985	0.7538	1.0002	0.00464	0.00496
ten	0.9717	0.9966	0.0755	0.9947	0.00439	0.00449
tex	1.0037	1.0149	0.0482	1.0144	0.00341	0.00345
uth	1.0040	1.0142	0.4010	1.0101	0.00524	0.00563
vmt	0.9889	1.0018	0.8232	0.9912	0.00454	0.00479
vir	1.0009	1.0058	0.1753	1.0049	0.00338	0.00354
was	1.0142	1.0121	0.1305	1.0123	0.00418	0.00434
wev	0.9942	0.9877	0.1452	0.9887	0.00603	0.00628
wis	1.0173	1.0032	0.2877	1.0073	0.00325	0.00348
wyo	1.0361	1.0127	0.3992	1.0221	0.00882	0.00963
dcl	1.0375	1.0474	0.2191	1.0452	0.01081	0.01125

Table 1 (concluded)

STATE	Y	ML				
		MDLFT	WGHT	F12	RMPE1	RMPE2
ala	0.9965	1.0037	0.1190	1.0028	0.00415	0.00427
aka	1.0288	1.0175	0.4241	1.0223	0.00850	0.00933
arz	1.0204	1.0157	0.0608	1.0160	0.00448	0.00459
ark	0.9895	0.9963	0.1161	0.9955	0.00506	0.00525
cal	1.0307	1.0224	0.0559	1.0228	0.00314	0.00319
col	1.0033	1.0198	0.1617	1.0171	0.00446	0.00466
con	0.9886	1.0079	0.0849	1.0063	0.00398	0.00412
del	0.9938	1.0107	0.4050	1.0039	0.00697	0.00771
fla	1.0144	1.0120	0.0644	1.0121	0.00271	0.00276
gga	0.9955	1.0046	0.1368	1.0034	0.00375	0.00385
hai	1.0111	1.0105	0.2378	1.0106	0.00629	0.00679
idh	1.0125	1.0070	0.5099	1.0098	0.00507	0.00559
ill	1.0211	1.0103	0.0967	1.0113	0.00242	0.00248
ind	0.9936	1.0026	0.1174	1.0015	0.00309	0.00323
iow	0.9932	1.0034	0.1230	1.0021	0.00418	0.00438
kan	1.0056	1.0091	0.1870	1.0085	0.00432	0.00460
kty	0.9845	0.9874	0.1264	0.9870	0.00486	0.00502
lou	1.0234	1.0086	0.0214	1.0089	0.00446	0.00449
mne	1.0201	0.9993	0.3222	1.0060	0.00557	0.00608
mld	1.0242	1.0139	0.0583	1.0145	0.00376	0.00384
mas	0.9882	1.0068	0.1634	1.0037	0.00302	0.00319
mch	1.0079	1.0081	0.1335	1.0081	0.00242	0.00252
min	1.0111	1.0049	0.2386	1.0064	0.00339	0.00362
mis	1.0097	1.0085	0.1060	1.0087	0.00526	0.00541
mou	1.0080	1.0011	0.1404	1.0021	0.00326	0.00341
mon	1.0144	1.0059	0.3299	1.0087	0.00656	0.00717
neb	1.0008	1.0071	0.4587	1.0042	0.00420	0.00461
nev	1.0265	1.0150	0.2439	1.0178	0.00692	0.00746
nwh	0.9842	1.0033	0.2646	0.9983	0.00637	0.00691
nwj	1.0130	1.0105	0.0736	1.0106	0.00283	0.00290
nwm	1.0236	1.0254	0.2826	1.0249	0.00582	0.00617
nwy	1.0166	1.0119	0.0663	1.0122	0.00231	0.00235
noc	1.0118	0.9998	0.0614	1.0005	0.00401	0.00408
nod	1.0005	0.9970	0.8710	1.0000	0.00310	0.00324
oho	1.0108	1.0045	0.1055	1.0051	0.00236	0.00245
okl	0.9977	1.0018	0.1356	1.0013	0.00396	0.00416
ore	1.0027	1.0088	0.2421	1.0074	0.00408	0.00436
pen	0.9972	1.0014	0.1227	1.0008	0.00239	0.00248
rhi	1.0089	0.9940	0.3660	0.9995	0.00591	0.00645
soc	1.0632	1.0041	0.0176	1.0051	0.00519	0.00523
sod	1.0008	0.9985	0.7122	1.0002	0.00452	0.00490
ten	0.9717	0.9967	0.0619	0.9951	0.00413	0.00422
tex	1.0037	1.0148	0.0393	1.0144	0.00329	0.00332
uth	1.0040	1.0141	0.3512	1.0105	0.00498	0.00536
vmt	0.9889	1.0019	0.7901	0.9916	0.00445	0.00477
vir	1.0009	1.0058	0.1467	1.0051	0.00317	0.00330
was	1.0142	1.0120	0.1082	1.0123	0.00391	0.00406
wev	0.9942	0.9879	0.1207	0.9886	0.00567	0.00590
wis	1.0173	1.0033	0.2461	1.0067	0.00306	0.00328
wyo	1.0361	1.0127	0.3494	1.0209	0.00829	0.00909
dcl	1.0375	1.0470	0.1849	1.0452	0.01036	0.01078

it is clear that, from a national perspective, prediction is not very sensitive to estimation methods for τ^2. (Cressie (1990) reaches the same conclusion based on a similar comparison of $\hat{\tau}^2_{m\ell}$ and $\hat{\tau}^2_{mm}$.) However, from Table 1, it is equally clear that estimated root mean squared prediction errors are considerably more sensitive.

Cressie (1990) gives expressions for the risks of adjusting using $\hat{p}(\underset{\sim}{Y};\tau^2)$ and of not adjusting. When $\hat{\tau}^2_{r\ell}$ and $\hat{\underset{\sim}{\beta}}(\hat{\tau}^2_{r\ell})$ are substituted into those expressions, the risk of adjusting is 3,253, while the risk, of not adjusting is 34,134. That is, not adjusting leads to a 949% increase in risk (provided the model defined by (1.4), (1.5) and (1.6) holds).

5.2 Simulation

To check the asymptotic distribution theory of the REML (and m.l.) estimator of τ^2, a simulation was carried out on the linear model described in Section 5.1, with parameter values:

$$\beta_0 = 1.0330, \qquad \beta_1 = 0.000712, \qquad \beta_5 = -0.000110, \qquad \tau^2 = 95.00. \tag{5.3}$$

The simulation,

$$\underset{\sim}{Y} \sim \mathrm{Gau}(X\underset{\sim}{\beta}, \Delta + \tau^2 D), \tag{5.4}$$

where Δ is given by (1.4) the same values of $\delta_1^2, \ldots, \delta_{51}^2$, as used in Section 5.1 and Cressie in 1990, are used here and D is given by (1.6), was performed 500 times, and each time the estimates, $\hat{\tau}^2_{m\ell}$, $\hat{\tau}^2_{mm}$, and $\hat{\tau}^2_{r\ell}$ were computed. (Whenever a negative value was obtained, the estimate was set equal to zero.) The stem-and-leaf plots of the three sets of estimates are presented in Figures 1a, 1b and 1c, respectively. Notice the relatively larger number of zeros for the m.l. estimates (Figure 1a).

Figure 1. Stem-and-leaf plots of estimated variance parameter τ^2, based on 500 simulations of (5.4): (a) maximum likelihood (Section 3.1), (b) method-of-moments (Section 3.2) and (c) restricted maximum likelihood (Section 3.3).

```
 0   0000000000000001155556667
 1   0001223566667889
 2   000112356677899
 3   0011122344555557779999
 4   0011111112222333444455555566667777788888899999
 5   00001222333333344555666666778888889999999
 6   00000111111112222222222333334444455666777777788889999
 7   00011111111222223344444455555666677777888888999
 8   0011122222223333333333444555667777777788889999
 9   0000111122222223333345556777777788
10   000011111112333344444566777777888899
11   000111222223444445666 7899
12   00011112222333333667778 8899
13   1223345556677999
14   000122233444566 6799
15   00001222334455 8999
16   157899
17   001122233589
18   2568
19   145
20   7
21   2
22   88
```

Figure 1a

```
 0  000000377778
 1  011113344446679999
 2  11144455557778888
 3  00002222222233333355555666666888899999
 4  122222224466667777777999999
 5  0002222333333455577777778888
 6  0000000111333334444444444666667777999999999
 7  111122222224444455555555777778888888
 8  00000222333333355555556666666688999
 9  1112222222244444466666777779999
10  000000222333333555777777888888
11  00000000111113333444444666677789999
12  1111122222224444444445777778
13  00002233666888888999
14  11122244667777999
15  002233558888
16  000001133444777799
17  12222224555788
18  00003335566899
19  26799
20  02258
21  37
22  11558
23  5
24  79
25
26
27  5
28  5
29  2
30  78
31  3
32
33  2
```

Figure 1b

The means ($\bar{X}$) and standard deviations (S) of the distributions shown in Figure 1 are:

$\hat{\tau}^2_{m\ell}$	$\hat{\tau}^2_{mm}$	$\hat{\tau}^2_{r\ell}$
$\bar{X} = 83.56$	$\bar{X} = 96.85$	$\bar{X} = 94.27.$
$S = 45.65$	$S = 57.46$	$S = 49.17.$

The means should be compared to the true value of $\tau^2 = 95.00$. The bias in $\hat{\tau}^2_{m\ell}$ is apparent; $\hat{\tau}^2_{r\ell}$ has very little bias and has a small advantage over $\hat{\tau}^2_{mm}$. With regard to standard deviations, the advantage of $\hat{\tau}^2_{r\ell}$ over $\hat{\tau}^2_{mm}$ is considerable, but it is at some disadvantage over $\hat{\tau}^2_{m\ell}$. For reasons explained in Section 3.3, that are not all statistical, bias is more of a concern than variance, and so REML estimation of τ^2 should be considered a serious alternative to m.l.

Asymptotic distribution theory for m.l. and REML can be checked from the simulations. (The method of moments is at a disadvantage in that no asymptotic distribution theory is readily available.) Substituting $\tau^2 = 95.00$ into (3.13) yields,

0	00000001234777799
1	0012334567789
2	012225556888899
3	112234444556889
4	0013344445555666777888888899
5	000011222333333444445566777788888999
6	00011122222223344444444445666677778888999999
7	00000011111112222222223333444445556777788888899999
8	00000001111222333455555566666777778889
9	000111222222222333333444455555566689999999
10	00000000011122333444555566777888899
11	00011222333444556666777888888899
12	0001111112223334445556 7789
13	00013333455555555556788
14	00011123444445667789
15	0011122234456678 8
16	001112222335555 7999
17	011235556
18	00112566777779
19	117
20	013478
21	123
22	7
23	6
24	
25	02

Figure 1c

$$\{\mathrm{var}(\hat{\tau}^2_{m\ell})\}^{1/2} \simeq 48.73,$$

which should be compared to $S = 45.65$. Finally, substituting $\tau^2 = 95.00$ into (3.29) yields,

$$\{\mathrm{var}(\hat{\tau}^2_{r\ell})\}^{1/2} = 50.14,$$

which should be compared to $S = 49.17$.

The opportunity also exists to use the simulation to look at "actual" errors of prediction and to assess the performance of $M_1(\hat{\tau}^2)$ and $M_2(\tau^2)^*$. If the parameter values (5.3) were estimated from the original data, then this amounts to a parametric boostrap.

6. CONCLUSIONS AND DISCUSSION

Model-based prediction of undercount relies on careful checking of model fit. Diagnostic plots based on standardized residuals have already been suggested at the end of Section 2. The standardized BLUP residuals $\{Y_i - \hat{p}_i(\underset{\sim}{Y}; \hat{\tau}^2)\}/\{[M(\hat{\tau}^2)]_{ii}\}^{1/2}$; $i = 1, \ldots, n$, also have a role to play. They could either be used in a quantile-quantile plot (*e.g.* Cressie 1991, p. 225) or, as suggested by Calvin and Sedransk (1991), plotted against $\hat{p}_i(\underset{\sim}{Y}; \hat{\tau}^2)$; $i = 1, \ldots, n$.

One could also extend the model (1.4) to include an unknown variance-component parameter σ^2:

$$\underset{\sim}{Y} \sim \mathrm{Gau}(\underset{\sim}{F}, \sigma^2\Delta), \tag{6.1}$$

where $\Delta = \text{diag}\{\delta_1^2, \ldots, \delta_n^2\}$. Upon fitting the more general model (6.1), (1.5) and (1.6), one could then test whether the REML estimate $\sigma_{r\ell}^2$ is significantly different from $\sigma^2 = 1$, which would provide a check on model misspecification. (In this case, REML estimation is recommended over m.l. estimation, since any bias will seriously affect inference on σ^2.)

Restricted maximum likelihood (REML) estimation of variance-matrix parameters is less likely to lead to empirical Bayes predictors that put too much weight on the regression model (1.5). The price paid is slightly larger mean squared prediction errors. Using asymptotic distribution theory for REML (which is checked by simulation), improved estimators of the mean squared prediction errors can also be obtained. Based on the model (1.4), (1.5) and (1.6), it can be concluded that there are accurate and precise ways to make inference on adjustment factors $\{F_i : i = 1, \ldots, n\}$; the predictors $\{\hat{p}_i(\underset{\sim}{Y};\hat{\tau}_{r\ell}^2): i = 1, \ldots, n\}$ yield true-count and undercount predictors,

$$T_i^{\text{prd}} = \hat{p}_i(\underset{\sim}{Y};\hat{\tau}_{r\ell}^2)C_i \quad \text{and} \quad U_i^{\text{prd}} = 100[1 - (\hat{p}_i(\underset{\sim}{Y};\hat{\tau}_{r\ell}^2))^{-1}]; \quad i = 1, \ldots, n,$$

respectively. Their biases and mean-squared prediction errors can be obtained using the δ-method (*cf.* Cressie 1991, Section 3.2.2).

ACKNOWLEDGEMENTS

The research assistance of Robert Parker is gratefully acknowledged. This research was supported by Joint Statistical Agreement JSA 90-41, between the U. S. Bureau of the Census and Iowa State University. The findings and opinions expressed in this article are of the author alone, and do not necessarily reflect those of the Census Bureau.

REFERENCES

CALVIN, J.A., and SEDRANSK, J. (1991). Bayesian and frequentist predictive inference for the patterns of care studies. *Journal of the American Statistical Association*, 86, 36-48.

CRESSIE, N. (1988). Estimating census undercount at national and subnational levels. In *Proceedings of Bureau of the Census Fourth Annual Research Conference*, Washington, DC: U. S. Bureau of the Census, 127-150.

CRESSIE, N. (1989). Empirical Bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.

CRESSIE, N. (1990). Weighted smoothing of estimated undercount. In *Proceedings of Bureau of the Census 1990 Annual Research Conference*, Washington, DC: U. S. Bureau of the Census, 301-325.

CRESSIE, N. (1991). *Statistics for Spatial Data*. New York: Wiley.

CRESSIE, N., and LAHIRI, S.N. (1991). The asymptotic distribution of REML estimators. *Statistical Laboratory Preprint 91-20*, Iowa State University, Ames, Iowa.

EATON, M.L. (1985). The Gauss-Markov Theorem in multivariate analysis. In *Multivariate Analysis – VI*, (Ed. P.R. Krishnaiah). Amsterdam: Elsevier, 177-201.

ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-109.

ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1989). Adjusting the 1980 Census of population and housing. *Journal of the American Statistical Association*, 84, 927-944.

FAY, R.E., III, and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

FELLNER, W.H. (1986). Robust estimation of variance components. *Technometrics*, 28, 51-60.

FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 Census. *Statistical Science*, 1, 3-17.

GELFAND, A.E., and SMITH, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

GROENEVELD, R.A., and MEEDEN, G.D. (1977). The mode, median and mean inequality. *American Statistician*, 31, 120-121.

HARVILLE, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61, 383-385.

HARVILLE, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-340.

HARVILLE, D.A. (1985). Decomposition of prediction error. *Journal of the American Statistical Association*, 80, 132-138.

KASS, R.E., and STEFFEY, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84, 717-726.

KITANIDIS, P.K. (1983). Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research*, 19, 909-921.

LAIRD, N.M., and LOUIS, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82, 739-750.

MARDIA, K.V., and MARSHALL, R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71, 135-146.

MILLER, J.J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *Annals of Statistics*, 5, 746-762.

PATTERSON, H.D., and THOMPSON, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 58, 545-554.

PATTERSON, H.D., and THOMPSON, R. (1974). Maximum likelihood estimation of components of variance. *Proceedings of the 8th International Biometric Conference*. Washington, DC: Biometric Society, 197-207.

PRASAD, N.G.N., and RAO, J.N.K. (1990). On the estimation of mean square error of small area predictors. *Journal of the American Statistical Association*, 85, 163-171.

RAO, C.R. (1979). MINQE theory and its relation to ML and MML estimation of variance components. *Sankhyā B*, 41, 138-153.

WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.

ZIMMERMAN, D.L. (1989). Computationally efficient restricted maximum likelihood estimation of generalized covariance functions. *Mathematical Geology*, 21, 655-672.

ZIMMERMAN, D.L., and CRESSIE, N. (1991). Mean-squared prediction error in the spatial linear model. *Annals of the Institute of Statistical Mathematics*, 43, forthcoming.

Survey Methodology, June 1992
Vol. 18, No. 1, pp. 95–108
Statistics Canada

Hierarchical and Empirical Bayes Methods for Adjustment of Census Undercount: The 1988 Missouri Dress Rehearsal Data

G.S. DATTA, M. GHOSH, E.T. HUANG, C.T. ISAKI, L.K. SCHULTZ and J.H. TSAY[1]

ABSTRACT

The present article discusses a model-based approach towards adjustment of the 1988 Census Dress Rehearsal Data collected from test sites in Missouri. The primary objective is to develop procedures that can be used to model data from the 1990 Census Post Enumeration Survey in April, 1991 and smooth survey-based estimates of the adjustment factors. We have proposed in this paper hierarchical Bayes (HB) and empirical Bayes (EB) procedures which meet this objective. The resulting estimators seem to improve consistently on the estimators of the adjustment factors based on dual system estimation (DSE) as well as the smoothed regression estimators.

KEY WORDS: Post Enumeration Survey; Adjustment factors; Dual system estimation; Hierarchical Bayes; Empirical Bayes; Variance components; EBLUP's; Regression estimates; Standard errors.

1. INTRODUCTION

The present article discusses a model-based approach towards adjustment of the 1988 Census dress rehearsal data collected from test sites in Missouri. The main objective behind this exercise is to develop procedures that can be used to model data from the 1990 Census Post Enumeration Survey (PES) in April, 1991, and smooth survey-based estimates of the so-called "raw adjustment factors". These raw adjustment factors which are ratios of estimates of the unknown total population to the corresponding 1990 Census count, are computed at various levels of aggregation (geographic areas such as cities, suburbs, *etc.*) crossed by various demographic categories (such as age, sex, race, *etc.*). The cross-classified categories are called poststrata.

Before proceeding further, a brief historical anecdote is in order. Adjustment of 1980 decennial census counts in the United States has been a topic of heated debate for nearly a decade. Despite the intensive efforts and the massive expenditure incurred by the U.S. Bureau of the Census to achieve near-complete coverage in the 1980 Census, there have been many lawsuits against the Bureau by individual states and cities demanding revision of the reported counts. In one such instance of litigation, by now well-publicized to the Statistics community in the articles of Ericksen and Kadane (1985) and Freedman and Navidi (1986), New York City among others sued the Census Bureau, and many reputed statisticians appeared as expert witnesses on either side. In particular Ericksen and Kadane appeared on the plaintiff's side, and proposed a model-based approach towards the adjustment of census counts. They advocated shrinking the adjustment factors calculated on the basis of the PES data towards some suitable regression model. This approach documented in Ericksen and Kadane (1985) is similar to the one considered in Fay and Herriot (1979) or Morris (1983). Despite criticism of the Ericksen-Kadane approach by some statisticians (most severely by Freedman and

[1] G.S. Datta, University of Georgia, Athens, GA 30602; M. Ghosh, University of Florida, Gainesville, FL 32611; E.T. Huang, C.T. Isaki, L.K. Schultz and J.H. Tsay, U.S. Bureau of the Census, Washington, DC 20233.

Navidi (1986)), most people recognize the importance of the model-based approach for adjustment. Indeed, in this article, barring a few differences in the assumptions, to be pointed out later in section 2, we use the Fay-Herriot or the Ericksen-Kadane model for the analysis of the 1988 Missouri Dress Rehearsal data. A different model-based approach which does not include co-variates is given in Cressie (1989).

A good description of the PES conducted as part of the 1988 Missouri Dress Rehearsal can be found in Childers and Hogan (1990). Hogan and Wolter (1988) discuss the categories of error that occur in a PES and a means of their evaluation. Basically, the PES design consists of a single stage stratified sample of blocks and dual system estimation of the number of persons by poststrata.

In the present article, we begin at the point where a set of estimated raw adjustment factors and their covariances from the PES are available for modelling based on the 1988 Census Dress Rehearsal Data from the Missouri test sites. It is also assumed that a set of possible explanatory variables defined at the poststrata level and to be used in regression are also available. There are two geographic areas under consideration: the city of St. Louis which is a large central city, and East Central Missouri, which is a collection of areas of moderate population size. In defining the poststrata in St. Louis, persons were classified into the following demographic categories: (i) race: white non-hispanic and others, (ii) owners and non-owners (renters) of dwellings, (iii) sex: male and female, (iv) age groups: 0-9, 10-19, 20-29, 30-44, 45-64 and 65+. This led to a total of $2 \times 2 \times 2 \times 6 = 48$ adjustment factors for St. Louis. In East Central Missouri, the sex and the age-group categories remained the same as in St. Louis, but instead of (i) and (ii), a new category (i)' classifying persons as (a) White non-Hispanic in Tape Address Register (TAR) areas, (b) White non-Hispanic in non-TAR areas, and (c) others in all areas were introduced. For East Central Missouri, a total of $3 \times 2 \times 6 = 36$ adjustment factors were calculated. Thus, a total of 84 adjustment factors were used for modelling. Within each area, estimated adjustment factors were correlated due to the use of a block cluster sampling scheme. This led to a block-diagonal sample covariance matrix of the adjustment factors of dimensions 48×48 and 36×36 corresponding to St. Louis and East Central Missouri, respectively.

In Section 2 of this article, we describe a general model-based method for obtaining smoothed adjustment factors, and the associated standard errors. Both the hierarchical and empirical Bayes methods are used. The EB method can also be regarded as a variance components method (see for example Harville (1985)). The formulas for posterior standard errors associated with the HB estimators are also provided. We may point out here that an EB method when employed naively can lead to serious underestimates of the associated standard errors. This is due to the fact that a naive EB method does not take into account the uncertainty due to estimation of the unknown variance components. However, Kackar and Harville (1984), and Prasad and Rao (1990) have suggested interesting approximations to the estimated mean squared errors (MSE's) of the EB estimators. Following their principle, we have derived formulas for the estimated MSE's in the present context. We have also pointed out in this section how some (though not all) of the criticisms levelled against the Ericksen-Kadane (1985) procedure by Freedman and Navidi (1986) can be avoided in the present context.

In Section 3, we have analyzed the actual data. The sample estimates, the HB estimates, the EB estimates and the regression estimates of the adjustment factors are all provided. Also, the associated standard errors are given. Both the HB method and the EB methods which take into account the uncertainty due to unknown prior parameters stand on par in their performance, and enjoy a clear-cut superiority over the raw estimates as well as the regression estimates in reducing the estimated standard errors.

Finally, some of the technical details of this paper are given in the Appendix.

2. HB AND EB ESTIMATION

This section describes the general HB and EB estimation procedures for certain hierarchical models. The specific application to estimation of adjustment factors is considered in Section 3.

The following hierarchical model is considered:

I. $Y \mid \Theta, \beta, \sigma^2 \sim N(\Theta, V)$, where V is a known $m \times m$ positive definite matrix;

II. $\Theta \mid \beta, \sigma^2 \sim N(X\beta, \sigma^2 I)$;

III. β and σ^2 are marginally independent with β uniform (R^p) and σ^2 uniform $(0,\infty)$.

The HB analysis is based on I-III. In the absence of precise prior information on β and σ^2, we prefer the use of diffuse priors in III. We also analyzed the data with the prior *pdf* of σ^2 proportional to σ^{-2} on $(0, \infty)$. The results were quite similar and are not reported. The following theorem is proved.

Theorem 1. Consider the model given in (I) - (III). Write $\Sigma = V + \sigma^2 I$. Suppose $m \geq p + 3$. Then (i) the conditional pdf of Θ given σ^2 and $Y = y$ is $N(GV^{-1}y, G)$, where

$$G = V - V\Sigma^{-1}V + V\Sigma^{-1}X[X^T\Sigma^{-1}X]^{-1}X^T\Sigma^{-1}V; \tag{2.1}$$

(ii) the conditional pdf of σ^2 given $Y = y$ is

$$f(\sigma^2 \mid y) \propto \mid \Sigma \mid^{-1/2} \mid X^T\Sigma^{-1}X \mid^{-1/2} \exp(-1/2\, y^T F y), \tag{2.2}$$

where

$$F = \Sigma^{-1} - \Sigma^{-1}X[X^T\Sigma^{-1}X]^{-1}X^T\Sigma^{-1}. \tag{2.3}$$

The proof of the theorem is deferred to the appendix. Using formulas for conditional expectations and variances, one then gets

$$E(\Theta \mid y) = E[E(\Theta \mid \sigma^2, y) \mid y] = (E(GV^{-1} \mid y))\, y; \tag{2.4}$$

$$V(\Theta \mid y) = V[E(\Theta \mid \sigma^2, y) \mid y] + E[V(\Theta \mid \sigma^2, y) \mid y] = V(GV^{-1} y \mid y) + E(G \mid y). \tag{2.5}$$

Using (2.2) and (2.3), one obtains $E(\Theta \mid y)$ and $V(\Theta \mid y)$ from (2.4) and (2.5) via numerical integration.

The calculations involved in (2.1) - (2.3) can be somewhat simplified when one uses the spectral decomposition theorem for V. Thus, $V = PDP^T$, where $D = \text{Diag}(d_1, \ldots, d_m)$, d_i being the eigenvalues of V, and $P = (\xi_1 \ldots, \xi_m)$, ξ_i being the corresponding orthonormal eigenvectors. Using the orthogonality of P, one now gets

$$|\Sigma| = |\sigma^2 I + D| = \prod_{i=1}^{m} (\sigma^2 + d_i);$$

$$\Sigma^{-1} = P(\sigma^2 I + D)^{-1} P^T;$$

$$X^T \Sigma^{-1} X = (P^T X)^T (\sigma^2 I + D)^{-1} (P^T X);$$

$$F = P(\sigma^2 I + D)^{-1} P^T - P(\sigma^2 I + D)^{-1} (P^T X) \times$$

$$\left[(P^T X)^T (\sigma^2 I + D)^{-1} (P^T X) \right]^{-1} (P^T X) (\sigma^2 I + D)^{-1}.$$

The actual numerical integration over σ^2 which needs evaluation of the integrand at different values of σ^2, is somewhat simplified since P and X are known and $\sigma^2 I + D$ is a diagonal matrix.

Next we consider EB estimation. Then, one does not use III. First a Bayes estimator, *i.e.* the posterior mean of Θ is obtained from I and II assuming β and σ^2 to be known. This estimator is given by

$$\begin{aligned} \hat{\Theta}^B &= \mathrm{E}(\Theta \mid Y, \beta, \sigma^2) \\ &= (V^{-1} + \sigma^{-2} I)^{-1} (V^{-1} Y + \sigma^{-2} X\beta) \\ &= \Sigma^{-1} (\sigma^2 Y + VX\beta). \end{aligned} \tag{2.6}$$

The corresponding posterior variance is given by

$$V(\Theta \mid Y, \beta, \sigma^2) = (V^{-1} + \sigma^{-2} I)^{-1} = V - V\Sigma^{-1} V.$$

However, in practice, β and σ^2 are unknown, and are estimated via the maximum likelihood method from the marginal distribution of Y which is $\mathrm{N}(X\beta, \Sigma)$. These MLE's are denoted by $\hat{\beta}$ and $\hat{\sigma}^2$, where $\hat{\beta} = (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} Y$, $\hat{\Sigma} = V + \hat{\sigma}^2 I$. Substituting such estimators of Σ, σ^2 and β in (2.6), an EB estimator of Θ is found as

$$\hat{\Theta}^{\mathrm{EB}} = \hat{\Sigma}^{-1} (\hat{\sigma}^2 Y + VX\hat{\beta}) = X\hat{\beta} + \hat{\sigma}^2 \hat{\Sigma}^{-1} (Y - X\hat{\beta}). \tag{2.7}$$

The estimator given in (2.7) is also obtainable as an estimated best linear unbiased predictor (EBLUP). First assume that σ^2 is known, and find the BLUP $\hat{\Theta}^{\mathrm{BLUP}} = X\tilde{\beta} + \sigma^2 \Sigma^{-1} (Y - X\tilde{\beta})$ of Θ where $\tilde{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$. Next estimate σ^2 by $\hat{\sigma}^2$, its MLE and correspondingly Σ by $\hat{\Sigma}$. Substitution of $\hat{\sigma}^2$, and $\hat{\Sigma}$ in place of σ^2 and Σ in $\hat{\Theta}^{\mathrm{BLUP}}$ results in the EBLUP $\hat{\Theta}^{\mathrm{EB}}$.

A naive EB estimator of the variance matrix of $\hat{\Theta}^{\mathrm{EB}}$ is $V - V\hat{\Sigma}^{-1} V$. This is a gross underestimation of the variance matrix since uncertainty due to estimation of β and σ^2 is not taken into account. If σ^2 is assumed known, and β is assigned a uniform prior on $R^p (m \geq p + 3)$, then the HB estimator of Θ is the same as $\hat{\Theta}^{\mathrm{BLUP}}$, and the posterior variance matrix is then $M = V - V\Sigma^{-1} V + V\Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} V$. This implies immediately that $\mathrm{E}[(\hat{\Theta}^{\mathrm{BLUP}} - \Theta)(\hat{\Theta}^{\mathrm{BLUP}} - \Theta)^T] = M$, where expectation is taken over the joint distribution of Y and Θ given in I and II. Thus, in the Bayesian language, $V\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}V$ can be interpreted as the excess in the posterior variability due to the uncertainty involved in β, while using the classical terminology, the same phenomenon can be interpreted as the excess in the MSE due to the same uncertainty.

We have the additional problem of tackling unknown σ^2. The Bayesian method enables us to find the posterior distribution of σ^2 given $Y = y$, while even without introducing a prior for Θ, it is still possible to find an approximation to the MSE of $\hat{\Theta}^{EB}$ by adapting an argument of Kackar and Harville (1984) or Prasad and Rao (1990).

The necessary theorem whose proof is deferred to the Appendix is given below.

Theorem 2. An approximate estimate of MSE of $\hat{\Theta}^{EB}$ is given by

$$\widehat{\mathrm{MSE}}\left(\Theta^{\mathrm{EB}}\right) \doteq V - V\hat{K}V + \left(V\hat{K}^3V\right)\left[2(\mathrm{tr}\hat{\Sigma}^{-2})^{-1}\right], \tag{2.8}$$

where

$$\hat{K} = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}X\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}. \tag{2.9}$$

The third term in the right hand side of (2.8) can be interpreted as the excess in the mean squared error due to uncertainty in estimating σ^2. A general decomposition of the prediction error is given in Harville (1985).

Although the posterior variances $V(\Theta \mid y)$ associated with the HB estimator $\hat{\Theta}^{HB}$ of Θ and the estimated MSE of the EB estimator $\hat{\Theta}^{EB}$ of Θ are motivated from two distinct inferential philosophies, one common thread tying the two is that they both attempt to incorporate the uncertainty due to estimation of the model variance. For a better understanding of this, note that writing $K = \Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$,

$$\mathrm{E}\left[V(\Theta \mid \sigma^2, y)\right] = G = V - VKV \tag{2.10}$$

and $\mathrm{E}(G \mid y)$ is approximated by $V - V\hat{K}V$ which is one of the two terms given in (2.8). Also, $\mathrm{E}(\Theta \mid \sigma^2, y) = GV^{-1}y$, and it can be shown after some simplification that $GV^{-1} = I - VK$. Thus, $V(GV^{-1}y \mid y) = VV(K \mid y)V$, and $V(K \mid y)$ is apparently approximated by $\hat{K}^3[2(\mathrm{tr}\hat{\Sigma}^{-2})^{-1}]$. However, as evidenced later in the numerical calculations of Section 3, MSE approximation of $\hat{\Theta}^{EB}$ need not match $V(\Theta \mid y)$ perfectly.

In Ericksen and Kadane (1985) one assumption involved was that of known σ^2. Freedman and Navidi (1986) insisted on estimation of σ^2, and we have in Theorems 1 and 2 accounted for this source of uncertainty both in a Bayesian and frequentist way. It should be noted that unlike previous work that addressed the estimation of net undercount of total population at the city and balance of state level, our interests lay in the estimation of adjustment factors at finer levels of detail. Operationally, adjustment at the finer levels allows for considerable savings in terms of time and computer costs as census files need to be used only once. Adjustment models using higher levels of geography would require several passes through the census data because they would require a method of distributing the undercount to lower levels of geography. Finally, correlation in the error structure allows the possibility of a non-diagonal V, another important generalization of the Fay-Herriot (1979) or Ericksen-Kadane (1985) model. Thus, the Freedman-Navidi criticism of lack of correlation across estimated adjustment factors does not hold against the present set up. The remaining main criticism of assuming the components of V to be known, whereas in reality these are sample based estimates, is yet to be resolved. Efforts are now being made to model the components of V as a function of variables such as the number of sample persons, the initial regression predictor, *etc*. It is hoped that such models will stabilize the estimated variances by reducing their variance.

Along with the HB and EB estimators of $\boldsymbol{\Theta}$, there are also the regression estimators given by $\hat{\boldsymbol{\Theta}}^{\mathrm{REG}} = X(X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}Y$. The associated variance-covariance matrix is given by $M_o - \hat{\sigma}^2(M_o\hat{\Sigma}^{-1} + \hat{\Sigma}^{-1}M_o^T - I)$, where $M_o = X(X^T\hat{\Sigma}^{-1}X)^{-1}X^T$.

3. DATA ANALYSIS

Let $Y_i = \mathrm{DSE}_i/Census_i$ = adjustment factor i, $i = 1, \ldots, 84$, and $Y = (Y_i, \ldots, Y_{84})^T$. The set of explanatory variables X is quite large when all possible interactions are considered. To simplify the analysis, experts at the Census Bureau were consulted and a reduced set of 22 potential explanatory variables were considered for modelling purposes. (See Huang *et al.* 1991). The number of potential explanatory variables was also limited by the capability of the computer. The present model was selected using a best subset regression procedure with minimum Mallows' C_p as the criterion over a set of 22 possible explanatory variables. Because the computer software required the input data to be in the ordinary least squares situation, we transformed the dependent and explanatory variables in the usual manner. Also, because σ^2 is unknown, an interative procedure was used.

As an aside, in selecting explanatory variables in the modelling process of adjustment factors for the 1990 Census, a slightly different procedure was used. In 1990, several explanatory variables were forced into the model and a best subset procedure was used to select additional explanatory variables. The change in procedure was made to counteract the potential for understating σ^2. (See Isaki *et al.* 1991).

The X matrix obtained via best subsets regression is of the form $X = (1_{84}, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10})$. All of the explanatory variables in X are obtained from the 1988 Dress Rehearsal Census and defined at the poststrata level, the unit of analysis. 1_{84} is a unit vector; X_2 is the indicator variable for St. Louis; X_3 is the indicator variable for renters or is the proportion of renters for the East Central Missouri poststrata; X_4 through X_7 are indicator variables for age groups 0-9, 10-19, 20-29 and 30-44, respectively; X_8 is an indicator or proportion variable for males aged 20-64 that rent; X_9 is an indicator variable for other males aged 20-64; and X_{10} is an indicator variable for other persons in St. Louis.

Using the above design matrix, we obtained $\hat{\beta} = (.9812, -.0271, .0485, .0699, .0695, .0533, .0386, .0628, .0475, .0778)^T$ and $\hat{\sigma}^2 = .000574$. The EB's or the EBLUP's and the associated approximate standard errors can now be computed using formulas derived in Section 2. For consistency, the HB analysis was also performed with the same X matrix (we do not require $\hat{\beta}$ or $\hat{\sigma}^2$ for that analysis).

In Figures 1 and 2 we plot the estimated adjustment factors and standard errors by poststrata. The first 12 poststrata refer to white non-Hispanic non-owners in St. Louis; poststrata 13-24 refer to all other non-owners in St. Louis; poststrata 25-36 refer to white non-Hispanic owners in St. Louis and poststrata 37-48 refer to all other owners in St. Louis. Poststrata 49-60 refer to white non-Hispanic persons in Tape Address Register (TAR) areas in East Central Missouri; poststrata 61-72 refer to white non-Hispanic persons in non-TAR areas in East Central Missouri; poststrata 73-84 refer to all other persons in East Central Missouri.

Within each group of 12 poststrata, the first six refer to males by age 0-9, 10-19, 20-29, 30-44, 45-64 and 65+. We note in Figure 1 that the raw adjustment factors for the other group tend to be higher than those for the white non-Hispanic except for TAR area in East Central Missouri. The same observation nearly holds in Figure 2 concerning the raw standard errors. In Figure 3 a plot of the estimated standard errors versus the adjustment factors is provided.

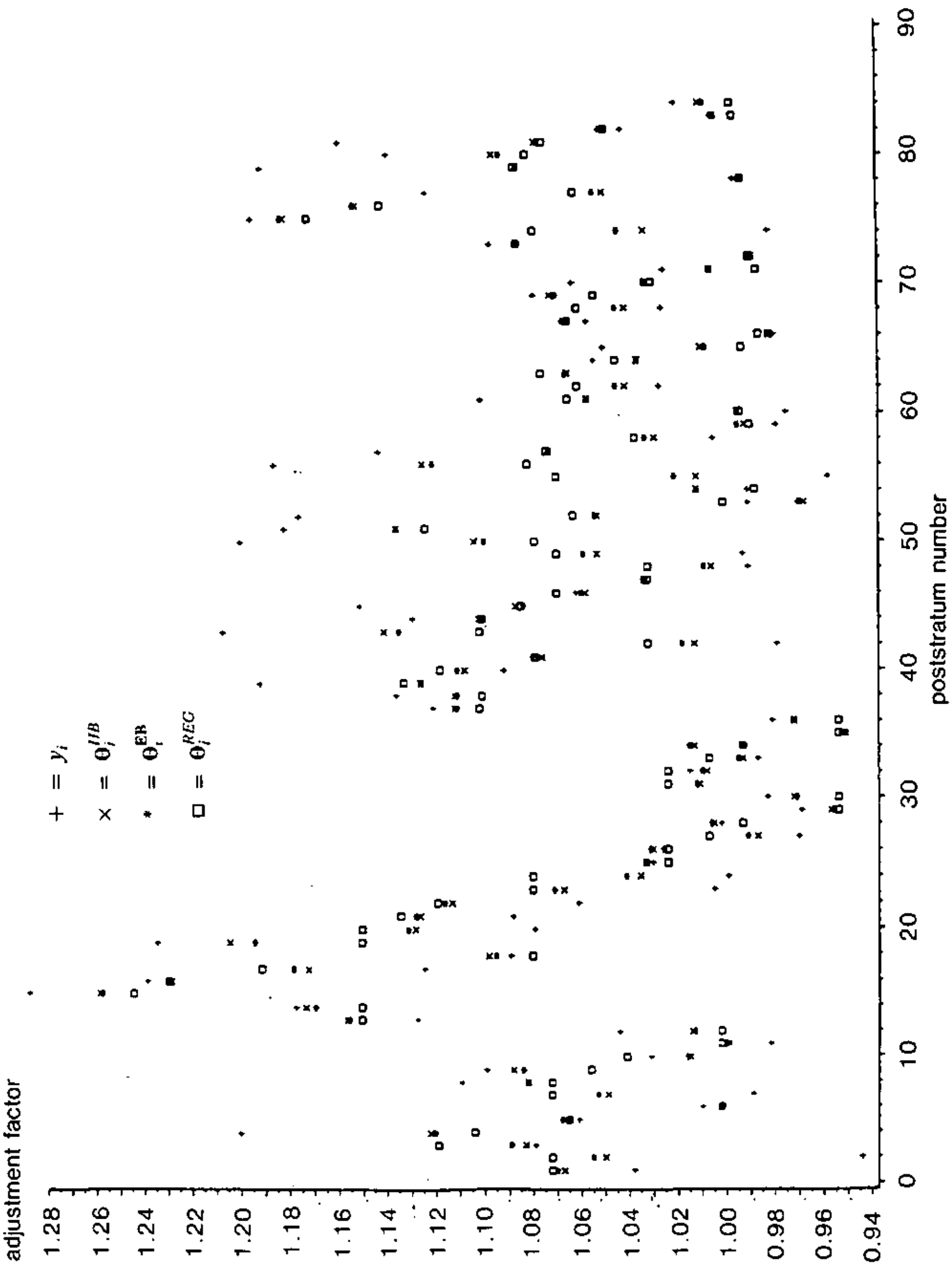

Figure 1. Adjustment Factors by Poststrata.

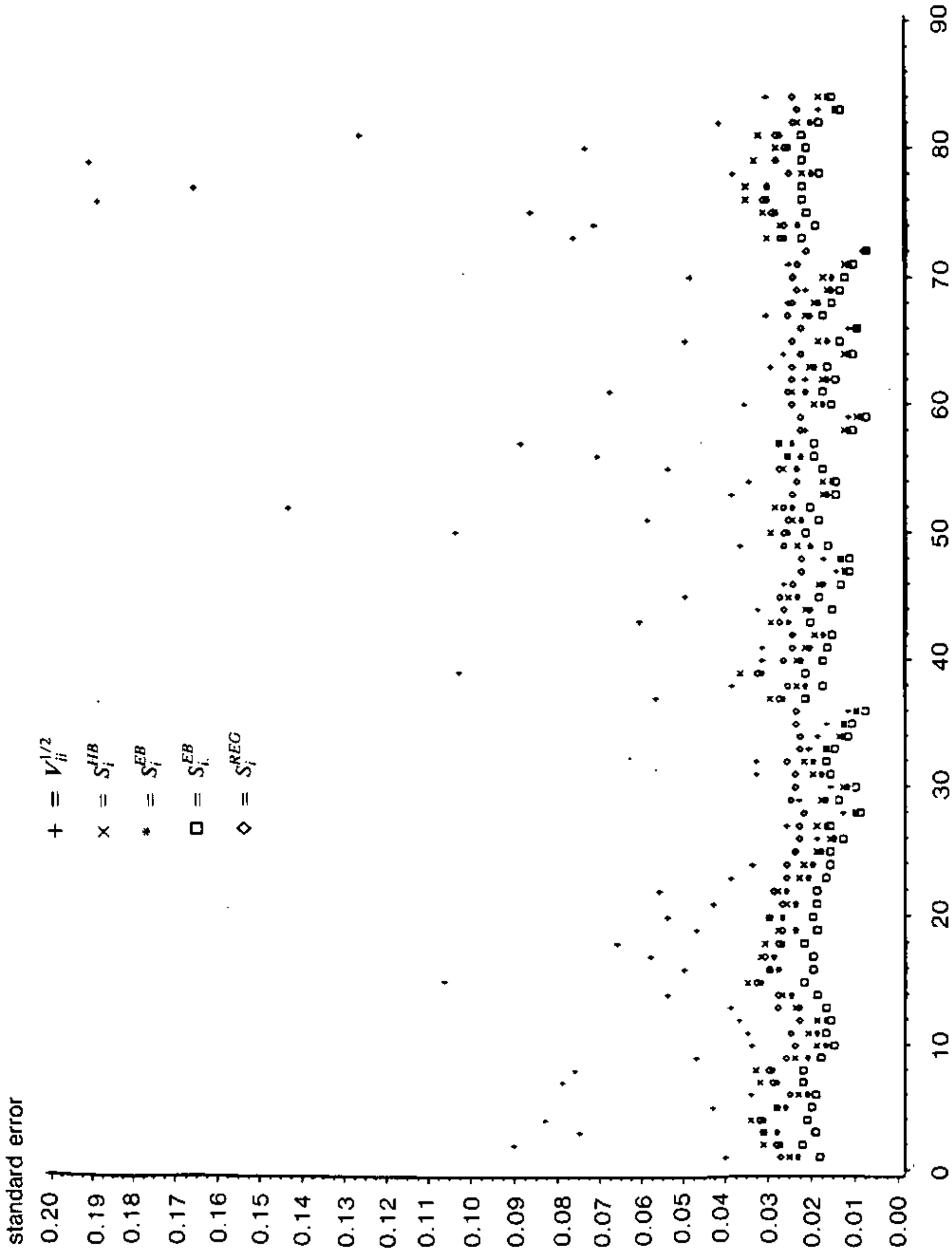

Figure 2. SE of Adjustment Factors by Poststrata.

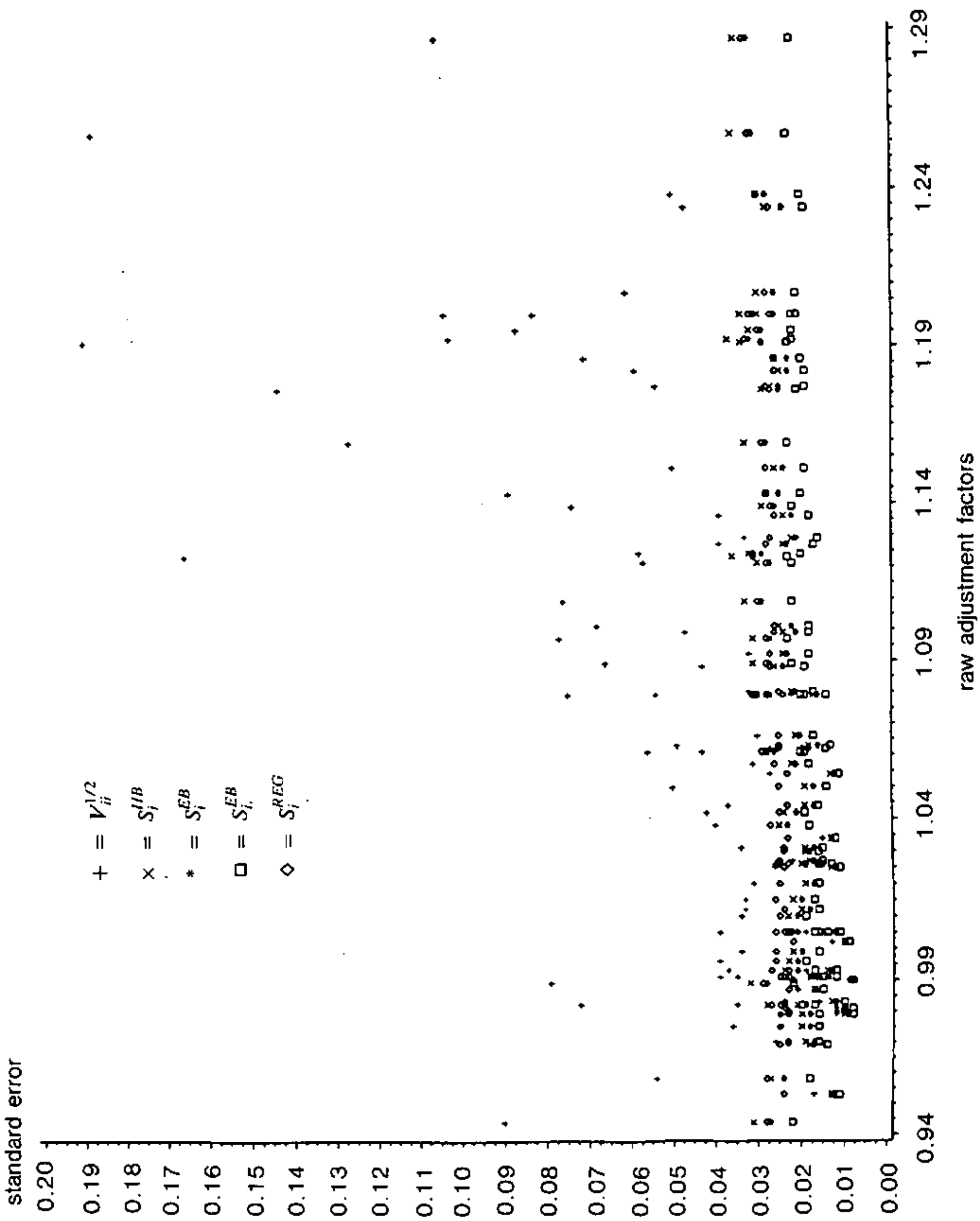

Figure 3. SE of Adjustment Factors by Raw Adjustment Factors.

Figures 1 to 3 lead to several interesting conclusions.

(1) For every stratum, the estimated standard errors of the HB and the EB estimators of the adjustment factors are much smaller than the standard errors of the raw adjustment factors when compared to the unadjusted DSE's.

(2) The EB estimators improve on the regression estimators for all the 84 strata by providing reduced estimated standard errors. Although the HB estimators do not improve on the regression estimators for all the strata, the improvement is substantial for most of the strata.

(3) The data plots demonstrate that the difference between the point estimates $\hat{\theta}_i^{EB}s$ and $\hat{\theta}_i^{HB}s$ is quite small. Indeed, the percentage difference is always less than (and most often far less than) 1%.

(4) The posterior standard errors associated with the HB estimates (s_i^{HB}) are always bigger than the approximate MSE's of the EB estimates (s_i^{EB}). As discussed earlier, the two need not be the same. It is our feeling that the approximate standard errors of the EB estimates are often slight underestimates. However, a comparison of s_i^{EB} and $s_{i.}^{EB}$ reveals that a naive EB procedure (with associated estimated standard errors $s_{i.}^{EB}$) can grossly underestimate the estimated standard errors by failing to incorporate uncertainty due to estimation of σ^2. This deficiency is largely rectified by s_i^{EB} which is based on second order approximations.

At the time of revision of this article, adjustment of the 1990 Decennial Census was completed. The EB estimation procedure was used. Basically, most of the same steps followed in modelling the adjustment factors in the 1988 Dress Rehearsal Census were used. However, there were several differences. In 1990 adjustment, the estimated adjustment factors were modelled by each of four census regions and a special set for Indian reservations. The number of adjustment factors ranged from 12 for the Indian set to 456 in one of the regions. In addition, estimated variances of the raw adjustment factors were smoothed via regression models. Smoothing of the estimated variances tended to reduce large estimated variances and increase small estimated variances. The net effect was an increase in the contribution of the associated adjustment factors with large estimated variances to the EB estimates and vice versa. Other differences were that outlier detection procedures were used in both the variance and adjustment factor smoothing. Finally, the EB estimates at the poststratum level were ratio adjusted to regional total population estimates derived from the raw adjustment factors. The ratio adjusted smoothed factors were then applied to related census population counts at the census block level. The results were then integer rounded by collection of blocks in such a manner that each cell within a block is rounded up or down to an integer and that control totals are off by at most one person.

The procedures used to adjust the 1990 Census counts were pre-specified and the entire operation was conducted under a very tight time schedule. The Bureau of the Census recommended that the 1990 Census adjusted counts be used. A special panel selected by the Secretary of Commerce was evenly divided in this issue. Upon weighing the evidence, the Secretary decided against using the adjusted counts. The issue is now subject to litigation. A current issue is the possible use of adjusted counts for use in postcensal estimation. Research in obtaining better adjusted counts for use in postcensal estimation is currently underway.

APPENDIX - PROOFS OF THE THEOREMS

Proof of Theorem 1. We provide only an outline of the proof. The details appear in Datta *et al.* (1991). The joint (improper) pdf of Y, Θ, β and σ^2 is given by:

$$f(y,\Theta,\beta,\sigma^2) \propto \exp\left[-1/2(y-\Theta)^T V^{-1}(y-\Theta)\right]\sigma^{-m}\exp\left[-1/(2\sigma^2)\|\Theta - X\beta\|^2\right], \quad \text{(A.1)}$$

where $\|\cdot\|$ denotes the Euclidean norm. Writing $P_X = X(X^TX)^{-1}X^T$, $\|\Theta - X\beta\|^2 = \left[\beta - (X^TX)^{-1}X^T\Theta\right]^T(X^TX)\left[\beta - (X^TX)^{-1}X^T\Theta\right] + \Theta^T(I - P_x)\Theta$.

Now, integrating with respect to β in (A.1), it follows that the joint improper pdf of Y, Θ and σ^2 is

$$f(y,\Theta,\sigma^2) \propto \sigma^{-(m-p)}\exp\left[-1/2(y-\Theta)^T V^{-1}(y-\Theta) - 1/(2\sigma^2)\Theta^T(I-P_x)\Theta\right]. \quad \text{(A.2)}$$

Next writing $E = V^{-1} + \sigma^{-2}(I - P_x)$, it follows after some simplifications that

$$(y-\Theta)^T V^{-1}(y-\Theta) + \sigma^{-2}\Theta^T(I-P_x)\Theta =$$

$$(\Theta - E^{-1}V^{-1}y)^T E(\Theta - E^{-1}V^{-1}y) + y^T(V^{-1} - V^{-1}E^{-1}V^{-1})y. \quad \text{(A.3)}$$

Hence, the posterior distribution of Θ given σ^2 and $Y = y$ is $\mathrm{N}(E^{-1}V^{-1}y, E^{-1})$. Using the familiar matrix inversion formula $(A + BDB^T)^{-1} = A^{-1} - A^{-1}B(D^{-1} + B^TA^{-1}B)^{-1}B^TA^{-1}$ (see for example Exercise 2.9, p. 33 of Rao (1973)), one gets $E^{-1} = G$. This completes the proof of the first part of the Theorem. Next, using (A.3) and integrating with respect to Θ in (A.2), one gets the joint (improper) pdf of Y and σ^2 is

$$f(y,\sigma^2) \propto \sigma^{-(m-p)}|E|^{-1/2}\exp\left[-(1/2)y^T(V^{-1} - V^{-1}E^1V^{-1})y\right]. \quad \text{(A.4)}$$

Using Exercise 2.4, p. 32 of Rao (1973), it follows that

$$|E| = |V^{-1} + \sigma^2(I - P_x)| = |\Delta| \div |\sigma^2X^TX|$$

which on simplification reduces to

$$|V^{-1}|\,|I + \sigma^{-2}V|\,|X^T(I+\sigma^{-2}V)^{-1}X| \div |X^TX| \propto |I + \sigma^{-2}V|\,|X^T\Sigma^{-1}X|. \quad \text{(A.5)}$$

Also, after some calculations, it follows that

$$V^{-1} - V^{-1}E^{-1}V^{-1} = F. \quad \text{(A.6)}$$

The proof of part (ii) of Theorem 1 follows now from (A.4) - (A.6) and noting that $f(\sigma^2 \mid y) \propto f(\sigma^2, y)$. Note, however that the posterior pdf of σ^2 given $Y = y$ is proper.

Proof of Theorem 2. Once again, only a sketch of the proof is given. The details are available in Datta *et al.* (1991).

Recall

$$\tilde{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y.$$

Define,

$$\hat{\boldsymbol{\theta}} = X\tilde{\beta} + \sigma^2 \Sigma^{-1}(Y - X\tilde{\beta}).$$

Now, observe that (i) $\hat{\boldsymbol{\theta}}$ is the best unbiased predictor of $\boldsymbol{\theta}$ (due to normality) for every fixed σ^2, and (ii) $\mathrm{E}(\hat{\boldsymbol{\theta}}^{\mathrm{EB}} - \hat{\boldsymbol{\theta}}) = \mathbf{0}$ since $\hat{\sigma}^2$ is the MLE of σ^2 (*cf* Kackar and Harville (1984)). Now using Lemma 3.3.1 of Datta (1990), $\hat{\boldsymbol{\theta}}^{\mathrm{EB}} - \hat{\boldsymbol{\theta}}$ is uncorrelated with $\hat{\boldsymbol{\theta}}$. Hence,

$$\mathrm{E}\left[(\hat{\boldsymbol{\theta}}^{\mathrm{EB}} - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{\mathrm{EB}} - \hat{\boldsymbol{\theta}})^T\right] =$$

$$\mathrm{E}\left[(\hat{\boldsymbol{\theta}}^{\mathrm{EB}} - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{\mathrm{EB}} - \hat{\boldsymbol{\theta}})^T\right] + \mathrm{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\right]. \tag{A.7}$$

Next, write $\hat{\boldsymbol{\theta}}^{\mathrm{B}} = X\beta + \sigma^2 \Sigma^{-1} (Y - X\beta)$. Then standard arguments give

$$\mathrm{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\right] = \mathrm{E}\left[(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{\mathrm{B}})(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{\mathrm{B}})^T\right] + \mathrm{E}\left[(\hat{\boldsymbol{\theta}}^{\mathrm{B}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}^{\mathrm{B}} - \boldsymbol{\theta})^T\right]. \tag{A.8}$$

Our previous calculations yield

$$\mathrm{E}\left[(\hat{\boldsymbol{\theta}}^{\mathrm{B}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}^{\mathrm{B}} - \boldsymbol{\theta})^T\right] = V - V\Sigma^{-1}V. \tag{A.9}$$

Further,

$$\mathrm{E}\left[(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{\mathrm{B}})(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{\mathrm{B}})^T\right] = V\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}V. \tag{A.10}$$

Finally, write $\hat{\boldsymbol{\theta}} = g(\sigma^2)$ and $\hat{\boldsymbol{\theta}}^{\mathrm{EB}} = g(\hat{\sigma}^2)$. Using first order Taylor approximation, one gets

$$\mathrm{E}\left[(\hat{\boldsymbol{\theta}}^{\mathrm{EB}} - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{\mathrm{EB}} - \hat{\boldsymbol{\theta}})^T\right] \doteq \mathrm{E}\left[(\hat{\sigma}^2 - \sigma^2)^2 \frac{dg(\sigma^2)}{d\sigma^2} \frac{dg(\sigma^2)^T}{d\sigma^2}\right]. \tag{A.11}$$

Since $g(\sigma^2) = Y - V\Sigma^{-1}[Y - X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y]$, using matrix differentiation, techniques, one gets

$$\frac{dg}{d\sigma^2} = V\left[\Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\right]\Sigma^{-1}(Y - X\hat{\beta}). \tag{A.12}$$

$$\mathrm{E}\left[\frac{dg}{d\sigma^2}\frac{dg^T}{d\sigma^2}\right] = VK\Sigma^{-1}\mathrm{E}\left[(Y - X\hat{\beta})(Y - X\hat{\beta})^T\right]\Sigma^{-1}KV. \tag{A.13}$$

But, simple algebra gives

$$\mathrm{E}\left[(Y - X\hat{\beta})(Y - X\hat{\beta})^T\right] = \Sigma - X(X^T\Sigma^{-1}X)^{-1}X^T = \Sigma K \Sigma. \tag{A.14}$$

Hence, from (A.13),

$$\mathrm{E}\left[\frac{dg}{d\sigma^2}\frac{dg^T}{d\sigma^2}\right] = VK^3V. \tag{A.15}$$

Using, one more approximation, it follows from (A.11) and (A.15) that

$$\mathrm{E}\left[(\hat{\boldsymbol{\theta}}^{\mathrm{EB}} - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{\mathrm{EB}} - \hat{\boldsymbol{\theta}})^T\right] \doteq \mathrm{E}(\hat{\sigma}^2 - \sigma^2)^2 VK^3V. \tag{A.16}$$

To estimate $\mathrm{E}(\hat{\sigma}^2 - \sigma^2)^2 = \mathrm{MSE}(\hat{\sigma}^2)$, we proceed as follows.

Since $Y \sim \mathrm{N}(X\beta, \Sigma)$, write the likelihood function as

$$L(\sigma^2) \propto |\,\Sigma\,|^{-1/2} \exp\left[-1/2(Y - X\beta)^T\Sigma^{-1}(Y - X\beta)\right]. \tag{A.17}$$

Hence,

$$\frac{d\log L}{d\sigma^2} = -1/2\,\frac{d}{d\sigma^2}\log|\,\Sigma\,| - 1/2\,\frac{d}{d\sigma^2}\left[(Y - X\beta)^T\Sigma^{-1}(Y - X\beta)\right]; \tag{A.18}$$

$$\frac{d^2\log L}{d(\sigma^2)^2} = -1/2\,\frac{d^2}{d(\sigma^2)^2}\log|\,\Sigma\,| - 1/2\,\frac{d^2}{d(\sigma^2)^2}\left[(Y - X\beta)^T\Sigma^{-1}(Y - X\beta)\right]. \tag{A.19}$$

As before, denote by $d_1, \ldots, d_m$ the eigenvalues of V.

Then, $\log|\,\Sigma\,| = \Sigma_{i=1}^m \log(\sigma^2 + d_i)$. Hence

$$\frac{d^2}{d(\sigma^2)^2}\log|\,\Sigma\,| = -\Sigma_{i=1}^m\left(\sigma^2 + d_i\right)^{-2} = -\operatorname{tr}\left(\Sigma^{-2}\right). \tag{A.20}$$

Using (A.20) and matrix differentiation, it follows from (A.19) that

$$\frac{d^2\log L}{d(\sigma^2)^2} = 1/2\operatorname{tr}\left(\Sigma^{-2}\right) - \left(Y - X\beta\right)^T\Sigma^{-3}\left(Y - X\beta\right). \tag{A.21}$$

Thus,

$$\mathrm{E}\left[-\frac{d^2\log L}{d(\sigma^2)^2}\right] = -1/2\operatorname{tr}\left(\Sigma^{-2}\right) + \operatorname{tr}\left(\Sigma^{-2}\right) = 1/2\operatorname{tr}\left(\Sigma^{-2}\right).$$

Approximating $\mathrm{E}[(\hat{\sigma}^2 - \sigma^2)^2]$ by

$$\left(\mathrm{E}\left[-\frac{d^2\log L}{d(\sigma^2)^2}\right]\right)^{-1},$$

justifiable by the asymptotic theory of maximum likelihood, one gets, from (A.16),

$$\mathrm{E}\left[(\hat{\boldsymbol{\theta}}^{\mathrm{EB}} - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{\mathrm{EB}} - \hat{\boldsymbol{\theta}})^T\right] \doteq 2\left(\operatorname{tr}(\Sigma^{-2})\right)^{-1} VK^3V. \tag{A.22}$$

Combining (A.7) - (A.10) and (A.22), one gets

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}^{\mathrm{EB}}) \doteq V - V\Sigma^{-1}V + V\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}V + VK^3V\left[2(\operatorname{tr}\Sigma^{-2})^{-1}\right] \tag{A.23}$$

Substitution of $\hat{\Sigma}$ for Σ yields the approximation given in (2.8). This completes the proof of Theorem 2.

ACKNOWLEDGEMENTS

This paper is based upon work supported by the National Science Foundation under grant SES 87-13643, "On-Site Research to Improve the Government-Generated Social Science Data Base." The research was conducted at the U.S. Bureau of the Census while the first two authors were participants in the American Statistical Association/Census Bureau Research Program, which is supported by the Census Bureau and through the NSF grant. Any opinions, findings and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Bureau of the Census.

REFERENCES

CHILDERS, D.R., and HOGAN, H. (1990). Results of the 1988 dress rehearsal post enumeration survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 547-552.

DATTA, G.S. (1990). Bayesian prediction in mixed linear models with applications in small area estimation. Unpublished Ph.D. dissertation. University of Florida.

DATTA, G.S., GHOSH, M., HUANG, E.T., ISAKI, C.T., and SCHULTZ, L.K. (1990). Hierarchical and Empirical Bayes methods for adjustment of census undercount: The 1988 Missouri Dress-Rehearsal data. Technical Report No. 376. Department of Statistics, University of Florida.

ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond (with discussion). *Journal of the American Statistical Association*, 80, 98-131.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for Small Places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census (with discussion). *Statistical Science*, 1, 3-39.

HARVILLE, D.A. (1985). Decomposition of prediction error. *Journal of the American Statistical Association*, 80, 132-138.

HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a Post-Enumeration Survey. *Survey Methodology*, 14, 99-116.

HUANG, E.T., ISAKI, C.T., and TSAY, J.H. (1991). Modelling PES adjustment factors using 1988 dress rehearsal data. *Proceedings of the Social Statistics Section, American Statistical Association*, to appear.

ISAKI, C.T., HUANG, E.T., and TSAY, J.H. (1991). Smoothing adjustment factors from the 1990 post enumeration survey. *Proceedings of the Social Statistics Section, American Statistical Association*, to appear.

KACKAR, R.N., and HARVILLE, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79, 853-862.

MORRIS, C. (1983). Parametric Empirical Bayes inference and applications. *Journal of the American Statistical Association*, 78, 47-65.

PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of mean squared error of small area predictors. *Journal of the American Statistical Association*, 85,163-171.

RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*, (2nd. Edition). New York: Wiley.

Survey Methodology, June 1992
Vol. 18, No. 1, pp. 109–125
Statistics Canada

A Comparison of Some Estimators of a Set of Population Totals

DON ROYCE[1]

ABSTRACT

The Population Estimates Program of Statistics Canada has traditionally been benchmarked to the most recent census, with no allowance for census coverage error. Because of a significant increase in the level of undercoverage in the 1986 Census, however, Statistics Canada is considering the possibility of adjusting the base population of the estimates program for net census undercoverage. This paper develops and compares four estimators of such a base population: the unadjusted census counts, the adjusted census counts, a preliminary test estimator, and a composite estimator. A generalization of previously-proposed risk functions, known as the Weighted Mean Square Error (WMSE), is used as the basis of comparison. The WMSE applies not only to population totals, but to functions of population totals such as population shares and growth rates between censuses. The use of the WMSE to develop and evaluate small-area estimators in the context of census adjustment is also described.

KEY WORDS: Census adjustment; Undercoverage; Small area estimation.

1. INTRODUCTION

The Population Estimates Program of Statistics Canada provides a wide variety of detailed information about the characteristics and distribution of the Canadian population during the five-year period between each census. Intercensal estimates of population have many important uses, including the calculation of billions of dollars of transfer payments from the federal to provincial governments, the estimation of important demographic statistics such as birth and mortality rates, the planning of future levels of immigration, and the weighting of current population surveys such as the monthly Labour Force Survey.

Traditionally, the estimates program is based on the most recent census, with no allowance for coverage error. In the 1986 Census, however, undercoverage increased significantly compared to previous censuses, and continued to be distributed unevenly across geographic and demographic groups. This caused considerable disruption to the estimates program and to the many other programs which use population estimates. As a result, a project was initiated in early 1989 to investigate whether, and if so how, the population estimates in the post-1991 Census period should be adjusted for estimated census coverage error. The research described in this paper was conducted as part of this project. For a more general description of the project, see Royce (1992).

It should be noted that only the population estimates would be affected by this adjustment. The 1991 Census data will be published with no adjustment for undercoverage, other than the small adjustments that have traditionally been made to correct for underenumeration of temporary residents and for persons missed because their dwelling was misclassified as vacant. From the technical point of view, however, the question is quite similar to the issue of census adjustment that has been of interest to many statistical agencies in recent years.

Two key questions in the adjustment issue are the degree to which census counts are improved by adjustment, and which adjustment methods are best. In this paper, we compare the accuracy of several different estimators of a set of population totals, using a weighted mean square error as our criterion.

[1] Don Royce, Chief, Census Data Quality Section, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

Section 2 introduces the topic by considering the simple case of a single population total. We derive and compare the Mean Square Errors of four possible estimators: the unadjusted census count, the adjusted census count, a preliminary test estimator, and a composite estimator. Section 3 extends the results to multiple population totals and to functions of population totals, such as population shares and growth rates. In Section 4, we consider methods for small-area estimation, specifically, the use of synthetic estimation and a special case of synthetic estimation known as across-the-board adjustment. Section 5 concludes with a description of areas for further research.

In developing the estimators described in this paper, two assumptions were made. First, adjustment must result in estimates that are consistent across all geographic and demographic levels, as well as across time. Users consider it to be essential that parts add up to totals, and that there be no major breaks in the time series of estimates. Second, adjustment will be based on the combined results of Statistics Canada's two coverage measurement studies: the Reverse Record Check, which measures gross undercoverage, and the Overcoverage Study, which measures gross overcoverage. Both studies are subject to sampling errors and non-sampling errors.

2. SINGLE POPULATION TOTAL

We first describe and compare four estimators for the case of a single population total. In comparing the estimators, we use the Mean Square Error (MSE) as our criterion.

Let: Y be the known census count;
T be the unknown true population total to be estimated;
U be the true net undercoverage, *i.e.*, $U = T - Y$;
$\hat{U}$ be an estimate of U from the coverage measurement studies;
σ^2 be the variance of $\hat{U}$; and
R be the relative bias of $\hat{U}$, *i.e.* $R = E(\hat{U})/U - 1$.

In the case of all four estimators, our estimate of T can be written as the census count plus some estimate of U. Thus, the MSE of our estimate of T will be the same as that of the corresponding estimate of U. The MSEs (and the WMSEs in later sections) are taken over hypothetical repetitions of the coverage measurement studies, treating the Census counts as fixed quantities.

2.1 Unadjusted Census Estimator

The unadjusted census estimate of U is zero. It has a bias equal to $-U$ and zero variance. Therefore $\text{MSE}(\hat{U}^c) = U^2$.

2.2 Adjusted Census Estimator

The adjusted census estimator of U is $\hat{U}$. It has a bias of UR and a variance equal to σ^2. Thus $\text{MSE}(\hat{U}^A) = \sigma^2 + U^2R^2$.

2.3 Preliminary Test Estimator

A comparison of the MSEs of the previous two estimators suggests that we would use the adjusted census count in preference to the unadjusted census count whenever

$$\sigma^2 < U^2(1 - R^2). \tag{1}$$

Although the parameters in this inequality are unknown, they can (with the exception of R) be estimated from the coverage measurement studies. This suggests the possibility of using these estimates to develop a statistical test of the hypothesis that the inequality holds. The result of the test is then used to choose which estimator to use (thus the term preliminary test, or pretest, estimator).

Specifically, assume that $|R| < 1$, (obviously necessary for (1) to hold) and $\hat{U} \sim N(U(1 + R), \sigma^2)$, where σ^2 is known. Then $\hat{U}^2/\sigma^2$ has a non-central $\chi^2_{(1)}$ distribution with non-centrality paramater $\lambda = U^2(1 + R)^2/2\sigma^2$. The null hypothesis $H_0 : \sigma^2 \geq U^2(1 - R^2)$ is equivalent to the hypothesis $H_0 : \lambda \leq (1 + R)/2(1 - R)$. One approach, therefore, could be to adjust whenever $\hat{U}^2/\sigma^2 > c$, where the critical value $c \geq 0$ is chosen so that

$$\alpha = \Pr\left\{\chi^2\left(1, \frac{1+R}{2(1-R)}\right) \geq c\right\}, \tag{2}$$

where α is the significance level of the test. This is a special case of a more general test suggested by Toro-Vizcorrondo and Wallace (1968).

Note that $\hat{U}^2/\sigma^2$ is the inverse of the square of the estimated coefficient of variation (CV) of $\hat{U}$. Thus, the criterion for adjustment can be interpreted in terms of a requirement to have a sufficiently small (in absolute value) CV.

In practice, we would have to substitute some prior estimate of the relative bias, say r, for R in (2). The sensitivity of c to various values of R is examined in Royce (1991) for the case of a one-sided test (a normal distribution was used instead of a χ^2 in this case). For example, with a significance level of 2.5%, it was found that the critical CV was only reduced from 33.8% to 27.1% even when the relative bias was as much as 50%.

If σ is not known but an estimate $\hat{\sigma}$ is available, then a similar test can still be constructed by assuming that

$$\frac{\nu\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(\nu)} \tag{3}$$

independent of $\hat{U}$. This leads to a test based on a non-central F distribution. Further details on the construction of such tests are given in Judge and Bock (1978).

In order to determine the MSE of the preliminary test estimator, we note that it can be written as $\hat{U}^P = I\hat{U}$ where

$$\begin{aligned} I &= 1 \quad \text{if} \quad \frac{\hat{U}^2}{\sigma^2} > c \\ &= 0 \quad \text{if} \quad \frac{\hat{U}^2}{\sigma^2} \leq c. \end{aligned} \tag{4}$$

When σ^2 is known, the MSE of this estimator can be shown to be (see, for example, Judge and Bock 1978, p. 72)

$$\begin{aligned} \text{MSE}(\hat{U}^P) = \sigma^2 + U^2R^2 + (2U^2(1 + R) - \sigma^2)\Pr\{\chi^2_{(3,\lambda)} \leq c\} - \\ U^2(1 + R)^2\Pr\{\chi^2_{(5,\lambda)} \leq c\}. \end{aligned} \tag{5}$$

Note that as $c \to \infty$, *i.e.* as the chance of adjustment goes to zero, the MSE approaches U^2, the MSE of the unadjusted census. Similarly, as $c \to 0$, *i.e.* as the chance of adjustment goes to certainty, the MSE approaches $\sigma^2 + U^2R^2$, the MSE of the adjusted census estimator. Thus, the two previous approaches of adjustment or no adjustment can be interpreted as extreme cases of the pretest estimator procedure.

Figure 1 shows MSE/σ^2 for the preliminary test estimator as a function of U^2/σ^2, for various values of c, in the unbiased case ($R = r = 0$). The MSEs/σ^2 of the unadjusted census and the adjusted census are also shown. In all cases, the MSE of the preliminary test estimator starts out higher than that of the unadjusted census, crosses the MSE of the adjusted census, reaches a maximum, and then approaches the MSE of the adjusted census. As the value of c decreases and the level of significance α of the test therefore increases, the MSE of the preliminary test estimator approaches that of the adjusted census more quickly, but at the expense of being higher for small values of U^2/σ^2. Thus, the performance of the preliminary test estimator over the range of possible values of U^2/σ^2 depends on the level of significance that is chosen for the test.

Figures 2 and 3 show similar plots in the case where $R = .5$ and $R = -.5$ respectively (since we may feel we have no information on which to base an estimate of R, we have set $r = 0$). Again, the MSEs of the preliminary test estimators approach those of the adjusted census as U^2/σ^2 increases. With a positive bias the MSE of the preliminary test estimator approaches the MSE of the adjusted census more quickly than in the unbiased case, while for a negative bias the reverse is true.

What is the "best" value of c for the test? Ideally, we would like to choose c so that the MSE of the preliminary test estimator is as close as possible to the minimum of the MSEs of the adjusted census and the unadjusted census. One approach, due to Sawa and Hiromatsu (1973) and extended by Brook (1976), is to minimize the maximum difference between the MSE of the preliminary test estimator and the minimum of the MSEs of the adjusted census and unadjusted census. For the unbiased case this criterion gives an optimal value of c of approximately 1.88. This corresponds to a critical CV (in absolute value) for the estimated undercoverage of 73%. The MSE of this estimator is shown in Figure 4.

Judge and Bock (1978) also describe other approaches to choosing the optimal value of c, such as minimizing the average distance (rather than the maximum difference) and Bayesian approaches.

2.4 Composite Estimator

The preliminary test estimator was written as $\hat{U}^P = I\hat{U}$, where I took on only the values 0 or 1. Because of this inherent discontinuity, however, it has been shown that the preliminary test estimator is inadmissible (Cohen 1965). As an alternative, we might consider letting the multiplier of $\hat{U}$ take any value between 0 and 1. That is, instead of using the data to tell us **whether** to adjust, we use the data to tell us **how much** to adjust. This type of estimator has been suggested by Spencer (1980) and more recently by Andrews (1991). We define $\hat{U}^\alpha = \alpha\hat{U}$ where $0 \le \alpha \le 1$. For a given alpha, this estimator has a MSE equal to

$$\mathrm{MSE}(\alpha\hat{U}) = \alpha^2\sigma^2 + U^2(\alpha(1 + R) - 1)^2, \tag{6}$$

which is minimized when

$$\alpha = \frac{U^2(1 + R)^2}{(1 + R)(\sigma^2 + U^2(1 + R)^2)}. \tag{7}$$

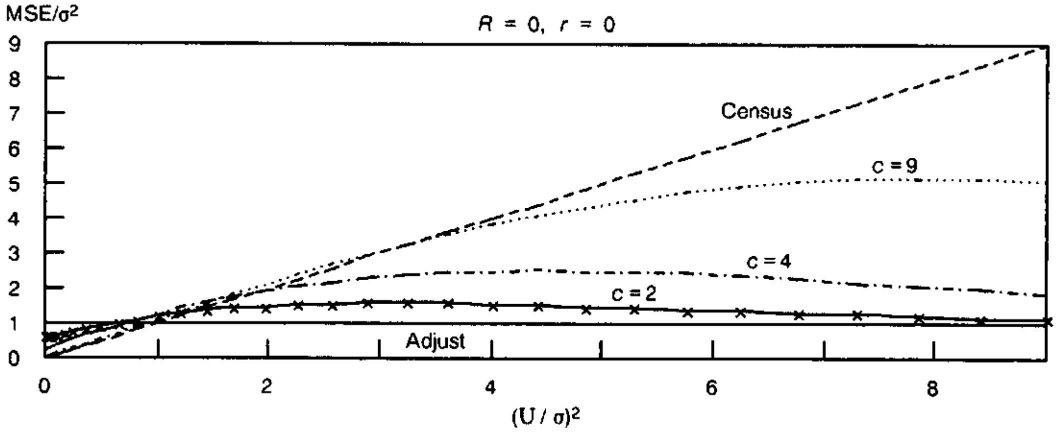

Figure 1 Comparison of MSEs

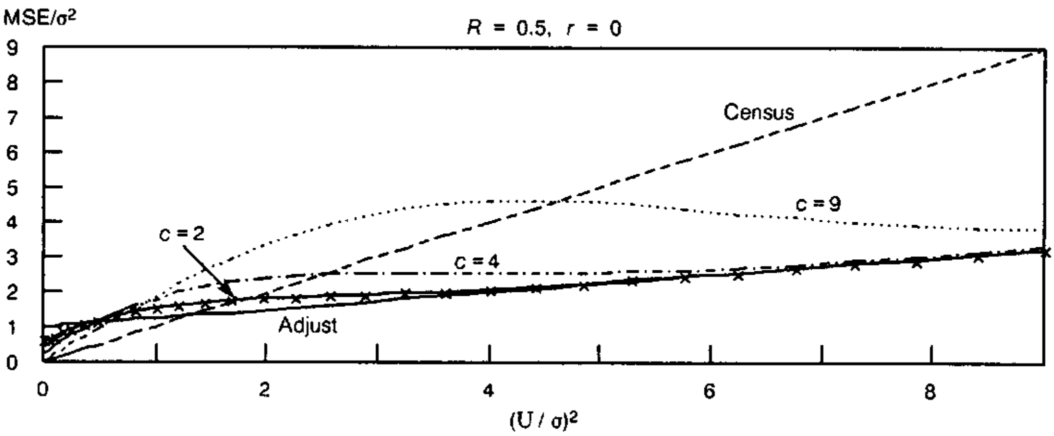

Figure 2 Comparison of MSEs

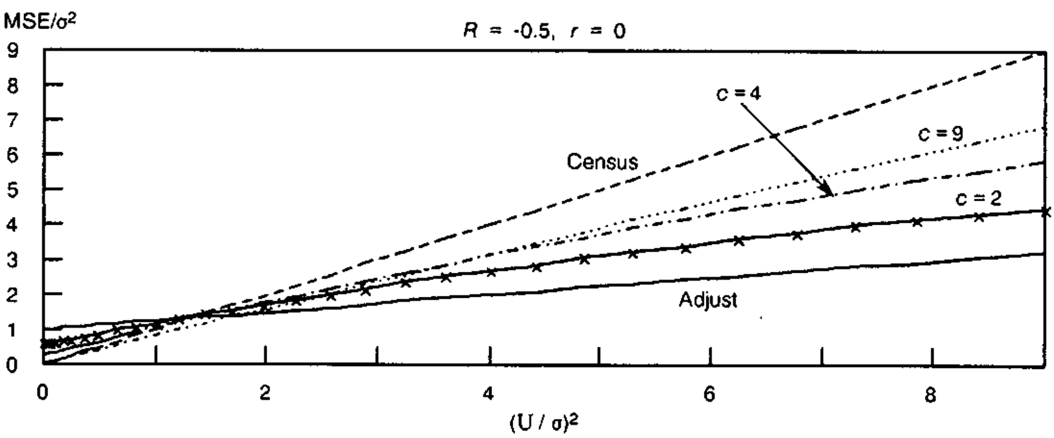

Figure 3 Comparison of MSEs

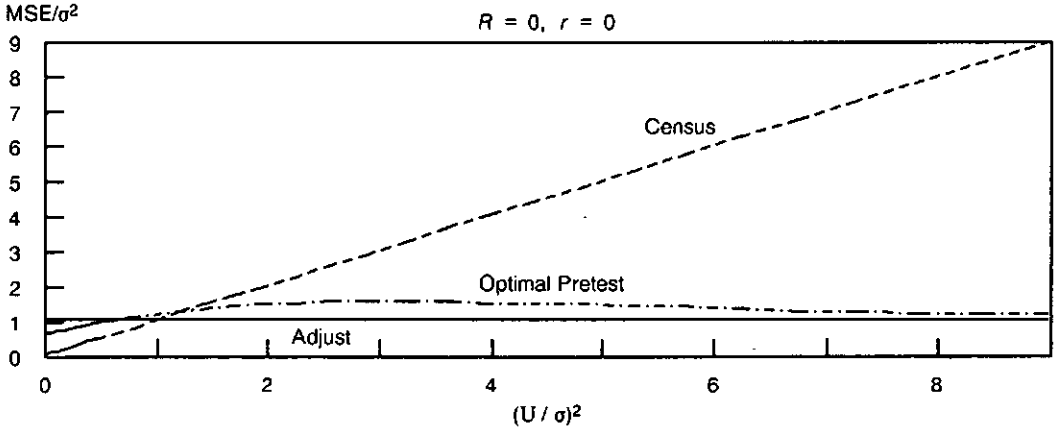

Figure 4 Comparison of MSEs

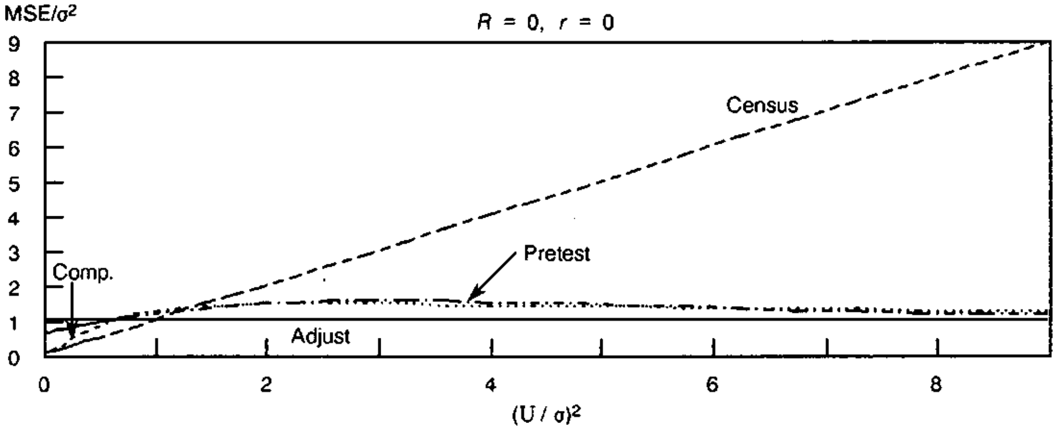

Figure 5 Comparison of MSEs

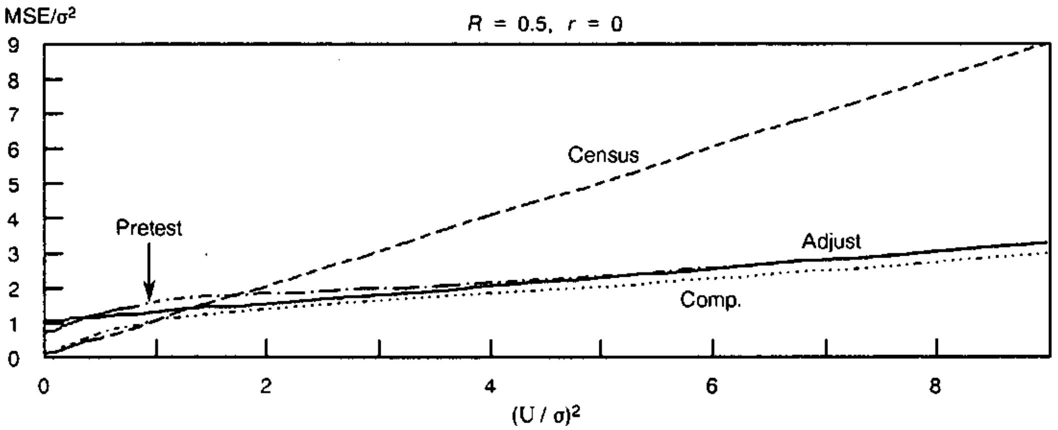

Figure 6 Comparison of MSEs

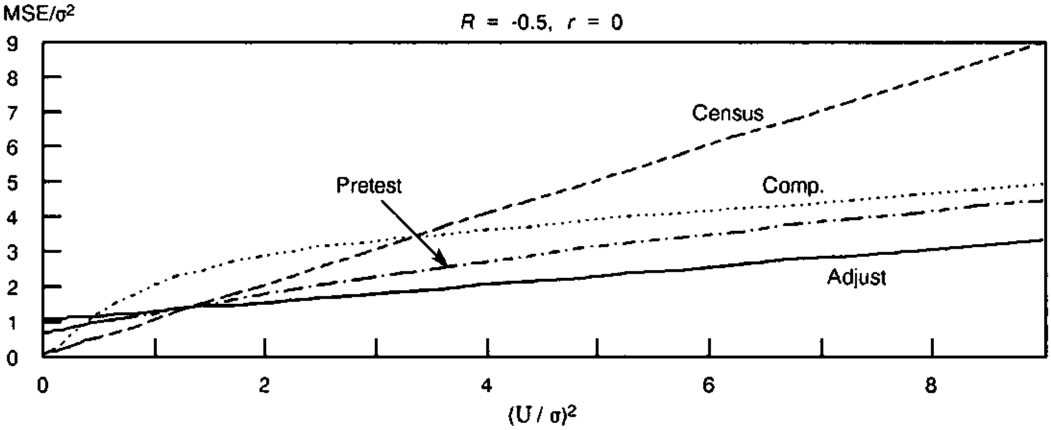

Figure 7 Comparison of MSEs

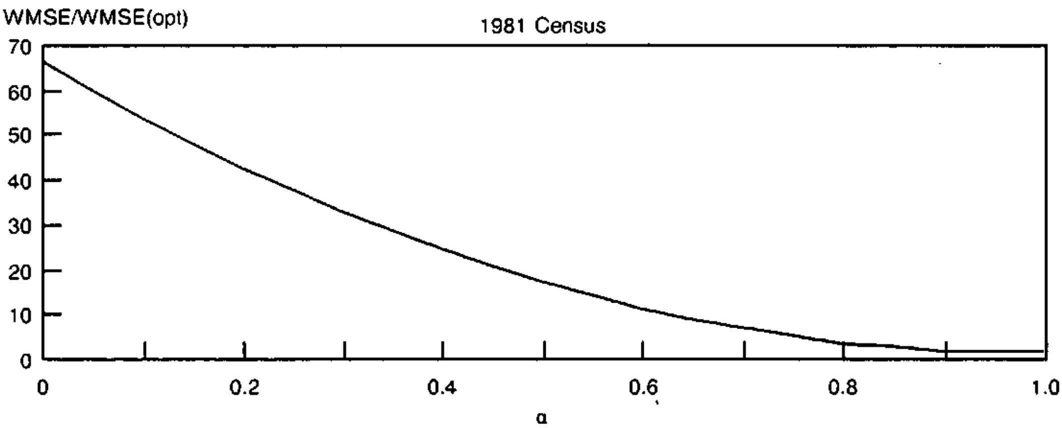

Figure 8 WMSEs for Totals

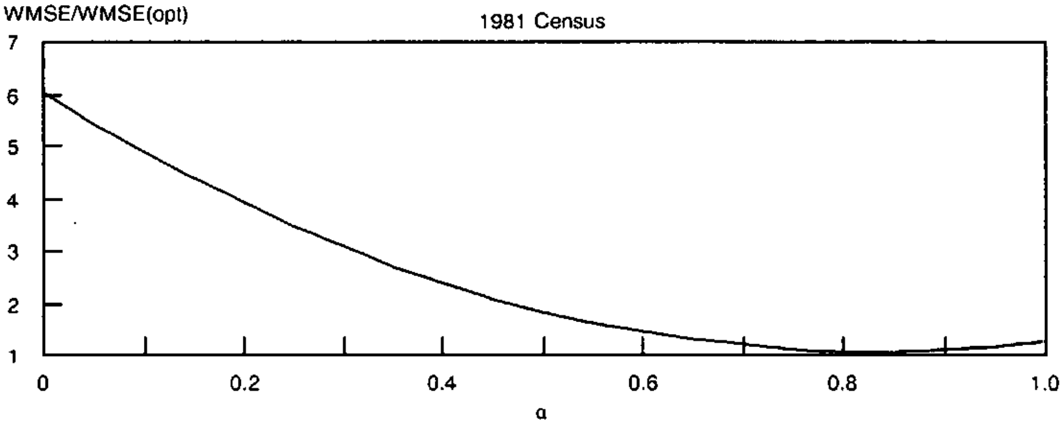

Figure 9 WMSEs for Shares

If σ is assumed known, then a possible estimator of α is

$$\hat{\alpha} = \frac{\hat{U}^2}{(1 + r)(\sigma^2 + \hat{U}^2)} \tag{8}$$

and thus

$$\hat{U}^{\hat{\alpha}} = \frac{\hat{U}^3}{(1 + r)(\sigma^2 + \hat{U}^2)}. \tag{9}$$

The approximate MSE of this estimator can be found using a Taylor series approximation. Letting

$$h(U,\sigma^2) = \frac{U^3}{(1 + r)(\sigma^2 + U^2)} \tag{10}$$

we get (dropping terms higher than those involving the first derivative)

$$\mathrm{MSE}(\hat{U}^{\hat{\alpha}}) \doteq (h(U,\sigma^2) - U)^2 + \left(\frac{\partial h(U,\sigma^2)}{\partial U}\right)^2 (U^2R^2 + \sigma^2)$$

$$+ 2(h(U,\sigma^2) - U)\left(\frac{\partial h(U,\sigma^2)}{\partial U}\right) UR. \tag{11}$$

This approximation can also be extended to the case where σ is unknown by making the assumption given in (3). The MSE is then increased by the additional term

$$\left(\frac{\partial h(U,\sigma^2)}{\partial \sigma^2}\right)^2 \frac{2\sigma^4}{\nu}. \tag{12}$$

Figures 5, 6 and 7 show the MSE of the composite estimator as a function of U^2/σ^2, as well as the MSEs of the unadjusted census, adjusted census, and the optimal preliminary test estimator from Section 2.3. In the unbiased case (Figure 5) and the positive bias case (Figure 6), the composite estimator outperforms the optimal preliminary test estimator. When the bias is negative, however, (Figure 7) the MSE of the composite estimator can be much higher than any of the other estimators over a considerable portion of the range of U^2/σ^2.

3. MORE GENERAL ESTIMATORS

In this section, we generalize the four estimators examined in Section 2 in two ways. First, instead of a single population total, we consider a vector of population totals, denoted as $\underline{T} = (T_1, T_2, \ldots, T_N)$. Second, we consider not only the population totals themselves, but also functions of the population totals, denoted by $\underline{g}(\underline{T}) = (g_1(\underline{T}), g_2(\underline{T}), \ldots, g_K(\underline{T}))$ where in general $K \neq N$. Typical functions of interest include population shares, used in the transfer of funds from the federal to provincial governments, as well as growth rates between censuses, differences in growth rates among different provinces, and so on.

In evaluating the overall accuracy of some estimate $\underline{g}(\underline{\hat{T}^*})$ for $\underline{g}(\underline{T})$, we will make use of a loss function. The use of loss functions for evaluating the effects of census adjustment is described in Fellegi (1980), Citro and Cohen (1985), Spencer (1986), and Wolter and Causey (1991) to name just a few. The specific loss function used in this paper is a generalization of previously-proposed loss functions for population totals and shares. Specifically, the risk (expected loss) of the estimator $\underline{g}(\hat{\underline{T}}^*)$ is the Weighted Mean Square Error, defined as

$$\text{WMSE}\,(\underline{g}(\hat{\underline{T}}^*)) = E\left\{\sum_{k=1}^{K} w_k\left(g_k(\hat{\underline{T}}^*) - g_k(\underline{T})\right)^2\right\}, \tag{13}$$

where w_k is a user-specified weight reflecting the importance of the k-th component of the loss function.

Since $\underline{g}$ may be complex in practice, it is useful to work instead with an approximation to the WMSE derived by expanding $\underline{g}(\hat{\underline{T}}^*)$ in a Taylor series around $\underline{T}$. This yields:

$$\text{WMSE}\underline{g}(\hat{\underline{T}}^*) \doteq \sum_{i=1}^{N}\sum_{j=1}^{N} \omega_{ij}[\text{Cov}(\hat{U}_i^*, \hat{U}_j^*) + \text{Bias}(\hat{U}_i^*)\text{Bias}(\hat{U}_j^*)] \tag{14}$$

where the weight ω_{ij} is given by

$$\omega_{ij} = \sum_{k=1}^{K} w_k \frac{\partial g_k}{\partial T_i}\frac{\partial g_k}{\partial T_j}. \tag{15}$$

(Note that the approximate WMSE can also be written as the expectation of the quadratic form $(\hat{\underline{U}}^* - \underline{U})'\Omega(\hat{\underline{U}}^* - \underline{U})$ where ω_{ij} is the ij-th element of Ω.)

This formulation conveniently splits each component of the risk function into two parts: a weight ω_{ij} that depends only on the w_k and the function $\underline{g}$, and the portion in square brackets which depends only on the particular estimator being used.

While the choice of the w_k can be arbitrary, considerations of equity have often led to the choice $w_k = 1/T_k$. In the case of population totals and shares, for example, the risk function (14) then becomes equivalent to those proposed by Fellegi (1980) and also used by Wolter and Causey (1991), among others. Other choices for the weights that have been suggested in the literature include $w_k = 1/Y_k$, $w_k = 1/\hat{T}_k$, and $w_k = 1$. For further discussion on the merits of these various weightings, see the references cited above. Table 1 shows some examples of ω_{ij} for different functions.

In the case of population growth rates, the first pair of subscripts on the omega refer to the population quantity of interest (*e.g.* province) while the second pair refer to the census at time 1 or time 2 respectively. The second subscript on the T_i also refer to the census at time 1 or 2.

In the remainder of this section, we illustrate the use of the WMSE in developing and evaluating the unadjusted census, adjusted census, preliminary test estimator, and composite estimator.

3.1 Unadjusted Census

The WMSE of the unadjusted census is $\text{WMSE}(\underline{\hat{U}^C}) = \sum_{ij}\omega_{ij}U_iU_j$.

3.2 Adjusted Census

The WMSE of the adjusted census is $\text{WMSE}(\underline{\hat{U}^A}) = \sum_{ij}\omega_{ij}[\sigma_{ij} + b_ib_j]$ where $\sigma_{ij} = \text{Cov}(\hat{U}_i,\hat{U}_j)$ and $b_i = \text{Bias}(\hat{U}_i)$.

Table 1

Examples of Weights ω_{ij} in the Approximate WMSE for Various Functions

Function	ω_{ij}
Set of Population Totals	$\omega_{ii} = w_i$ $\omega_{ij} = 0 \qquad i \neq j$
Set of Population Shares	$\omega_{ii} = \frac{1}{T^4}\left(\sum_k w_k T_k^2 + w_i T^2 - 2w_i T T_i\right)$ $\omega_{ij} = \frac{1}{T^4}\left(\sum_k w_k T_k^2 - T(w_i T_i + w_j T_j)\right) \qquad i \neq j$
Set of Growth Rates	$\omega_{ii11} = \frac{w_i T_{i2}^2}{T_{i1}^4}$ $\omega_{ii12} = -\frac{w_i T_{i1} T_{i2}}{T_{i1}^4} = \omega_{ii21}$ $\omega_{ii22} = \frac{w_i T_{i1}^2}{T_{i1}^4}$ $\omega_{ij11} = \omega_{ij12} = \omega_{ij21} = \omega_{ij22} = 0 \qquad i \neq j$

3.3 Preliminary Test Estimator

As in Section 2.3, we would use the adjusted census in preference to the unadjusted census if the WMSE of the adjusted census is less than the WMSE of the unadjusted census, *i.e.*, if

$$D = \sum_{ij} \omega_{ij}[U_i U_j - \sigma_{ij} - b_i b_j] > 0. \tag{16}$$

Tests for this type of hypothesis were suggested by Fellegi (1980) for the specific cases of population totals and population shares, but the ideas generalize quite readily to any function $\underline{g}$. The left hand side of the inequality (16) is estimated by $\hat{D} = \sum_{ij} \omega_{ij}[\hat{U}_i \hat{U}_j - 2\sigma_{ij}]$ where the ω_{ij} are assumed to be known. (In practice the ω_{ij} are estimated by substituting either the census counts or the adjusted census counts in (13). Fellegi claimed that minor variations in the weights were unlikely to substantially change the test results.) It is then easy to show that $E(\hat{D}) = D + 2\sum_{ij} \omega_{ij} b_i (U_j + b_j)$. For the case of totals and shares, Fellegi presented arguments why it could be assumed that the second term was non-positive, *i.e.*, $\sum_{ij} \omega_{ij} b_i (U_j + b_j) \leq 0$ so that $\hat{D}$ would tend to underestimate D. Fellegi also derived an approximate variance for $\hat{D}$. This, along with the assumption that $\hat{D}$ was normally distributed, permitted the construction of a test for the hypothesis given in (16).

Table 2

z Values for Fellegi's Tests for Adjustment of Provincial Population Totals and Shares, Reverse Record Check, 1976, 1981 and 1986

Function	1976	1981	1986
Totals	9.3	10.1	13.1
Shares	3.1	1.8	1.5

In the more general case, the approximate variance of $\hat{D}$ is given by $\text{Var}(\hat{D}) \doteq 4\sum_{ij}\sigma_{ij}(\sum_{i'j'}\omega_{ij'}\omega_{i'j}U_{i'}U_{j'})$. An estimate of $\text{Var}(\hat{D})$ can then be derived by substituting estimates of the U_i and σ_{ij} in this formula.

In the case of totals, for example, the test statistic (z value) is given by

$$z = \frac{\hat{D}}{\sqrt{\hat{\text{Var}}(\hat{D})}} = \frac{\sum_i \frac{\hat{U}_i^2 - 2\sigma_i^2}{Y_i}}{2\sqrt{\sum_i \frac{\hat{U}_i^2\hat{\sigma}_i^2}{Y_i^2}}}, \tag{17}$$

where in this case the inverse of the census counts have been used as the weights. A similar expression can be derived for population shares.

Table 2 shows the z values calculated for the censuses of 1976, 1981 and 1986 for provincial population totals and shares. The data come from the Reverse Record Checks conducted in these censuses.

The case for adjusting population totals is much stronger than the case for adjusting shares, reflecting the fact that estimates of differences in undercoverage rates among provinces are less accurate than estimates of the undercoverage rates themselves. Further numerical results are given in Royce and Luc (1990).

3.4 Composite Estimator

A natural extension of the composite estimator of Section 2.4 would at first seem to be $\alpha_i\hat{U}_i$. However the use of different amounts of adjustment for each value of i introduces problems of consistency. For example, it would imply that more adjustment should be done at the Canada level than at the province level, since the estimates of undercoverage at the province level will be less accurate than for the national level. If this were done, the provincial totals would not add up to the Canada total.

In practice, therefore, we constrain ourselves to a single value of alpha, *i.e.* $\underline{\hat{U}}^\alpha = \alpha\underline{\hat{U}}$, where again $0 \leq \alpha \leq 1$. The WMSE of this estimator is

$$\text{WMSE}(\underline{\hat{U}^\alpha}) = \sum_{ij} \omega_{ij}[\alpha^2(\sigma_{ij} + b_ib_j) + (\alpha - 1)^2U_iU_j + 2\alpha(\alpha - 1)U_ib_j], \tag{18}$$

which is minimized when

$$\alpha = \frac{\sum_{ij} \omega_{ij} U_i(U_j + b_j)}{\sum_{ij} \omega_{ij}[\sigma_{ij} + (U_i + b_i)(U_j + b_j)]}. \tag{19}$$

If, as was done in Section 3.3, we make the assumption that $\sum_{ij}\omega_{ij}b_i(U_j + b_j) \leq 0$ then a lower bound for the optimal alpha is given by

$$\alpha_L = \frac{\sum_{ij} \omega_{ij}(U_i + b_i)(U_j + b_j)}{\sum_{ij} \omega_{ij}[\sigma_{ij} + (U_i + b_i)(U_j + b_j)]}, \tag{20}$$

which we estimate by

$$\hat{\alpha}_L = \frac{\sum_{ij} \omega_{ij}\hat{U}_i\hat{U}_j}{\sum_{ij} \omega_{ij}[\hat{\sigma}_{ij} + \hat{U}_i\hat{U}_j]}, \tag{21}$$

assuming the ω_{ij} are known. In practice, as we did for the preliminary test estimator, we would estimate the ω_{ij} by substituting census counts or adjusted census counts in (15).

In the case of population totals, for example, the estimated amount of adjustment is

$$\hat{\alpha}_L = \frac{\sum_{i} \hat{T}_i\hat{U}_i^2}{\sum_{i} \hat{T}_i[\hat{\sigma}_i^2 + \hat{U}_i^2]}, \tag{22}$$

where $\hat{U}_i$ is the estimated undercoverage **rate**, *i.e.* $\hat{U}_i/(Y_i + \hat{U}_i)$, and $\hat{\sigma}_i^2$ is its estimated variance.

For shares, the amount of adjustment is given by

$$\hat{\alpha}_L = \frac{\sum_{i} \hat{T}_i\hat{U}_i^2 - \hat{T}\hat{U}^2}{\sum_{i} \hat{T}_i[\hat{\sigma}_i^2 + \hat{U}_i^2] - \hat{T}(\hat{\sigma}^2 + \hat{U}^2)}, \tag{23}$$

where $\hat{U}$ is the estimated undercoverage rate for the total population, *i.e.* $\sum_i\hat{U}_i/\sum_i(Y_i + \hat{U}_i)$ and $\hat{\sigma}^2$ is its estimated variance. The inverse of the adjusted census counts have been used as the weights in these two examples.

3.5 Numerical Comparisons

In the case of a single population total, it was possible to derive exact or approximate formulae for the MSEs of the four estimators as a function of U^2/σ^2, R, r and (in the case of the preliminary test estimator), the critical value of the test. Unfortunately, it has not yet been possible to derive similar expressions for the WMSEs of complex functions of a vector of population totals.

In the case of the unadjusted census, adjusted census, and composite estimator, however, it is possible to estimate the WMSEs by substituting estimates of undercoverage and their estimated variances into equation (18) (if estimates of the bias terms are available they can be used, but in what follows we assume they are zero). For example, Figures 8 and 9 show, for the 1981 Census, the estimated ratio of the WMSE to the optimal WMSE, as a function of α, where the provinces are again the units indexed by i. The extremes of $\alpha = 0$ and $\alpha = 1$ correspond to the unadjusted and adjusted census counts respectively, while the minimum point on the curve corresponds to the optimal α. Figure 8 is for totals and Figure 9 is for shares. The optimum values of α were computed using formulae (22) and (23).

In each case, the optimal degree of adjustment is close to 1.0, and results in a WMSE considerably lower than the WMSE corresponding to no adjustment (*e.g.* by a factor of almost 70 for totals). The optimal degree of adjustment is less for shares than for totals, again reflecting the fact that estimates of differences in coverage rates between provinces are less accurate than the estimates of the rates themselves. It is also interesting to note that the WMSE for full adjustment is only slightly higher than that of the optimal degree of adjustment. This can have important practical significance, since it is much easier to explain a full adjustment to data users than to explain a partial adjustment.

4. SMALL AREA ESTIMATION

The previous two sections considered the case where direct estimates of undercoverage, and estimates of their variances, were available from the coverage measurement studies. This situation applies, for example, for provinces, for some major Census Metropolitan Areas, and for broad demographic groups (*e.g.* age by sex, age by marital status) at the national level. However the Population Estimates Program produces estimates at very detailed levels, such as single years of age by sex by marital status for some 260 Census Divisions. Direct estimates of undercoverage generally do not exist at such levels.

Nevertheless, the need to maintain consistency of the estimates requires that any adjustment made at a higher level be "carried down" to the detailed levels used by the estimates program. In this section, we consider the use of synthetic estimation for this purpose, and show how the WMSE can again be used to develop preliminary test estimators and composite estimators.

The synthetic estimator is based on the assumption that net undercoverage is uniform within each of a number of "adjustment groups", indexed by a. The synthetic estimate is then given by $\hat{U}_i^S = \sum_a \lambda_{ia} \hat{U}_a$ where $\lambda_{ia} = Y_{ia}/Y_a$. For example, the adjustment groups might correspond to age-sex groups, for which estimates of undercoverage $\hat{U}_a$ are available at some higher level.

A special case of the synthetic estimator arises when there is only one adjustment group. Wolter and Causey (1991) have called this the across-the-board estimator. It is defined as $\hat{U}_i^{ATB} = \lambda_i \hat{U}$ where $\lambda_i = Y_i/Y$. WMSEs for the across-the-board and the synthetic estimator can be derived using equation (14). Since the ω_{ij} do not depend on the particular estimator used, only the portion in square brackets changes. Table 3 compares the estimators of U_i and their covariance and bias terms for the census, adjusted census, across-the-board and synthetic estimators.

Table 3

Examples of Covariance and Bias Terms in the Approximate WMSRE for Various Estimators

Estimator	$\hat{U}_i^*$	$\mathrm{Cov}(\hat{U}_i^*, \hat{U}_j^*)$	$\mathrm{Bias}(\hat{U}_i^*)$
Census	0	0	$-U_i$
Adjusted Census	$\hat{U}_i$	σ_{ij}	b_i
Across-the-Board	$\lambda_i \hat{U}$	$\lambda_i \lambda_j \sigma^2$	$\lambda_i(U+b) - U_i$
Synthetic	$\sum_a \lambda_{ia} \hat{U}_a$	$\sum_{aa'} \lambda_{ia'} \lambda_{ja'} \sigma_{aa'}$	$\sum_a \lambda_{ia}(U_a + b_a) - U_i$

where $b = \sum_i b_i$ and similarly b_a is the bias of $\hat{U}_a$.

4.1 Preliminary Test Estimators

As was the case in Sections 2 and 3, the WMSE can be used to develop statistical tests to decide between two competing estimators. As an example, consider the situation where we wish to choose between the unadjusted census and the across-the-board estimator for population totals (shares are of course unchanged by across-the-board adjustment). On comparing the WMSEs of these two estimators, we find that we would use the across-the-board estimator in preference to the census counts if

$$\sigma^2 < U^2(1 - R^2)\left[1 - \frac{2TB}{U(1 - R)}\right], \tag{24}$$

where

$$B = 1 - \frac{1}{\sum_i \frac{\lambda_i^2}{\tau_i}} \tag{25}$$

and $\tau_i = T_i/T$. This condition was given, in a different form, by Wolter and Causey (1991). B is a measure of the heterogeneity of undercoverage; it is non-negative, and is equal to zero if and only if the undercoverage is completely uniform.

Noting that this inequality is the same as (1) except for the additional term in square brackets, we can derive a test very similar to the test described in Section 2.3. The critical value of the coefficient of variation will depend on the chosen significance level and the relative bias as before, but will also depend on $B/\bar{U}$, the ratio of the heterogeneity of undercoverage to the overall undercoverage rate.

Royce (1991) showed that, in practice, the effect of this additional factor on the critical CV was likely to be negligible. Thus, if adjustment is justified at some higher level, then carrying down the adjustment to lower levels is almost certainly justified as well. Similar results were found in a simulation study reported by Wolter and Causey (1991).

4.2 Composite Estimators

In Sections 2 and 3 we considered composite estimators where the two extremes were the unadjusted census and the adjusted census. With the addition of the synthetic and across-the-board estimators, the number of possible composite estimators increases considerably. For example, we might consider composite estimators involving the unadjusted census and the synthetic estimator, the adjusted census and the across-the-board estimator, the across-the-board estimator and the synthetic estimator, and so on. Consequently, we present below a method which can be used to derive a composite estimator involving any two estimators.

Our general composite estimator is defined as $\underline{\hat{U}^*} = \alpha\underline{\hat{U}_1} + (1-\alpha)\underline{\hat{U}_2}$ where $\underline{\hat{U}_1}$ and $\underline{\hat{U}_2}$ are two estimators. The WMSE of this estimator is

$$\text{WMSE}(\underline{g}(\underline{\hat{T}^*})) = \alpha^2\text{WMSE}(\underline{g}(\underline{\hat{T}_1})) + (1-\alpha)^2\text{WMSE}(\underline{g}(\underline{\hat{T}_2})) + 2\alpha(1-\alpha)\text{WMXPE}(\underline{g}(\underline{\hat{T}_1},\underline{g}(\underline{\hat{T}_2})), \tag{26}$$

where

$$\text{WMXPE}(\underline{g}(\underline{\hat{T}_1},\underline{g}(\underline{\hat{T}_2})) = \sum_{ij} \omega_{ij}[\text{Cov}(\hat{U}_{1i},\hat{U}_{2j}) + \text{Bias}(\hat{U}_{1i})\text{Bias}(\hat{U}_{2j})] \tag{27}$$

is defined to be the Weighted Mean Cross-Product Error of $\underline{g}(\underline{\hat{T}_1})$ and $\underline{g}(\underline{\hat{T}_2})$. The WMSE of our composite estimator is minimized when

$$\alpha = \frac{\text{WMSE}(\underline{g}(\underline{\hat{T}_2})) - \text{WMXPE}(\underline{g}(\underline{\hat{T}_1}),\underline{g}(\underline{\hat{T}_2}))}{\text{WMSE}(\underline{g}(\underline{\hat{T}_1})) + \text{WMSE}(\underline{g}(\underline{\hat{T}_2})) - 2\text{WMXPE}(\underline{g}(\underline{\hat{T}_1}),\underline{g}(\underline{\hat{T}_2}))}. \tag{28}$$

To obtain an estimate of α, we substitute estimates of the WMSEs and the WMXPE into the above.

As an example of how this approach could be used, suppose a decision has been taken to adjust a provincial population total. To carry down the adjustment, we might consider using either across-the-board adjustment (*i.e.* adjust all sub-provincial quantities by the same factor), or a synthetic adjustment, where the adjustment is done separately within several age-sex groups. The across-the-board method has the advantage that it uses only the provincial estimate of undercoverage, which is likely to be more reliable than the estimates of undercoverage by age and sex at the province level. On the other hand, if undercoverage varies considerably among age and sex groups, and if the sub-provincial quantities indexed by i also differ in their age-sex composition, then the synthetic estimator may be better.

If estimates of the U_i are available from some source, then all covariance and bias components of the WMSEs and the WMXPE can be estimated (using formulae such as those in Table 3), and the optimum composite estimator involving the across the board and synthetic estimators can be estimated. Although for sub-provincial quantities the U_i will not usually exist in practice, the method can be investigated at higher levels. For example we could use the provinces as the quantities indexed by i and use across-the-board and synthetic adjustment factors computed at the Canada level. A second possibility is to construct an artificial population (*e.g.* as in Shirm and Preston (1987) or Wolter and Causey (1991)) where the U_i are assumed to be known.

5. FURTHER WORK

The results presented in this paper represent only a start to the investigation and comparison of the performance of various estimators of a set of population totals. There are several areas where considerable work is yet required.

First, further investigation of the WMSEs for the preliminary test and composite estimators in the more general cases described in Sections 3 and 4 is required. Although attempts to derive analytic expressions for these WMSEs have not yet been successful, the more general results for preliminary test estimators and Stein-rule estimators described by Judge and Bock (1978) may yet be found to apply. If so, this would help to answer questions such as: Can optimal critical values be found for the Fellegi-type preliminary test estimators of Sections 3.3 and 4.1? How does the WMSRE of the preliminary test estimator compare in practice to those of the other three estimators?

Second, more work is needed to explore the sensitivity of the results to different weightings in the loss function. The results of Section 3 were based on the use of a weight equal to the inverse of the census count or the adjusted census count for each province. If the provinces had been weighted differently, the results would change. A more general weight we might want to consider is $w_k = Y_k^\gamma$, where γ is some type of power parameter. The sensitivity of the results in Section 3 to various of values of γ could then be studied.

Finally, while the methods described in this paper provide a framework for developing and evaluating various estimators, the exact manner in which the methods will be applied has yet to be decided. Specific issues that must be resolved include:

1. What is the relative importance of different types of functions such as totals, shares and growth rates? Different functions give rise to different results, but in the end a single estimator must be chosen in order to maintain consistency.

2. At what geographic and demographic levels should these methods be applied? For example, should the preliminary test estimator or composite estimator described in Section 3 be applied at the province level, at the province by age group and sex level, or at even more detailed levels? The results obtained depend on the level of analysis used.

3. Could we even consider composite estimators for "high profile" estimators such as the provincial population totals? It might be difficult to explain to users why the adjustments do not coincide with the published estimates of undercoverage.

Because the resolution of issues such as these will require professional judgement, the decision about whether to adjust (and how to adjust) cannot be an automatic one based on completely pre-specified criteria. While the methods described in this paper can provide useful guidance, the final decision will require a careful balancing of the potential improvement in the accuracy of the estimates with consideration of how easily the methods can be communicated to and understood by users of the estimates program.

ACKOWLEDGEMENTS

I would like to thank the editor, the two referees, and Richard Carter for several valuable comments that helped improve the quality of this paper.

REFERENCES

ANDREWS, D. (1991) Invited discussion at the Meeting of Statistics Canada's Advisory Committee on Statistical Methods, October 1991.

BROOK, R.J. (1976). On the use of a regret function to set significance points in prior tests of estimation. *Journal of the American Statistical Association*, 71, 353, 126-131.

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

COHEN, A. (1965). Estimates of the linear combination of parameters in the mean vector of a multivariate distribution. *Annals of Mathematical Statistics*, 36, 78-87.

FELLEGI, I.P. (1980). Should the census count be adjusted for allocation purposes? Equity considerations. In *Proceedings of the 1980 Conference on Census Undercount*, U.S. Bureau of the Census.

JUDGE, G.G., and BOCK, M.E. (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam-New York-Oxford: North-Holland Publishing Company.

ROYCE, D., and LUC, M. (1990). Recalculation of Fellegi's test statistics on census adjustment for the 1981 and 1986 censuses. Internal Statistics Canada report.

ROYCE, D. (1991). Technical criteria for adjusting the population estimates program for census coverage error. Internal Statistics Canada report.

ROYCE, D. (1992). Incorporating estimates of census coverage error into the Canadian population estimates program. *Proceedings of the Eighth Annual Research Conference*, Bureau of the Census, Washington, DC (to appear).

SAWA, T., and HIROMATSU, T. (1973). Minimax regret significance points for a preliminary test in regression analysis. *Econometrica*, 41, 1093-1101.

SHIRM, A.L., and PRESTON, S.H. (1987). Census undercount adjustment and the quality of geographic population distributions. *Journal of the American Statistical Association*, 82, 400, 965-978.

SPENCER, B. (1980). Implications of equity and accuracy for undercount adjustment: A decision-theoretic approach. In *Proceedings of the 1980 Conference on Census Undercount*, Bureau of the Census, Washington, DC.

SPENCER, B. (1986). Conceptual issues in measuring improvement in population estimates. In *Proceedings of the Second Annual Research Conference*, Bureau of the Census, Washington, DC, 393-407.

TORO-VIZCORRONDO, C., and WALLACE, T.D. (1968). A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association*, 63, 322, 558-572.

WOLTER, K.M., and CAUSEY, B.D. (1991). Evaluation of procedures for improving population estimates for small areas. *Journal of the American Statistical Association*, 86, 414, 278-284.

Survey Methodology, June 1992
Vol. 18, No. 1, pp. 127–141
Statistics Canada

The Creation of a Residential Address Register for Coverage Improvement in the 1991 Canadian Census

L. SWAIN, J.D. DREW, B. LAFRANCE and K. LANCE[1]

ABSTRACT

The Address Register is a frame of residential addresses for medium and large urban centres covered by Geography Division's Area Master File (AMF) at Statistics Canada. For British Columbia, the Address Register was extended to include smaller urban population centres as well as some rural areas. The paper provides an historical overview of the project, its objective as a means of reducing undercoverage in the 1991 Census of Canada, its sources and product, the methodology required for its initial production, the proposed post-censal evaluation and prospects for the future.

KEY WORDS: Address Register; Census undercoverage; Geographical Information Systems (GIS).

1. OBJECTIVE

The concept of an Address Register at Statistics Canada dates back to the 1960s. Fellegi and Krótki (1967) first considered building one for the 1971 Census using administrative source files as the base. Their approach was mostly manual and yielded a very complete set of addresses with minimal undercoverage and overcoverage. In the mid-1970s (Booth 1976), the idea resurfaced in planning for the 1981 Census. This time the approach started with data capture of addresses from the previous Census and was augmented with information from Canada Post. In both cases, the generated address lists were being considered as a frame for a mail-out Census. However, costs of creation were high and would have needed offsetting reductions in other Census operations to be effective. In addition, the risks associated with changing the traditional enumeration method were considered too great. As a result, the construction of an Address Register was suspended in each case.

A renewed interest in the concept of an Address Register emerged from the International 1991 Census Planning Conference (Royce 1986, 1987) in October 1985. This interest derived from the potential for automation of Fellegi and Krótki's approach due to technological developments, such as the availability of machine readable administrative files with addresses and postal codes and the development of in-house software to parse addresses into standard components, to assign postal codes and to link postal codes to Census geography. It followed as well from the development of a statistical theory for record linkage (Fellegi and Sunter 1969) and computer systems based on this theory (Hill and Pring-Mill 1985).

As a result, a project was initiated in 1986 with the first research (Gamache-O'Leary *et al.* 1987) investigating the use of an Address Register for a mail-out Census rather than the traditional drop-off approach. It concluded that the new Census data collection approach would be less expensive only if the quality of the Address Register required minimal field updating prior to the Census. Two small pilot registers created in early 1987 put Address Register coverage at 90-95%, which was unacceptable without field updating (Drew *et al.* 1987), ruling out the use of an Address Register for a mail-out Census.

---

[1] L. Swain and B. Lafrance, Social Survey Methods Division; J.D. Drew, Household Surveys Division; K. Lance, Geography Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

However, the two pilot registers revealed the potential for an Address Register to aid in coverage improvement when used in conjunction with the traditional drop-off methodology. This fitted well with the emergence of coverage improvement as one of the top priorities for the 1991 Census. The results of the Reverse Record Check for the 1986 Census had indicated a dramatic rise in the undercoverage rate compared to previous Censuses (from 2.01% in 1981 to 3.21% in 1986 for the national total population; from 2.08% in 1981 to 3.28% in 1986 for the national urban population) (Statistics Canada 1990). It was therefore decided that *the research project should concentrate on the development of the Address Register to use in coverage improvement of the 1991 Census.*

The next section describes the two major tests conducted to develop and refine the procedures used to create the Address Register for the 1991 Census. As well, the second section outlines the joint agreement with the Province of British Columbia to extend the Address Register. The third section presents the administrative and geographic sources used in the production process and the structure and content of the Address Register booklets, the end product used by Census Representatives in the field. The fourth section describes the methodology used to exploit the sources in order to produce the Address Register booklets. In the fifth section, the proposed post-censal evaluation is discussed while the last section presents future prospects for the Address Register. A separate future report will detail an evaluation of the methodology.

2. BACKGROUND

2.1 The November 1987 Test of Coverage Improvement Methods

A substantial test of the use of the Address Register (AR) as a coverage improvement tool was conducted in November 1987 in five large Regional Office cities. It was designed to estimate both undercoverage and overcoverage of dwelling units for the traditional Census method of listing and for two experimental methods of using an AR for Census coverage improvement: Post-list and Pre-list. The Post-list approach had the enumerator compile the dwelling list in the usual Census manner (creating a Visitation Record) then reconcile it with a dwelling list for the Enumeration Area (EA) derived from the AR. Field follow-ups were done where necessary on any address discrepancies between lists. In the Pre-list method, the enumerator was given the AR in advance and updated it during a canvass of the EA to create the final dwelling list.

The results (van Baaren 1988) concluded that the Post-list method was the more effective in improving coverage. This approach as a simple add-on to the standard Census enumeration process was fail-safe. If for some reason we failed to produce the AR (either in whole or in part) on time for the 1991 Census, the AR reconciliation step could simply be dropped without affecting the traditional enumeration process. The test data also provided estimates of the degree of coverage improvement and costs (Royce and Drew 1988). It was estimated that 34,000 occupied dwellings and 68,000 persons would be added by the AR to the medium and large urban centres for which it would be constructed (these urban centres representing those areas for which an Area Master File exists, *i.e.*, covering about 65% of the Canadian population). This would represent an improvement in coverage of 0.26 percentage points (the national undercoverage rate in 1986 being estimated as 3.21 percent). Relative to the two previous attempts at AR construction, costs were demonstrated to be low to the Census due to the highly automated approach and the proven benefit. As well, the risk was minimized since the traditional data collection method would still be used. Based on this cost, benefit and risk assessment, approval was given for creation of an AR for the 1991 Census.

From the November 1987 test, two concerns presented themselves. First, the ordering of the addresses in the AR booklets produced for each Enumeration Area (EA) didn't correspond to the order in the Visitation Records which made reconciliation a tedious and time-consuming task. Second, the overall overcoverage at 17% still seemed too high and more effort was required to eliminate erroneously placed or duplicate records. Both these problems were addressed by improving the methods for matching the AR to Census geography. Instead of linking addresses merely to EAs as had been done for the November test, procedures were developed to match the AR to the Area Master File (AMF) (Statistics Canada 1988) blockfaces. An algorithm was developed to sort addresses by block and within block in the same order they would be encountered by the enumerator in walking around the EA.

2.2 The September 1989 Test to Refine Procedures

Another substantial test was conducted in September 1989 involving four cities of various sizes: Moncton, Laval, Brampton and Calgary. Each was chosen because of unique difficulties that could arise based on the November 1987 test. The results (Dick 1990) showed a significant decrease in coverage from 84% in the 1987 test to 73%, a discouraging outcome. On the other hand, this test revealed a considerable reduction in overcoverage down from 17% to 8%. Importantly, despite the reduced coverage of the AR, its performance as a coverage improvement tool for the Census was still viable. On analysis, the new geocoding operation was found to be problematic, both in terms of its high costs, since it involved a great deal of clerical intervention, and in terms of its quality. The geocoding steps were therefore revamped for production, a key aspect of which was the adoption of CANLINK record linkage software (Statistics Canada 1989b) to improve quality and reduce costs of the AR/AMF linkage.

2.3 Joint Agreement with the Province of British Columbia

The Ministry of Finance and Corporate Relations in British Columbia was concerned about the high rate of undercoverage in their province in the 1986 Census (4.49% in 1986, up from 3.16% in 1981, for the provincial total population) (Statistics Canada 1990). Statistics Canada entered into a joint agreement with the Planning and Statistics Division (the provincial statistical agency) of the Ministry to help reduce undercoverage in British Columbia in the 1991 Census. Within this contract, the Address Register was expanded to include smaller urban areas in British Columbia, thereby increasing the population covered from 62% to 88%.

3. SOURCES AND PRODUCT

Production started in April 1990 and ended with the final Address Register (AR) booklet stapled in mid-May 1991, when 22,756 booklets had been compiled containing 6.6 million addresses for use in the Census data collection process.

3.1 Administrative Sources

In the September 1989 test, it was concluded that wherever possible the following four administrative sources ought to be used as sources of addresses to create the AR: telephone company billing files, municipal assessment rolls, hydro company billing files and the T1 Personal Income Tax file. However, the use of all four sources was possible only in Nova Scotia, New Brunswick, and eight major urban centres in Ontario (Ottawa, Toronto, Brampton, Etobicoke, London, Mississauga, Hamilton and Windsor). Because of the multiplicity of files, the cost of files and refusals, only three sources were used for Newfoundland, Québec, Manitoba, Alberta (telephone, hydro and tax files) and for Regina and the rest of Ontario (telephone, assessment and tax files). For Saskatoon, only telephone and tax files were available. The primary source files used by the British Columbia government were those of telephone and hydro, though motor vehicles, cable and Elections files were also used.

3.2 Geography Sources

In building the AR, extensive use was made of a Geography Division system and files.

i. The Area Master File (AMF) (Statistics Canada 1988) is a digitized feature network (covering streets, railroads, rivers, *etc.*) for medium and large urban areas, generally with populations of 50,000 or more. Of interest for the AR were the street features which contained street name and civic number ranges which could be used to locate individual addresses onto a blockface, the primary linkage.

ii. The Computer Assisted Mapping System (CAM) orders blockfaces into blocks and blocks into a Census Enumeration Area (EA). CAM was used for the sequencing of addresses in the AR booklets. The EA maps produced by CAM were used by the Census Representatives for the 1991 Census. For the AR, the maps for all AMF areas were used in the second clerical operation.

iii. The 1990 Postal Code Conversion File (PCCF) (Statistics Canada 1991) is a national file of all postal codes, each of which is linked to a 1986 Census EA or a series of 1986 EAs. This input was used for secondary linkage of addresses at the EA level.

iv. The 1986/1991 EA Correspondence File relates the 1986 EA geography to the 1991 geography. This file was used for the secondary linkage at the EA level and the second clerical operation.

3.3 Address Register Booklets

The end product consisted of a set of booklets of residential addresses, one for each Enumeration Area, covering urban areas of Canada for which an Area Master File existed. Figure 1 contains a fictitious example of a page from an AR booklet (reduced in size).

Each booklet was divided into two sections: a structured portion and an unstructured portion. The structured portion contained all the addresses tied to a blockface with all the blockfaces being sequenced into blocks within the EA. The sequencing mirrored that found on the map that the Census Representative (CR) used for listing the EA in his/her Visitation Record (VR). The unstructured portion contained the addresses that could be tied only to the EA rather than a blockface. These were sorted by odd/even civic numbers within street name. The volume of addresses split 90%-10% between structured and unstructured.

Besides the address data, each page in an AR booklet contained a series of columns to be used in the reconciliation operation between the AR and VR. In the reconciliation, the Census Representative manually compared the Visitation Record with the AR to identify matches and non-matches. If the address was only on the VR, it was added to the AR (undercoverage in the AR). If the address was only on the AR, field resolution was usually required by the CR, with the result that the address was designated either as a new address to be enumerated for the Census by the CR (undercoverage in the Census) or as an invalid address classified by type of error (overcoverage in the AR). Addresses were denoted as invalid if they were duplicates, if they lay outside the EA, or for any other reason. All valid addresses had the Census Household Number coded in the booklet by the CR. A telephone number for the address, if available,

ADDRESS REGISTER Protected PROVINCE 35 EA 261 Page 21 of 22
FED 038 VN 0

Block No.	Address			Hhld No.	Not Listed at Drop-off	Field Follow-up Required	Invalid			AR Ref No.	Telephone Number
	Civic No.	Street	Apt. No.				Duplicate	Outside EA	Other		
1	2	3	4	5	6	7	8	9	10	11	12
4	23	MAIN ST								1044566	5551111
4	19	MAIN ST								1044564	5561234
4	15	MAIN ST								1044562	5552321
4	11	MAIN ST								1044559	
4	9	MAIN ST								1044583	7475739
4	7	MAIN ST								1044581	5552222
5	30	CENTRE RD								1019615	5561029
5	34	CENTRE RD								1019617	
5	34	CENTRE RD	BT							1019618	5564261
5	60	CENTRE RD								1019627	
5	64	CENTRE RD								1019629	7478765
5	68	CENTRE RD								1019634	5556942
5	72	CENTRE RD								1019636	
5	76	CENTRE RD								1019640	
5	80	CENTRE RD								1019642	7476789
5	84	CENTRE RD								1019644	5568765
5	88	CENTRE RD								1019646	5559999
5	92	CENTRE RD								1019579	7473456
5	96	CENTRE RD								1019581	7450987
5	100	CENTRE RD								1019648	
5	108	CENTRE RD								1019579	5557171
5	112	CENTRE RD								1019581	5558888
5	116	CENTRE RD								1019583	7462009
5	120	CENTRE RD								1019586	7450235
5	124	CENTRE RD								1019588	5569630

Figure 1. Example of a Page from an AR Booklet (reduced in size).

was pre-printed in the last column of the booklet to assist the CR in any required Census follow-up operation.

4. METHODOLOGY

In this section, the creation of the Address Register (AR) is described. Figure 2 provides an overview of the steps involved.

4.1 Overview of the Methodology

The free-format addresses contained on the source files were first standardized into ordered component parts (steps 1 and 2) in preparation for the use of subsequent software. Then, postal codes were confirmed or corrected (step 3) so that those areas or worksites for which the AR was to be created could be selected from among all the addresses and locations contained on the source files (step 4). Because the same addresses could be contained on more than one file or more than once on the same file, unduplication of addresses based on both exact and probabilistic matching took place (steps 5 and 6).

Next, automated linkages were made of addresses to the blockface level using the Area Master File (step 7) or, where this was not possible, to Enumeration Area (EA) using the Postal Code Conversion File (step 8). After loading the addresses into a database management system (step 9), manual linkages were made of addresses to blockface (steps 10 and 11) or to EA (step 12). Addresses within each EA were then sequenced by and within blocks (step 13) before being printed and collated in booklets by EA (step 14) for use in the Census.

4.2 Address Standardization (Steps 1, 2 and 3)

The Postal Address Analysis System (PAAS - step 2 of Figure 2) (Statistics Canada 1989c) performed two tasks: it broke up the free-format addresses from the source files into their component parts (street name, civic number, street designator, street direction, apartment number, municipality, province, postal code) and composed the address search key (ASK). ASK is an ordered concatenation of all the components of an address and is used during unduplication.

Although PAAS was an excellent product, analysis from the 1989 prototype had revealed certain shortcomings that we felt could be resolved by grooming or filtering the administrative file contents prior to using the generalized software. This FILTER step (step 1) concentrated on the following tasks: eliminating special characters with which PAAS refused to deal, repackaging address components in a manner compatible with PAAS, translating street designator short forms to acceptable ones, introducing commas between the street and municipality components of the free-format address to improve PAAS's comprehension, eliminating leading zeroes from civic numbers and numeric street names, and adding municipality and province names.

The FILTER and PAAS steps were applied in an iterative fashion. The first step was to discover what anomalies needed filtering for each administrative source. If the PAAS error rate after filtering was greater than 5%, error records were reviewed to find recurring problems that could be successively eliminated by further filtering until an error rate of less than 5% was achieved. As any address record that failed address standardization was eliminated from further consideration, it was vital to have a PAAS success rate as high as possible.

The PCVERFY step (step 3) used the Automated Postal Coding System (PCODE) (Statistics Canada 1989a) package for confirmation and generation of postal codes. It was not quite as effective as the PAAS software at address analysis and could only confirm or add postal codes for 84% of the output from PAAS. It confirmed 78% of the postal codes and changed another 6%. Only .003% of the source administrative records had arrived with no existing postal code. It was crucial to have correct postal codes because these would be used for worksite selection in the subsequent step.

Two problems arose in the PCVERFY step during production. If an address was missing a municipality/province component, the software continued to attempt to find a postal code instead of suspending further processing. As a consequence, enormous amounts of processing time could be spent trying to find postal codes. This problem was solved by including in the FILTER a step to add municipality and province names. The second problem occurred when a street name was numeric, as the processing time per address increased fourfold. This problem was not resolved and will necessitate modifications to the PCODE software.

4.3 Worksite Selection (Step 4)

This step partitioned the country by postal code into manageable worksites for processing with the sizes of worksites being based on the efficiency of CANLINK software for linkage of multiple large files. A geographic partitioning into worksites was adopted so they had dwelling counts in the 100,000 to 150,000 range based on the 1986 Census. Worksites were formed from an individual AMF (for a medium sized city), collections of physically adjacent AMFs (for small towns/townships), or parts of an AMF (for a large city). Geography Division's

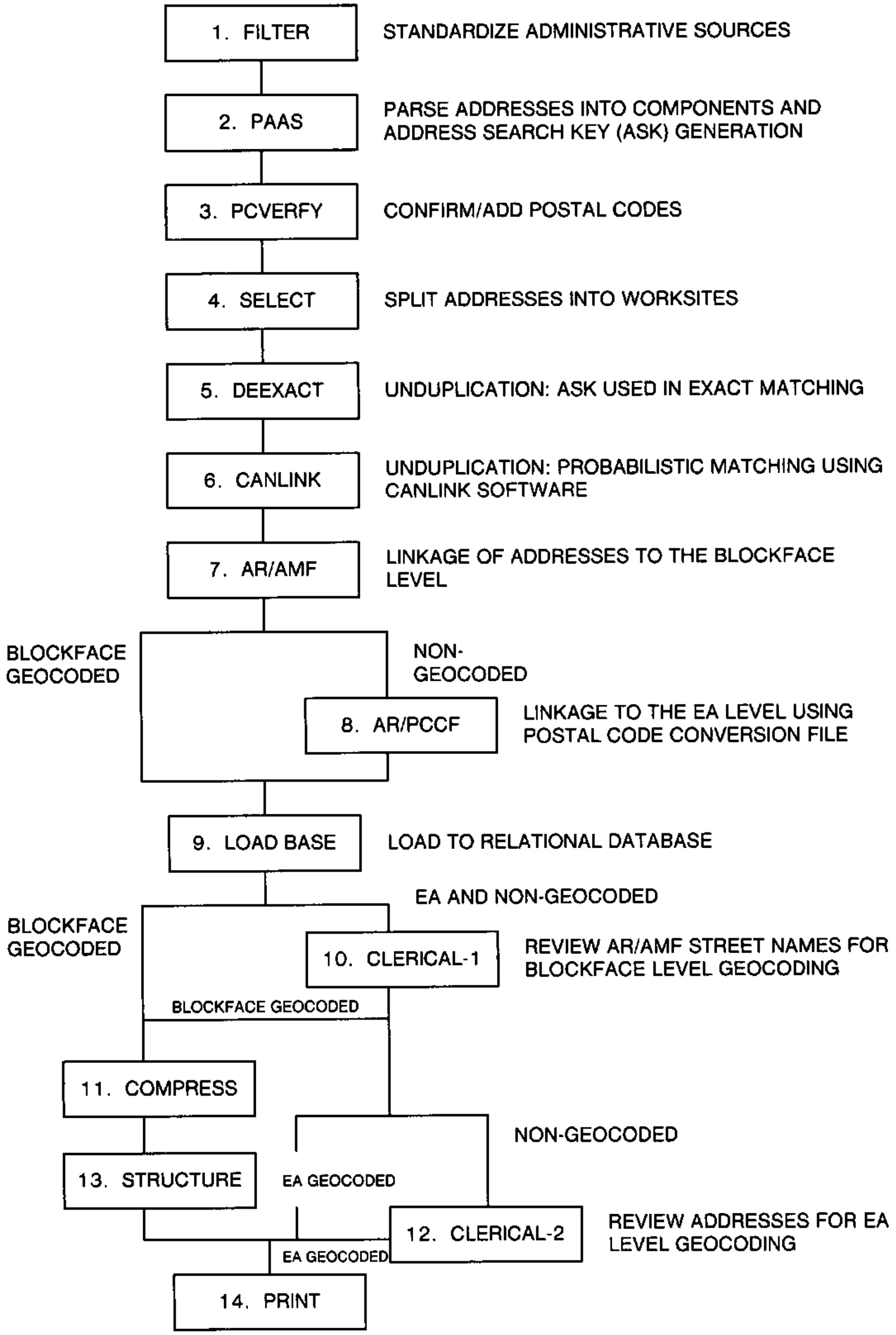

Figure 2. Overview of the Methodology.

Postal Code Conversion File (PCCF) which links postal codes and detailed Census geography was used to do this partitioning in the SELECT step (step 4). Once partitioning was completed, there were 105 distinct worksites and the original 43.4 million addresses had been reduced to 20.5 million addresses, with the dropped addresses having postal codes outside the AMF areas (*i.e.*, smaller cities and rural areas).

4.4 Unduplication (Steps 5 and 6)

In order to delete addresses included more than once on the source files, an unduplication process was conducted in two stages: an exact match with DEEXACT (step 5) and a probabilistic match using CANLINK software (step 6).

The DEEXACT step utilized the address search key (ASK) produced by the PAAS software and all records with an identical ASK were collapsed into a single record. With DEEXACT, the 20.5 million records from the SELECT step were reduced down to 10.1 million records. This reduction shows the importance of performing the address standardization.

Step 6 utilized the CANLINK generalized record linkage software (Statistics Canada 1989b). It clusters close records into groups called "pockets" and only records within the same pocket are actually matched together. For this application, civic number was used as the pocket. The components of the address (street name, municipality name, postal code, *etc.*) were used for matching purposes and weights were assigned for agreement or disagreement of each component. The development of levels of partial agreement for street name, municipality name and the last three characters of the postal code allowed for spelling variations and letter transpositions within the fields. The CANLINK step accounted for a further reduction to 6.7 million records. More details on the use of CANLINK in address unduplication are given in Drew *et al.* (1988), where its application in the November 1987 test is described.

4.5 AR/AMF Linkage (Step 7)

The major concern from the 1989 test was the strategy used to link addresses to their respective blockface. Because of the 11% drop in coverage from 84% to 73% compared to the 1987 test, a thorough investigation was needed and possibly a new approach. The other concern was that automated matching accounted for only 80% of the records matched while the other 20% were picked up clerically. This would have represented a daunting manual workload in full production. In order to circumvent these two concerns, another CANLINK application was developed for the AR/AMF linkage (step 7).

The original 1989 test files for Brampton still existed, so this became the test site for developing this step. The revised approach yielded 10% more matches, which increased the coverage back up to 1987 levels. As well, the automated matching was now responsible for 97% of the matches with 3% being picked up clerically, a significant improvement on the earlier 80%-20% split. Based on these results, the CANLINK approach was adopted for Census production.

In the construction of the new matching strategy, the first area of study involved a comparison of the contents of fields that would be used for matching purposes. This revealed certain anomalies that could be corrected prior to use to improve the number of linkages. The processing modifications to existing fields covered the following areas: removal of blanks between compound street names; alignment of street directions and civic numbers; conversion of numeric street names to numbers (on the AMF); removal of special characters in street names (on the AMF); correction of spelling variations in municipalities (on the AR); and a recreation of certain PAAS translations for street names (on the AR). Several new fields were also generated: NYSIIS (New York State Identification and Intelligence System) and SOUNDEX versions of the street name, employing two phonetic encoding packages used to eliminate the effects of common spelling errors (Statistics Canada 1989d); a duplicate street name flag (on the AMF) to identify situations where a street name was not unique; a unidirectional street flag (on the AMF) to identify streets that had only a single street direction coded; and an official street name flag (on the AR) to indicate that the street name matched an official AMF street name. The AMF records contained only street data so we appended the Census Subdivision name and a province code and then attempted to assign postal codes to blockface civic numbers. When the postal codes differed between the "from" and "to" civic numbers, we generated subblockfaces for each unique postal code.

For this application, three distinct pockets were created for each record, effectively triplicating the files. The primary pocket was the most stringent in nature and was designed to find all the good match possibilities quickly in the first pass of the files. It was composed of street name/Forward Sortation Area (FSA)/odd or even civic number flag. The second pocket was postal code/odd or even civic number flag which allowed for poorly parsed addresses to be matched on postal code. The third was the NYSIIS version of the street name/odd or even civic number flag which allowed records with spelling variations in street name and missing postal codes to be considered as potential matches.

The function rules established for partial matches for street name, municipality name and the last three characters of the postal code were taken directly from our existing CANLINK application used for internal unduplication where they had already demonstrated their effectiveness.

However, there were three AMFs to which we had difficulty matching in the course of production: Red Deer, St. Thomas and Charny. The problem with all three was missing civic number data on the AMF. Knowing that these would require heavy clerical intervention, a field operation was mounted in December 1990 to update the maps from the Computer Assisted Mapping System (CAM). CAM maps from Geography Division were sent to Regional Office staff who added the missing civic number ranges. These updated maps were subsequently forwarded to Geography Division for inclusion in the next round of updates to the AMF. For the creation of the AR, the civic number ranges for the three AMFs were used manually in the clerical operation.

Success in matching was quite similar across all provinces except for Québec. In Québec, the automatic matching to the blockface dipped by about 10-12% to 73% as it was not as effective at dealing with French addressing as it was with English addressing. Three situations were identified as causes for the drop in the automatic match rate: the use/non-use of articles within the street name (*e.g.*, Savane, de la Savane, la Savane), the use of complete personal names as street names with a high degree of spelling variability (*e.g.*, Jean-François Belanger, J.F. Belanger and Jean F. Belanger) and the lack of street designators. As a result, the clerical operations described below, especially the first one, were of increased importance for matching in Québec relative to the other provinces.

During the AR/AMF processing with the CANLINK software, the only problem that arose was in exceeding an internal pocket maximum on the number of records allowed. The solution was to identify the streets causing the problem from the pocket report (they were always major thoroughfares) and set up special pre-processing programs that would add the fifth digit of the postal code in calculating the pocket value for those streets to make it more discriminating. This had the effect of reducing the number of records within the pocket.

4.6 AR/PCCF Linkage (Step 8)

This step (step 8) attempted to obtain an automated link to the proper Enumeration Area (EA) for those addresses which could not be matched to the blockface using the AMF in step 7.

The principal inputs were the Postal Code Conversion File (PCCF), which gave the correspondence between postal codes and 1986 EAs, and the 1986 to 1991 EA Correspondence File. By matching the two together we could identify postal codes that were uniquely matched to a single 1991 EA, as well as postal codes matched to two or more possible 1991 EAs, requiring manual work to resolve later in step 12.

Again, Brampton became the test vehicle. The analysis of the postal code/EA matching revealed that 38% of the postal codes could be uniquely assigned to a 1991 EA. The linkage to these postal codes of the AR records unmatched to a blockface yielded a further 5% increase in total matches. Overall, the automated match rate increased to 89% (84% to the blockface and 5% to the EA), up from 64% in the September 1989 test, almost cutting in half the amount of manual intervention.

4.7 Loading the Base (Step 9)

To facilitate queries and in anticipation of future usage, ORACLE had been used in the 1989 test as the database management system and was used again for the 1991 production. The ORACLE load step (step 9) involved the transformation of the up-to-now sequential file into four separate component files, one for each of municipality, blockface, street and address.

4.8 Clerical Procedures (Steps 10, 11 and 12)

The clerical procedure for the 1989 test was a review of all unique combinations of street name/street designator/street direction from both AMF and AR records along with an AR record count for each street combination. The objective was to replace an unmatched AR street combination with the legimate AMF combination. By comparing similar street combinations and determining which ones should in fact have been identical, hitherto uncoded AR records could be matched manually to a particular blockface. This procedure had worked well in 1989 and had proved useful in two problem situations: those where there were large discrepancies in street name spelling and those where the AR street name field contained both the street name and a street designator short form that the PAAS software had not understood in parsing the address.

We expanded the capability of this clerical procedure (step 10) to compare AR street combinations with other similar AR street combinations to handle instances where a particular street might have a number of AR spelling variations with no AMF equivalent. This expansion permitted some additional manual coding of addresses to blockface.

To summarize, in this first clerical procedure (Clerical-1), all addresses not coded automatically to blockface in step 7 (that is, those coded automatically to EA in step 8 and those not yet coded) were examined for possible manual coding to blockface.

Following the Clerical-1 procedure, we added a Compress step (step 11), which was applied to all records coded to the blockface. For each unique value of street name/street designator/street direction within a worksite, all the corresponding address records were checked for uniqueness using the civic number/apartment number as the key. Where multiple records occurred, they were collapsed with all pertinent data blended into one single record, a further step of unduplication.

As a result, at the end of step 10, the database contained addresses coded automatically or manually to blockface, automatically to EA or uncoded as yet.

Step 12 now dealt with those residual addresses that could not be linked to a unique EA but could be matched to two or more possible EAs via step 8. A complete set of CAM-generated maps was produced for the AR project. The Clerical-2 step consisted of examining these maps for the candidate EAs to assign these residual addresses to the proper EA wherever possible.

Overall, the ratio of automated to manual matching was 91%-9%. The automated portion comprised 87% from the AR/AMF linkage to blockface, and 4% from the AR/PCCF linkage to EA. The manual portion was split 3% matched to the blockface from the Clerical-1 operation and 6% to the EA in Clerical-2.

Although ORACLE was an appropriate vehicle for the 1989 prototype, it proved to be costly and eventually a bottleneck once in full production with the AR as just one user on a Bureau-wide database. It allowed for only 8-10% of the worksites on-line at any one time, and had to export and import sites continuously to free up space and reload to carry on processing. A second ORACLE database was therefore set up for exclusive use of the AR team. In fairness to ORACLE, not all the processing being done was conducive to any database management system. The product was being built and as a consequence large portions of the tables were being examined to make sweeping field changes, to eliminate duplication and to select records for printing. ORACLE did offer tremendous flexibility to change software procedures quickly and generate new ones as production unfolded.

4.9 Use of the Computer Assisted Mapping System (Step 13)

The Computer Assisted Mapping System (CAM) was a new research initiative for the 1991 Census whose development ran concurrently with AR development. The system generated all the Enumeration Area maps within AMF coverage areas. This was a major departure from the manual map generation process of the past. CAM also provided a structure to EAs that located blockfaces within blocks and sequenced the blocks within the EA (step 13). An off-shoot to CAM for AR purposes was set up to sequence the dwellings on the blockface. This was necessary to organize the address lists in a manner corresponding more closely to the way the Census Representatives do their listing.

CAM was fully implemented by the time of AR production. In order to remain compatible with it, the same vintage of the AMF that CAM employed was used. However, a small portion of blockfaces had no structure data assigned to them. For any EA where this percentage was greater than 5%, either CAM was re-executed for that worksite if time permitted or an alternate system, Point-in-Polygon Assignments (PIPA), that locates blockfaces within their EA was executed. Although PIPA shifted addresses from the structured portion of the AR booklet (based on blockface coding) to the unstructured portion (EA coding), at least the affected addresses were not dropped during the print selection process, which was the case when sequencing data were missing.

4.10 Printing and Booklet Production (Step 14)

The last production step was the printing and gathering of booklets (step 14) for the almost 23,000 Enumeration Areas containing at this point 6.6 million addresses. Major concerns which were addressed included print speed and quality (a continuous-page printer was used), durability of booklets (the booklets had front and back covers and were stapled) and compilation costs (the booklets were gathered and attached in-house).

5. POST-CENSAL EVALUATION

The post-censal evaluation can be broadly categorized into four study areas: field operations, data capture of AR booklets, update of the AR and determination of the AR contribution to coverage improvements.

Evaluation of field operations will focus on the effectiveness of training, how complete the reconciliation work was, and causes of errors, with a view to improving the methodology for future Censuses.

The data capture operation will yield two separate outputs. First, addresses printed in the booklets will be deleted if invalid, and if valid their Census Household Number will be captured. Second, the new addresses added by the Census Representatives will be captured. It will then be possible to calculate the AR overcoverage and undercoverage rates and the AR contribution to Census coverage. Addresses placed in the wrong EA can be investigated and traced back to the source of error. Through the Census Household Number, the number of persons added and characteristics of dwellings and persons can be studied.

From a cost perspective, the unit cost per dwelling added by the AR will be calculated, in view of the cost of creating the AR and using it in the Census.

6. FUTURE DIRECTIONS

The Address Register (AR), although initially set up as one of the procedures for reducing Census undercoverage, is a developmental project with potential impact on other programs within Statistics Canada as well as other government agencies.

The more immediate objectives for the future development of the AR are as follows: to incorporate the addresses identified during Census enumeration; to evaluate the effectiveness of the AR in improving coverage of the 1991 Census; to document and evaluate the production activities; and to develop a longer-term plan for the AR addressing its cost-effectiveness as a household frame, the optimal updating strategy and its potential for use by external agencies.

Within these guidelines, a project plan was prepared and is presented below under six main topic areas.

6.1 Relationships between the Census and the Address Register

Besides the potential for coverage improvement, other ways in which the AR could contribute to the Census will be explored. Some preliminary thoughts in this regard include possibilities for the AR to be used as a processing control file, for telephone numbers to be used for follow-up purposes, for creation of control numbers of dwellings in an Enumeration Area, for certification of dwelling counts for processing, or for migration analysis. Consideration will be given to whether the AR should be used before or after Census Day, and to how the AR might be used for those addresses where only a higher level of geography than the EA can be ascertained.

6.2 Relationships between Geography and the Address Register

As is evident in the description of the methodology, the creation of the AR relied heavily on many of the products from Geography Division (*e.g.*, the Area Master File, the Postal Code Conversion File). Their contributions and limitations in building the AR will be reviewed. For any new products developed by Geography Division, their possible use in the AR will be investigated with a view to incorporating the AR needs directly into the new product. As well, the AR will be integrated into the Geography Division's Geographical Information System (GIS).

The AR may be able to provide update indicators to the Area Master File (AMF) or for the delineation of Enumeration Areas. The AR could be used to establish priorities especially in high-growth areas or in areas where there are poor civic number ranges in the AMF. The updating of the Postal Code Conversion File might be served by postal code/Enumeration Area or postal code/blockface combinations from the AR. After each Census, all Census households are encoded with blockface centroids. Since the bulk of AR records have already been geocoded prior to the Census, a link of the AR with the Census Household Number will reduce the amount of manual geocoding work after the Census. This last project is already in progress.

6.3 Documentation, Evaluation and Improvement of Procedures

A user guide documenting procedures and a technical guide to document programs, sample problems and solutions and quality assurance are being prepared for the work done to date.

As with any new project, much is learned during the creative process and procedures are developed as required and as time and budget permit. After the fact, there are usually efficiencies to be gained by reviewing these procedures.

For the automated procedures, projects already underway include a more efficient use of ORACLE or choice of another system, the use of desk-top computers rather than the Statistics Canada mainframe computer, standardization of the filter, enhancements to PAAS, amalgamation of sites into provincial databases, the dropping of some fields earlier in the process, consideration of other postal coding software, improvement of address place name matching and an improvement of the Area Master File linkage with French addresses.

For the manual procedures, improved handling of adjacent Enumeration Areas across boundaries of Federal Electoral Districts and of the lack of civic numbers on CAM maps are to be pursued. The editing system to correct addresses will be reviewed for possible improvement as well.

Telephone numbers were added at a later stage within the AR production. A thorough evaluation of their coverage and accuracy will be undertaken especially in view of the potential uses of telephone numbers in the Census and other Statistics Canada surveys. For the latter, initial emphasis will be placed on testing within the context of the upcoming redesign of the Labour Force Survey.

Computer systems developed for the initial production have already been cleaned up to a large extent for better efficiency of mainframe expenditures, for programs and disk and tape storage, for file manipulation, for output, libraries and file access. Better system controls will be prepared.

This AR was produced only for urban areas. Future methodological development will examine the potential for extension to rural areas.

6.4 Updating Methodology

The AR was created from among four sets of administrative files: telephone files, municipal assessment files, hydro files and the T1 tax file from Revenue Canada. As well, the AR is currently being updated to be consistent with the 1991 Census so that the Census is also a source. The relative contributions of these source files, both in volume and quality, will be investigated so that a decision on acquisition of files for updating can be made.

An integral part of the updating strategy is the development of a methodology for updating. The definition of an update will be needed along with an update system. The cost effectiveness of ongoing updating, dependent on the various needs which result from projects identified throughout these future directions, will be considered as well. Is ongoing updating cost effective when compared to updating only in time for the Census? What requirements will there be from other possible uses? Answers to these questions will lead to an updating strategy.

6.5 Other Uses of the Address Register in Statistics Canada

Besides the Census and geographical relationships presented earlier, a number of other uses are suggested within Statistics Canada. The potential use of the AR in the Labour Force Survey (LFS) will be investigated as part of the LFS Redesign Project. The possibility of using the AR in urban areas either to improve sampling under the existing area frame or as a list frame to reduce the number of stages in the sample design are two major areas highlighted for research. With telephone numbers on the AR, more telephone interviewing would be possible.

The use of the AR as a survey frame for other Statistics Canada surveys will be examined. In addition, since the AR currently uses telephone files as a primary source of information, it has these files on hand for further exploitation. The Special Surveys Program, the General Social Survey and the existing Labour Force Survey are areas which use or require telephone files.

Another potential application within Statistics Canada is as a housing database if the AR were enriched with housing data from the 1991 Census and data obtained from municipal assessment files, for example. The existence of such a database might reduce the amount of information on housing that would have to be collected in future Censuses. Data needs and availability have to be explored.

6.6 Uses of the Address Register External to Statistics Canada

If the AR is to be used outside Statistics Canada, issues of confidentiality of the source files and releasability of the AR must be addressed and meet the requirements of the Statistics Act. Some source files were provided to Statistics Canada in confidence, either contractually (*e.g.*, some files from Alberta) or legally (the T1 file from Revenue Canada).

6.7 Conclusion

The breadth and diversity of the ideas contained above in future directions demonstrate the potential of the Address Register as a geographical product with applications in many areas of Statistics Canada and elsewhere.

ACKNOWLEDGEMENTS

The authors would like to thank the many persons from the following areas for their dedication and perseverance in the creation of the Address Register: Phillip Reed and the AR Production Unit, Geography Division, the Labour Force Survey Sample Control Unit, Census Methodology, Survey Operations Division, the Main Computer Centre and Household Surveys Division. The authors would also like to thank the referee, Gordon Deecker, Peter Schut, Dick Carter, Phillip Reed and Carol Sol for their helpful suggestions for this paper.

REFERENCES

BOOTH, J.K. (1976). A summary report of all address register studies completed to date. Report E-414E, Statistics Canada.

DICK, P. (1990). Address register - September 1989 test. Draft internal report, Statistics Canada.

DREW, J.D., ARMSTRONG, J.B., and DIBBS, R. (1987). Research into a register of residential addresses for urban areas of Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 300-305.

DREW, J.D., ARMSTRONG, J., VAN BAAREN, A., and DEGUIRE, Y. (1988). Methodology for construction of address registers using several administrative sources. *Proceedings of the Symposium on the Statistical Uses of Administrative Data*, Statistics Canada, 181-190.

FELLEGI, I.P., and KRÓTKI, K.J. (1967). The testing program for the 1971 Census in Canada. *Proceedings of the Social Statistics Section, American Statistical Association*, 29-38.

FELLEGI, I.P., and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

GAMACHE-O'LEARY, V., NIEMAN, L., and DIBBS, R. (1987). Cost implications of mail-out of Census questionnaires using an address register. Internal report, Statistics Canada.

HILL, T., and PRING-MILL, F. (1985). Generalized iterative record linkage system. *Proceedings of the Workshop on Exact Matching Methodologies*, Arlington, Virginia, 327-333.

ROYCE, D. (1986). Address register research for the 1991 Census of Canada. *Journal of Official Statistics*, 2, 4, 447-455.

ROYCE, D. (1987). Applications of an address register in the Canadian Census. *Proceedings of the International 1991 Census Planning Conference*, Statistics Canada, 207-215.

ROYCE, D., and DREW, J.D. (1988). Address register research: Current status and future plans. Internal report, Statistics Canada.

STATISTICS CANADA (1988). Area Master File (AMF), User guide. Statistics Canada.

STATISTICS CANADA (1989a). Automated Postal Coding System (PCODE), User and retrieval guide. Statistics Canada.

STATISTICS CANADA (1989b). Generalized Iterative Record Linkage System, Concepts guide. Statistics Canada.

STATISTICS CANADA (1989c). Postal Address Analysis System (PAAS), User guide. Statistics Canada.

STATISTICS CANADA (1989d). Record linkage software, Reference guide. Statistics Canada.

STATISTICS CANADA (1990). *User's guide to the quality of 1986 Census data: Coverage*. Catalogue 99-135E, Statistics Canada.

STATISTICS CANADA (1991). Postal Code Conversion File (PCCF), the January 1991 version, User guide. Statistics Canada.

SWAIN, L., DREW, J.D., LAFRANCE, B., and LANCE, K. (1992). The creation of a residential address register at Statistics Canada. *Proceedings of the Symposiumon Spatial Issues in Statistics*, Statistics Canada.

VAN BAAREN, A. (1988). Report on the November 1987 address register test. Internal report, Statistics Canada.

Survey Methodology, June 1992
Vol. 18, No. 1, pp. 143–154
Statistics Canada

Bibliography on Capture-Recapture Modelling With Application to Census Undercount Adjustment

STEPHEN E. FIENBERG[1]

ABSTRACT

This article presents a selected annotated bibliography of the literature on capture-recapture (dual system) estimation of population size, on extensions to the basic methodology, and the application of these techniques in the context of census undercount estimation.

KEY WORDS: Capture-recapture; Census undercount; Dual system estimation; Loglinear models.

1. INTRODUCTION

The method of capture-recapture for estimating the size of a closed population has been in use since at least the nineteenth century, when Peterson (1896) developed the standard estimator that bears his name for the use with fish populations. Subsequent application to other types of populations include Geiger and Werner (1924) - physics; Lincoln (1930) - wildlife; Chandrasekar and Deming (1948) - vital statistics for human populations; Wittes and Sidel (1968), Wittes, Colton and Sidel (1974) - epidemiology; Sanathanan (1972b) - particle scanning in physics; Blumenthal and Marcus (1975) - life testing; Green and Stollmack (1981),Rossmo and Routledge (1990) - crimes and criminals. In the context of the study of human populations and demography the method is often referred to as dual system estimation. We have included virtually no references to the related problem of counting the number of species, which goes back to the work of R.A. Fisher in the 1940s and had an elegant formulation in Efron and Thisted's (1976) *Biometrika* paper on "How many words did Shakespeare know?".

The basic capture-recapture approach rests on a number of assumptions, *e.g.*: (1) the population under study is closed; (2) individuals (units) can be perfectly matched from capture to recapture; (3) capture probabilities are constant across the individuals (units) in the population; (4) the probability of inclusion of an individual (unit) in recapture sample is independent of inclusion in original census or sample. Beginning in the late 1930s various investigators began to explore extensions that allowed for departures from the assumptions. These methods typically require additional data such as a second recapture (or even a third) and the full capture-recapture history of each individual.

For human populations and the study of vital statistics the methodology has long been linked to census data, *e.g.*, see Tracy (1941) and Shapiro (1949, 1954). In connection with the 1950 decennial census of population, the U.S. Bureau of the Census introduced the use of a sample matched to the census records for coverage evaluation. This approach has evolved into what is currently known as the Post Enumeration Survey approach to undercount and overcount estimation, and it has been the focal point of the recent and ongoing controversy of the possible adjustment of the 1980 and 1990 censuses, *e.g.*, see Eriksen and Kadane (1985); Freedman and Navidi (1986, 1992); Freedman (1991); Wolter (1991).

[1] Stephen E. Fienberg, Office of Vice President (Academic Affairs), York University, North York, Ontario M3J 1P3, Canada.

This selected annotated bibliography presents an overview of published literature on capture-recapture estimation of population totals. It includes historical references, articles that explore departures from assumptions and extensions of the basic methodology, and is most complete in connection with papers that describe the dual and multiple system approaches in the context of census undercount estimation. In this regard, however, we have not included references to any of the unpublished memoranda and papers from the U.S. Bureau of the Census (primarily because most of these have been replicated in some form in the published literature). We have tended to exclude articles published in unrefereed proceedings for related reasons. Because the literature on specialized applications of capture-recapture techniques to wildlife populations is so extensive, and only some of it is of relevance for human populations, we have provided primarily references to reviews of this literature, *e.g.*, see Brownie *et al.* (1977); Otis *et al.* (1978); Seber (1973, 1982). Similarly we have included only a small number of references to the more specialized methods in use for life testing, *e.g.*, see Dahiya and Blumenthal (1986), as well as those in use for software reliability applications, *e.g.* Jelinski and Moranda (1972), and Duran and Wiorkowski (1981). The methods in this latter literature diverge in significant ways from those used in the basic capture-recapture and dual system approaches.

2. SELECTED BIBLIOGRAPHY

ALHO, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.

- Extends usual dual systems approach to allow for multiplicative stratification effects.

BAKER, S. G. (1990). A simple EM algorithm for capture-recapture data with categorical covariates (with discussion). *Biometrics*, 46, 1193-1200.

- Links cross-classification of covariates to the capture and recapture via loglinear models and then uses EM algorithm to estimate population size.

BIEMER, P.P. (1988). Modelling matching error and its effect on estimates of census coverage error. *Survey Methodology*, 14, 117-134.

- Develops models for evaluating impact of matching error on census coverage.

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.H. (1975). *Discrete Multivariate Analysis: Theory and Practice*, Chapter 6. Cambridge, MA: MIT Press.

- Monograph on loglinear models which includes a chapter on the relationship to capture-recapture models.

BLUMENTHAL, S., and MARCUS, R. (1975). Estimating population size with exponential failure. *Journal of the American Statistical Association*, 70, 913-922.

- Uses exponential distribution to estimate population size based on a subset of observations obtained by truncated sampling.

BOSWELL, M.T., BURNHAM, K.P., and PATIL, G. P. (1988). Role and use of composite sampling and capture-recapture sampling in ecological studies. In *Handbook of Statistics 6: Sampling,* (Eds. P.R. Krishnaiah and C.R. Rao). Amsterdam: North Holland, 469-488.

- Gives succinct summary of several basic variants on capture-recapture models and their estimation.

BROWNIE, C., ANDERSON, D.R., BURNHAM, K.P., and ROBSON, D.S. (1977). Statistical inference from band recovery data: a handbook. *U.S. Fisheries and Wildlife Service Resource Publication No. 131.*

- Describes a comprehensive range of capture-recapture models and appropriate goodness-of-fit tests, with emphasis on banding experiments.

BURGESS, R.D. (1988). Evaluation of reverse record check estimates of undercoverage in the Canadian Census of Population. *Survey Methodology*, 14, 137-156.

- Describes the survey-based accounting approach of the reverse record check for undercount estimation. Does not deal with issue of exclusion of individuals from census and other lists.

BURNHAM, K. P., ANDERSON, D.R., WHITE, G.C., BROWNIE, C., and POLLOCK,K.H. (1987). *Design and Analysis Methods for Fish Survival Experiments Based on Release-Recapture*. Bethesda, MD: American Fisheries Society.

- Combines methodology of Brownie *et al.* for band recovery with survival estimation under Jolly-Seber mark-recapture models.

BURNHAM, K.P., and OVERTON, W.S. (1978). Estimation of the size of a closed population when the capture probabilities vary among animals. *Biometrika*, 65, 625-633. Correction (1981) 68, 345.

- Develops a capture-recapture model with heterogeneity for animals but constant probabilities of capture across samples. Model induces dependencies amongst captures.

CASTELDINE, B.J. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, 67, 197-210.

- Develops a Bayesian approach using beta priors for traditional independence-based Schnabel census model for multiple recapture data.

CHAKRABORTY, P.N. (1963). On a method of estimating birth and death rates from several agencies. *Calcutta Statistical Association, Bulletin*, 12, 106-112.

- Extends Chandrasekar-Deming approach to three or more sources.

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

- Develops dual-system technique and suggests the use of stratification for eliminating heterogeneity. Applies approach to estimation of number of births and deaths in several Indian villages.

CHAO, A. (1987). Estimating the population size for capture-recapture data with unequal catch ability. *Biometrics*, 43, 783-791.

- Explores heterogeneous catchability model of Burnham and Overton using a moment inequality to get a lower bound on population size.

CHAO, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, 45, 427-438.

- Explores adequacy of estimator resulting from moment inequality for heterogeneous catchability model in settings involving sparse data.

CHAPMAN, D.G. (1951). Some properties of the hypergeometric distribution with applications to zoological sample censuses. *University of California Publications in Statistics*, 1, 131-160.

- Develops the hypergeometric sampling model for estimating the population size in capture-recapture studies.

CHOI, C.Y., STEEL, D.G., and SKINNER, T.J. (1988). Adjusting the 1986 Australian census count for under enumeration. *Survey Methodology*, 14, 173-189.

- Describes the use of dual system estimation and a post enumeration survey to adjust the results of the Australian census. Also applies Wolter sex-ratio technique to check on sensitivity of dual system estimator.

CHRISTENSEN, H.T. (1958). The method of record linkage applied to family data. *Marriage and Family Living*, 20, 38-43.

CITRO, C. F., and COHEN, M. L., (Eds.) (1985). *The Bicentennial Census. New Directions for Methodology* in 1990. Washington, DC: National Academy Press.

- Report of a panel of the Committee on National Statistics on census methodology including an examination of the dual systems approach to undercount correction.

COALE, A.J. (1961). The design of an experimental procedure for obtaining accurate vital statistics. *International Population Conference*, New York, 372-375.

- Proposes the use of two lists covering the same sample from a population.

COHEN, M.L. (1990). Adjustment and reapportionment - analyzing the 1980 decision. *Journal of Official Statistics*, 6, 241-250.

- Examines effect of bias and variability on accuracy of adjusted and unadjusted census counts and the impact on the reapportionment of the U.S. House of Representatives.

CORMACK, R. M. (1981). Loglinear models for capture-recapture experiments on open populations. In *The Mathematical Theory of the Dynamics of Biological Populations*, II (Eds. R.W. Hiorns and D. Cooke). London: Academic Press, 217-235.

- Introduces Poisson model for capture-recapture and uses it with loglinear models to extend standard approach to allow for birth, death, and trap dependency.

CORMACK, R. M. (1989). Log-linear models for capture-recapture. *Biometrics*, 45, 395-413.

- Uses Poisson model and loglinear representation for inclusion of birth, death, and trap dependency into standard capture-recapture approach.

CORMACK, R. M., and JUPP, P.E. (1991). Inference for Poisson and multinomial models for capture recapture experiments. *Biometrika*, 78, 911-916.

- Compares MLEs of parameters under the two models and presents relationship between the corresponding asymptotic variances and covariances.

COWAN, C.D., and MALEC, D.J. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81, 347-353.

- Extends dual systems approach to situation involving clustered observations as in the U.S. census coverage improvement program.

CRESSIE, N. (1988). When are census counts improved by adjustment? *Survey Methodology*, 14, 191-208.

- Proposes a PES-based model for undercount adjustment utilizing an empirical Bayes estimation scheme and a family of loss functions.

CRESSIE, N. (1989). Empirical Bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.

- Develops and applies empirical Bayes smoothing methods for census adjustment factors produced from dual systems approach for geographic by demographic stratification. Applies approach to state data from 1980 U.S. census.

CRESSIE, N., and DAJANI, A. (1991). Empirical Bayes estimation of U.S.undercount based on artificial populations. *Journal of Official Statistics*, 7, 57-67.

- Shows that synthetic estimation approach used by Isaki *et al.* is special case of empirical Bayes.

CROXFORD, A.A. (1968). Record linkage in education. In *Record Linkage in Medicine* (Ed. E.D. Acheson). London: E. and S. Livingstone, 351-356.

DAHIYA, R.C., and BLUMENTHAL, S. (1986). Population or sample size estimation. In *Encyclopedia of Statistical Sciences*, (Volume 7), (Eds. S. Kotz and N.L. Johnson). New York: Wiley, 100-110.

• Reviews theory underlying population size estimation from truncated sampling for discrete distributions and provides references to domains of application.

DARROCH, J.N. (1958). The multiple-recapture census I: Estimation of a closed population. *Biometrika*, 45, 343-359.

• Describes the maximum likelihood approach to the multiple recapture problem under complete independence.

DARROCH, J.N. (1959). The multiple-recapture census II: Estimation when there is immigration or death. *Biometrika*, 46, 336-351.

• Extends the maximum likelihood approach under independence to open populations with either immigration or death.

DARROCH, J.N. (1961). The two sample capture-recapture census when tagging and sampling are stratified. *Biometrika*, 45, 343-359.

• Extends the maximum likelihood approach with independence to the situation where the original captured individuals are stratified into s groups and the individuals in the recapture sample are stratified, but according to t (possibly different) strata.

DARROCH, J.N., FIENBERG, S.E., GLONEK, G.F.V., and JUNKER, B.W. (1992). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. Submitted for publication.

• Extends triple system estimation to allow for individual heterogeneity and selected forms of dependence. Applies estimators to triple system data from census dress rehearsal in St. Louis.

DARROCH, J.N., and RATCLIFF, D. (1980). A note on capture-recapture estimation. *Biometrics*, 36, 149-153.

• Presents an alternative estimator for capture-recapture problems with interesting asymptotic properties.

DASGUPTA, P. (1964). On the estimation of the total number of events and of the probabilities of detecting an event from information supplied by several agencies. *Calcutta Statistical Association, Bulletin*, 13, 89-100.

• Extends Chandrasekar-Deming approach to three or more sources.

DAVIDSON, L. (1962). Retrieval of misspelled names in an airline passenger record system. *Communications of the Association of Computer Machinery*, 5, 169-171.

DEMING, W.E., and KEYFITZ, N. (1967). Theory of surveys to estimate total populations. In *Proceedings of the World Population Conference*, Belgrade, 1965 (Vol. 3). New York: United Nations, 141-144.

• Extends of Chandrasekar-Deming approach to three sources.

DIFFENDAL, G. (1988). The 1986 test of adjustment related operations in Central Los Angeles County. *Survey Methodology*, 14, 71-86.

• Describes implementation of post-enumeration survey approach to dual system estimation in a test census.

DING, Y. (1990). Capture-Recapture Census with Uncertain Matching. Ph.D. dissertation, Department of Statistics, Carnegie Mellon University.

• Develops a probabilistic matching model for use with dual and multiple system estimation, and considers a Bayesian approach for estimating the population size. Illustrates techniques using data from test census results from Los Angeles.

DING, Y., and FIENBERG, S.E. (1992). Estimating population and census undercount in the presence of matching error. Submitted for publication.

- Develops a probabilistic matching model for use with dual system estimation and illustrates its application to data from test census results from Los Angeles.

DURAN, J.W., and WIORKOWSKI, J.J. (1981). Capture-recapture sampling for estimating software error content. *IEEE Systems Engineering*, 7, 147-148.

EFRON, B., and THISTED, R.A. (1976).Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63, 435-467.

- Adapts a parametric model due to Fisher and a nonparametric model for the classical species problem using empirical Bayes methods. Applies approach to the vocabulary of Shakespeare.

EL-KHORAZATY, M.N., IMREY, P.B., KOCH, G.G., and WELLS, H.B. (1977). Estimating the total number of events with data from multiple record systems: a review of methodological strategies. *International Statistical Review*, 45, 129-157.

- Review of literature and methods for dual- and multiple systems estimation. Includes sections comparing use of techniques and departures from assumptions in wildlife and human populations.

ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond (with discussion). *Journal of the American Statistical Association*, 80, 98-131.

- Applies dual system approach to 1980 census data, including the regression-based smoothing of undercount estimates and the estimation of adjusted odds ratios using demographic estimates.

ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1989). Adjusting the 1980 census of population and housing. *Journal of the American Statistical Association*, 84,927-944.

- Presents revisions and extensions to the Ericksen and Kadane methodology and a critique of Freedman and Navidi.

FAY, R. E., PASSEL, J. S., ROBINSON, J. G., and COWAN, C. D. (1988). *The Coverage of Population in the 1980 Census*. Bureau of the Census. Washington, DC: U. S. Department of Commerce.

- Official Bureau of the Census report on attempts to measure undercount in the 1980 U.S. decennial census.

FEIN, D.J., and WEST, K.K. (1988). The sources of census undercount: Findings from the 1986 Los Angeles Test Census. *Survey Methodology*, 14, 223-240.

- Attempts to test hypotheses regarding the causes of census undercount for a hard-to-enumerate Hispanic urban population.

FIENBERG, S.E. (1972). The multiple-recapture census for closed populations and the 2^k incomplete contingency table. *Biometrika*, 59, 591-603.

- Introduces a method for estimating dependencies among multiple lists using loglinear models and develops a general approach for estimation using results on 2^k incomplete contingency tables and conditional estimation.

FIENBERG, S.E. (1989). Undercount in the U.S. decennial census. In *Encyclopedia of Statistical Sciences*, (Supplemental Volume), (Eds. S. Kotz and N.L. Johnson). New York: Wiley, 181-185.

- Presents historical background on the differential undercount of the U.S. population and brief descriptions of demographic analysis and the dual system estimation approaches.

FREEDMAN, D. A. (1991). Policy forum: Adjusting the 1990 census. *Science*, 252, 1233-1236.

- Critique of dual systems approach to adjustment of the 1990 census.

FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models and adjusting the 1980 census (with discussion). *Statistical Science*, 1, 3-39.

- Critique of Ericksen and Kadane dual systems methodology as applied to 1980 census data.

FREEDMAN, D.A., and NAVIDI, W.C. (1992). Should we have adjusted the census of 1980? (with discussion). *Survey Methodology*, this issue.

- Continues critique of the use of dual system estimation and synthetic adjustment as applied to 1980 census.

GARTHWAITE, P.H., and BUCKLAND, S.T. (1990). Analysis of multiple-recapture census by computing conditional probabilities. *Biometrics*, 46, 231-238.

- Uses a recursive relationship to generate point and interval estimate for multiple-recapture census under independence.

GEIGER, H., and WERNER, A. (1924). Die Zahl der ion radium ausgesandsen a-Teilchen. *Zeitschrift für Physik*, 21, 187-203.

- Applies a capture-recapture method to radium ion particle detection estimation.

GOLDBERG, J.D., and WITTES, J.T. (1978). The estimation of false negatives in medical screening. *Biometrics*, 34, 77-86.

- Applies capture-recapture models to problems in medical screening.

GOUDIE, I. B. J. (1990). A likelihood-based stopping rule for recapture debugging software reliability. *Biometrika*, 77, 203-206.

GREEN, M.A., and STOLLMACK, S. (1981). Estimating the number of criminals. In *Models in Quantitative Criminology*, (Ed. J.A. Fox). New York: Academic Press, 1-24.

GREENFIELD, C.C. (1975). On the estimation of a missing cell in a 2 × 2 contingency table. *Journal of the Royal Statistical Society*, Series A, 138, 51-61.

- Introduces a non-zero value for the response correlation, by taking the mid-point of the range of permissible correlation values, and consequently derives a value for missing cell. Applies approach to census data from Malawi.

GREENFIELD, C.C. (1976). A revised procedure for dual record systems in estimating vital events. *Journal of the Royal Statistical Society*, Series A, 139, 389-401

- Applies bounds on correlation in a 2 × 2 table to dual system estimation in the presence of event correlation induced by heterogeneity.

GREENFIELD, C.C., and TAM, S.M. (1976). A simple approximation for the upper limit to the value of a missing cell in a 2 × 2 contingency table. *Journal of the Royal Statistical Society*, Series A, 139, 96-103.

- Uses approximation for upper bound for response correlation to derive an upper bound for missing cell.

HOGAN, H., and WOLTER, K.M. (1988). Measuring accuracy in a post-enumeration survey. *Survey Methodology*, 14, 99-116.

- Reports on Los Angeles Test of Adjustment Related Operations and estimates of sources of bias in post enumeration survey and census-based dual systems estimates.

HOLST, L. (1973). Some limit theorems with applications in sampling theory. *Annals of Statistics*, 1, 644-658.

- Applies results on successive sampling to derive asymptotic distribution of usual Peterson estimator when there are heterogeneous capture probabilities or the effects of matching.

HOOK, E., and REGAL R. (1982). Validity of Bernoulli census, log-linear, and truncated binomial models for correcting underestimates in prevalence studies. *American Journal of Epidemiology*, 116, 168-176.

- Applies different loglinear related methods used to study the number of infants born with Downs syndrome.

HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.

- Uses linear logistic models for capture probabilities for individuals and capture occasions.

HUGGINS, R.M. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, 47, 725-732.

- Uses linear logistic models for capture probabilities and exploits temporal order of captures to introduce dependence amongst captures and on measurable covariates for those captured at least once.

ISAKI, C.T. (1986). Bias of the dual system estimator and some alternatives. *Communications in Statistics, Theory and Methods*, 15, 1435-1450.

- Exploits upper bound on correlation bias to reduce the bias of the dual system estimator.

ISAKI, C.T., and SCHULTZ, L.K. (1986). Dual system estimation using demographic analysis data. *Journal of Official Statistics*, 2, 169-179.

- Uses demographic analysis data to get revised dual system estimates for 1980 census using different models for correlation bias.

ISAKI, C.T., and SCHULTZ, L.K. (1987). The effect of correlation and matching error in dual system estimation. *Communications in Statistics, Theory and Methods*, 16, 2405-2427.

- Develops a simple matching error model in the presence of correlation bias to compare three dual system estimators.

ISAKI, C.T. , SCHULTZ, L.K., DIFFENDAL, G.J., and HUANG, E.T. (1988). On estimating census undercount in small areas. *Journal of Official Statistics*, 4, 95-112.

- Develops simulation populations based on 1980 census and coverage evaluation results, evaluates regression-based synthetic undercount estimation methods, and shows superiority of synthetic approaches to raw census counts.

JABINE, T.B., and BERSHAD, M.A. (1968). Some comments on the Chandrasekar and Deming technique for the measurement of population change. Paper presented at CENTO Symposium on Demographic Statistics, Karachi, Pakistan.

- Shows that positive correlation bias produces a downward bias in estimate of total population size.

JARO, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 Test Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.

- Describes census methodology for matching census and post-enumeration survey records, with the results from their application to 1985 test census.

JELINSKI, Z., and MORANDA, P.B. (1972). Software reliability research. In *Statistical Computer Performance Evaluation*, (Ed. W. Freiberger). New York: Academic Press, 465-484.

- Proposes a model with exponentially distributed failures to estimate total number of program faults based on times of occurrence of failures in fixed time period.

JEWELL, W.S. (1985). Bayesian estimation of undetected errors. In *Bayesian Statistics 2*, (Eds. J.M. Bernardo, *et al.*). New York: Elsevier, 663-671.

JOLLY, G.M. (1965). Explicit estimates from capture-recapture data with both death and immigration – stochastic models. *Biometrika*, 52, 225-247.

- Estimation from multiple-recapture data for open populations.

KADANE, J.B., MEYER, M.M., and TUKEY, J.W. (1992). Correlation bias in the presence of stratum heterogeneity. Submitted for publication.

- Demonstrates the impact of correlation bias resulting from collapsing over heterogeneous strata with different catchability probabilities in each strata, subject to a monotonicity constraint.

KRÓTKI, K.J. (Ed.) (1978). *Developments in Dual System Estimation of Population Size and Growth*. Edmonton: University of Alberta Press.

- Reviews the use of dual system estimation for vital records in various countries. Includes technical details on the use of complex samples and elaborations on basic techniques.

LASKA, E.M., MEISNER, M., and SIEGEL, C. (1988). Estimating the size of a population from a single sample. *Biometrics*, 44, 461-472. Correction, (1989), 45, 1347.

- Estimates population size from the last of k lists.

LEWIS, C.E., and HASSANEIN, K.M. (1969). The relative effectiveness of different approaches to the surveillance of infection among hospitalized patients. *Medical Care*, 8, 379-384.

- Applies dual system estimation to surveillance of infectious diseases.

LINCOLN, F.C. (1930). Calculating waterfowl abundance on the basis of banding returns. *Circular of the U.S. Department of Agriculture*, 118, 1-4.

- Applies capture-recapture method to estimating size of waterfowl populations.

MANTEL, N. (1951). Evaluation of a class of diagnostic tests. *Biometrics*, 7, 240-246.

- Shows how heterogeneity induces correlation bias (event correlation) in the estimation of disease prevalence.

MARKS, E.S., SELTZER, W., and KRÓTKI, K.J. (1974). *Population Growth Estimation: A Handbook of Vital Statistics Measurement*. New York: Population Council.

- Comprehensive review of dual-systems estimation, assumptions, background, design, and problems. Contains claim that the basic method has been used for more than three centuries for estimating size of animal populations.

MAXIM, L.D., HARRINGTON, L., and KENNEDY, M. (1981). A capture-recapture approach for estimation of detection probabilities in aerial surveys. *Photogrammetric Engineering and Remote Sensing*, 47, 779-788.

MULRY, M. H., and SPENCER, B.D. (1988). Total error in the dual system estimator: the 1986 census of Central Los Angeles County. *Survey Methodology*, 14, 241-263.

- Develops a total error model for dual systems approach applied to Los Angeles Test of Adjustment Related Operations.

MULRY, M. H., and SPENCER, B.D. (1991). Total error in PES estimates of population (with discussion). *Journal of the American Statistical Association*, 86, 839-863.

- Extends earlier Mulry-Spencer development of total error model for dual systems approach and applies it to 1988 dress rehearsal census in St. Louis and east-central Missouri.

NICHOLS, J.D., and POLLOCK, K.H. (1983). Estimating taxonomic diversity, extinction rates, and speciation rates from fossil data using capture-recapture models. *Paleobiology*, 9, 150-163.

OTIS, D.L. BURNHAM, K.P, WHITE, G.C., and ANDERSON, D.R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monograph*, 62, Washington, DC: Wildlife Society.

- Reviews capture-recapture and related methods for wildlife populations.

PERKINS, W.M., and JONES, C.D. (1965). Matching for census coverage checks. Paper presented at the Meetings of the American Statistical Association, Philadelphia.

PETERSON, C.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. Report of the Danish Biological Station to the Ministry of Fisheries, 6, 1-48.

- Classic development of method of capture-recapture and its application to the estimation of the size of fish populations.

POLLACK, E.S. (1965). Use of census matching for study of psychiatric admission rates. *Proceedings of the Social Statistics Section, American Statistical Association*, 107-115.

RAJ, D. (1977). On estimating the number of vital events in demographic surveys. *Journal of the American Statistical Association*, 72, 377-381.

- Develops a formula for bias in dual system estimator under a general model of response errors and explores use of double sampling to correct bias.

ROSSMO, D.K., and ROUTLEDGE, R. (1990). Estimating the size of criminal populations. *Journal of Quantitative Criminology*, 6, 293-314.

RUBIN, D.B. , SCHAFER, J.L., and SCHENKER, N. (1988). Imputation strategies for missing values in post-enumeration surveys. *Survey Methodology*, 14, 209-221.

- Presents a methodology matching for undercount estimation which utilizes an imputation approach rooted in loglinear models in the presence of missing data.

SANATHANAN, L.P. (1972a). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43, 142-152.

- Demonstrates asymptotic equivalence of conditional and unconditional estimators for the population size.

SANATHANAN, L.P. (1972b). Models and estimation methods in visual scanning experiments. *Technometrics*, 14, 813-829.

- Develops latent model to estimate the number of particles in scanning records which allows for differential detectability and induces dependencies amongst detectors.

SANATHANAN, L.P. (1973). A comparison of some models in visual scanning experiments. *Technometrics*, 15, 67-78.

- Applies traditional capture recapture model and latent models to data from actual visual scanning experiments.

SANDLAND, R.L., and CORMACK, R.M. (1984). Statistical inference for Poisson and multinomial models for capture recapture experiments. *Biometrika*, 71, 27-33.

- Shows relationship between the asymptotic variances of the population size under general capture-recapture model for the two alternate sampling schemes.

SCHENKER, N. (1988). Handling missing data in coverage estimation with application to the 1986 Test of Adjustment Related Operations. *Survey Methodology*, 14, 87-98.

- Examines effect of missing data on dual system estimate applied to test census data.

SCHIRM, A.L., and PRESTON, S.H. (1987). Census undercount adjustment and the quality of geographic population distributions (with discussion). *Journal of the American Statistical Association*, 82, 965-990.

- Applies synthetic estimation approaches to 1980 census data to evaluate the impact of undercount estimation.

SCHNABEL, Z.E. (1938). The estimation of the total fish population of a lake. *American Mathematical Monthly*, 45, 348-352.

- Extends the basic capture-recapture approach to multiple recaptures, with information at each recapture on whether individuals were captured previously.

SCOTT, C. (1974). The dual record (PGE) system for vital rate measurement, some suggestions for further development. In *International Population Conference, Liege*, 1973, (Volume 2). Liege: International Union for the Scientific Study of Population.

SEBER, G.A.F. (1965). A note on the multiple-recapture census. *Biometrika*, 52, 249-259.

- Estimation from multiple-recapture data for open populations with time-specific parameters.

SEBER, G.A.F. (1973). *The estimation of Animal Abundance and Related Parameters*. New York: Hafner. Second edition (1982). New York: Macmillan.

- Contains an up-to-date review of capture-recapture techniques and their extensions for animal populations, with emphasis on applications.

SEBER, G.A.F. (1982). Capture-recapture methods. In *Encyclopedia of Statistical Sciences*, (Volume 1), (Eds. S. Kotz and N.L. Johnson). New York: Wiley, 367-374.

- Reviews capture-recapture approach for both closed and open populations and provides guide to wildlife and fisheries applications.

SHAPIRO, S. (1949). Estimating birth registration completeness. *Journal of the American Statistical Association*, 45, 261-264.

- Describes use of 1940 U.S. census data to check on completeness of birth registration using Chandrasekar-Deming estimator.

SHAPIRO, S. (1954). Recent testing of birth registration completeness in the United States. *Population Studies*, 8, 3-21.

- Describes use of U.S. census data from 1940 and 1950 to check on completeness of birth registration using Chandrasekar-Deming estimator.

SIRKEN, M. G. (1978). Dual systems estimators based on multiplicity surveys (with discussion). Chapter 4 in *Developments in Dual System Estimation of Population Size and Growth*, (Ed. K. J. Krótki). Edmonton: University of Alberta Press, 81-91.

- Adapts author's approach of multiplicity surveys for rare events to dual systems problem.

SMITH, P.J. (1988). Bayesian methods for capture-recapture surveys. *Biometrics*, 44, 1177-1189.

- Uses Poisson approximation and Gamma prior for a Bayesian approach to estimation under independence in multiple recapture model.

SMITH, P. J. (1991). Bayesian analyses for multiple capture-recapture model. *Biometrika*, 78, 399-407.

- Develops exact Bayesian posterior distribution for multiple recapture census under independence of recaptures.

SRINIVASAN, S.K., and MUTHIAH, S.A. (1968). Problems of matching births identified from two independent sources. *Journal of Family Welfare*, 14, 13-22.

TRACY, W.R. (1941). Fertility of the population of Canada. Reprinted from *Seventh Census of Canada*, 1931, (Vol. 2), Census Monograph no. 3. Ottawa: Cloutier.

- An early application of dual systems approach to census data.

WINKLER, W.E. (1989). Methods for adjusting for lack of independence in an application of the Felligi-Sunter model of record linkage. *Survey Methodology*, 15, 101-117.

- Examines methods of adjusting linkage rules for matching dual-system records when independence cannot be assumed.

WITTES, J.T. (1974). Applications of a multinomial capture-recapture model to epidemiological data. *Journal of the American Statistical Association*, 69, 93-97.

- Applies multiple systems and independence to estimate the size of a population of children with a congenital anomaly and other problems.

WITTES, J.T., COLTON, T., and SIDEL, V.W. (1974). Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *Journal of Chronic Diseases*, 27, 25-36.

- Applies multiple systems and independence to estimate the size of a population of children with a congenital anomaly.

WITTES, J.T., and SIDEL, V.W. (1968). A generalization of the simple capture-recapture model with applications to epidemiological research. *Journal of Chronic Diseases*, 21, 287-301.

- Uses capture-recapture approach to estimate number of hospital patients using methicillin.

WOLTER, K. M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.

- Describes basic version of dual systems approach as used for the 1980 census, including the elimination of erroneous enumerations.

WOLTER, K. M. (1990). Capture-recapture estimation in the presence of a known sex ratio. *Biometrics*, 46, 157-162.

- Uses sex ratio to get a modified capture-recapture estimator for data with stratification by sex, assuming either a common odds ratio, or independence in one stratum. Applies approach to animal data and describes application to census data.

WOLTER, K.M. (1991). Policy Forum: Accounting for America's uncounted and miscounted. *Science*, 253, 12-15.

- Describes 1990 census adjustment procedures and why they are statistically defensible.

WOLTER, K. M., and CAUSEY, B.D. (1991). Evaluation of procedures for improving population estimates for small areas. *Journal of the American Statistical Association*, 86, 278-284.

- Examines through simulations the improvements of synthetic adjustment over the census counts for small areas.

ZALAVSKY, A.M., and WOLFGANG, G.S. (1990). Triple system modelling of census, post-enumeration survey, and administrative list data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 668-673.

- Applies various loglinear and related models of dependence to analyze triple system data from census dress rehearsal in St. Louis.

Survey Methodology, June 1992
Vol. 18, No. 1, pp. 155–163
Statistics Canada

A Variation of the Housing Unit Method for Estimating the Population of Small, Rural Areas: A Case Study of the Local Expert Procedure

LINDA K. ROE, JOHN F. CARLSON and DAVID A. SWANSON[1]

ABSTRACT

This paper examines the suitability of a survey-based procedure for estimating populations in small, rural areas. The procedure is a variation of the Housing Unit Method. It employs the use of local experts enlisted to provide information about the demographic characteristics of households randomly selected from residential unit sample frames developed from utility records. The procedure is nonintrusive and less costly than traditional survey data collection efforts. Because the procedure is based on random sampling, confidence intervals can be constructed around the population estimated by the technique. The results of a case study are provided in which the total population is estimated for three unincorporated communities in rural, southern Nevada.

KEY WORDS: Survey-based; Utility records; Confidence intervals; Nevada.

1. INTRODUCTION

In its most recent survey of state and local agencies preparing population and housing estimates, the U.S. Bureau of the Census found that about 89 percent of the agencies surveyed use the Housing Unit Method (HUM) (Byerly 1990). This method was also found to be widely used in an earlier survey (U.S. Bureau of the Census 1978). The method has been found to provide accurate estimates of the total population (Lowe, Pittenger and Walker 1977; Lowe, Weisser and Myers 1984; Smith and Lewis 1980, 1983; Smith and Mandell 1984) as well as a strong conceptual and practical foundation for a municipal estimation system (Martin and Serow 1979; Rives and Serow 1984; Swanson, Baker and Van Patten 1983).

One of the strong features of the HUM is that it can be implemented in a variety of forms, which allows it to be adapted to a range of data environments (Swanson, Baker and Van Patten 1983). This adaptability has been exploited primarily by subnational demographic centers for purposes of revenue sharing and related programs (Martin and Serow 1978; Swanson, Baker and Van Patten 1983). However, as pointed out by Rives (1982), the method has potential uses in other arenas.

As an example, consider the case of environmental impact statements. Concerns over legal and environmental issues have resulted in decisions to locate unpopular facilities in sparsely populated rural areas for which census and other socioeconomic data are usually not available (Freudenburg 1982; Brown, Geertsen and Krannich 1989; Munsell 1988). As a consequence, it has become necessary to develop methods of inquiry, particularly suited for small, rural areas, that fully exploit available data, are less costly and, in many cases, less intrusive, than area, telephone, and mail surveys. We believe that the variation of the HUM that we propose in this paper contributes to this type of methodological development.

[1] Linda K. Roe and John F. Carlson, Science Applications International Corporation, 101 Convention Center Drive, Las Vegas, Nevada 89109; David A. Swanson, Center for Social Research and Department of Sociology, Pacific Lutheran University, Tacoma, Washington 98447-0003, USA.

The HUM variation that we describe in this paper combines two methods that are, in themselves, well known. However, they have largely been developed in isolation from each other, as well as from the HUM. These are: (1) random sampling; and (2) "local expert" interviews. As discussed later, these methods, combined with the HUM may lead to a means of obtaining the population size and, eventually, composition data required to meet the information needs of impact assessment projects and other activities affecting small, rural areas.

2. CONSIDERATIONS IN ASSESSING IMPACTS IN SMALL, RURAL AREAS

The location of new plants or industries in rural areas generally requires a work force exceeding that which is available in the local area. Population growth in the communities that are in close proximity to the site can be expected to vary according to the size of the project and the number of employees that will be hired to build, then to operate and maintain the completed facility. Whether rapid increases in the overall number of individuals are expected, or significant changes in the age and sex distribution, the altered population structure will have an effect on the type and amount of public services needed (Summers 1982). Thus, impact assessments require information regarding anticipated increases in school enrollment, housing requirements, health care needs, and other services. Before such projections can be made, however, information on the current population in the impacted area must be determined in order to have a "jump-off" or "launch" population for forecasting purposes (Carlson, Williams and Swanson 1990; Pittenger 1976; U.S. Department of Energy 1988).

The understanding of major factors affecting the distribution of people in isolated rural areas is critical in constructing demographic profiles and projections. These communities are likely to have been affected by periods of both boom growth and decline (Krannich and Greider 1984). Historical patterns of population change, as well as current trends, may differ substantially from averages derived from that of the county as a whole or even other sub-county areas. This presents a special problem because accurate demographic information is usually available only for years in which the Federal Census is conducted. However, census data, including information on households, are not typically available for unincorporated places with small populations. Since cost is usually a major factor, the possibility of conducting special censuses or large sample surveys, particularly on a regular basis, is often precluded, even in small, rural areas. An additional problem associated with such counts is that they require interviewers to contact individual households, which imposes on time and privacy and adds to disruption burdens that may be already high for local residents (Brown, Geertsen and Krannich 1989; Krannich, Berry and Greider 1989; Schleifer 1986).

The estimation of the size of the current population of a small, rural area could, in principle, be accomplished through several techniques. However, data limitations and a desire for accuracy severely curtail the range of candidates and, realistically, point to a single technique: HUM (Smith 1986; Smith and Mandell 1984; Lowe, Weisser and Myers 1984; Swanson, Van Patten and Baker 1983; Smith and Lewis 1983, 1980).

3. THE HOUSING UNIT METHOD

The concept of the HUM relies on the fact that nearly everyone sleeps under some kind of shelter. The U.S. Bureau of the Census, for example, chooses to define two classes of shelters: group quarters; and housing units. All persons are assigned to one shelter class or the other. The HUM holds that these shelters can be identified, counted, and classified as occupied or vacant. Also, all occupied shelters must have a specific number of occupants. Therefore, the population of any given place must be equal to the sum of the housing units times the occupancy rate times the average number of persons per occupied housing unit (household) plus the number of persons in group quarters. The four elements of the HUM provide an exact demographic identity, with the population of a given place given by

$$P = [(H) * (O) * (PPH)] + GQ,$$

where

$$\begin{aligned} P &= \text{total population,} \\ H &= \text{total housing units,} \\ O &= \text{proportion of occupied units,} \\ PPH &= \text{mean number of persons per household,} \\ GQ &= \text{group quarters population.} \end{aligned}$$

The key accuracy issue in using the HUM is in the determination of each of the components. Moreover, as Smith (1986, pp. 245-246) observes:

> "The Housing Unit Method is a robust, comprehensive, and extremely flexible form of population estimation with a number of characteristics that make it useful for small-area analysis. It is not confined to a single technique or type of data; rather, it can utilize a number of different techniques and data sources, including those that may be applicable in one area but not another."

As also noted by Smith (1986), there are two major approaches used to generate the "number of households" element of the HUM. One relies on measures of construction activity and the estimation of an occupancy rate; the other uses utility data, such as residential electrical customers. A major advantage of the second approach is that it can directly provide the number of households, which eliminates or substantially reduces a number of potential data inaccuracies, including the need to estimate time lags between when permits are issued and units are completed, completion rates, demolitions, conversions, and occupancy rates. Starsinic and Zitter (1968) as well as Rives and Serow (1984) find that the "utility data" approach to the HUM is advantageous, although they also acknowledge certain limitations.

Another advantage of using utility data is that the same data used to obtain total households can also be used as a complete frame from which samples can be drawn in order to obtain an estimate of the average number of persons per household (PPH). There are three forms that traditional data collection usually take in obtaining this type of sample information: mail, telephone, and personal interview. We propose that in their place "local experts" be used to minimize both cost and disruption burdens.

4. LOCAL EXPERTS

The local expert procedure (also refered to as the key informant procedure) of obtaining information about a community is well-established in the field of cultural anthropology. It is generally acknowledged as a "reliance on a small number of knowledgeable participants, who observe and articulate social relationships for the researcher " (Seidler 1974, p. 816). Further, Poggie (1972) finds that when the questions asked in the field relate to noncontroversial, concrete, and directly observable public phenomena, local experts are a highly reliable and precise source of information.

There are two key issues in using the local expert procedure in conjunction with utility records and the HUM. The first is to identify and recruit people who are truly local experts on the composition of the households presented to them in the sample. The second is to be able to obtain household identifying information that is familiar to the local experts (*e.g.*, a street address and the name of the householder instead of a utility company billing code).

5. CASE STUDY

The data collection activity on which our population estimates rely is part of a program to assess the socioeconomic characteristics of communities located near Yucca Mountain, Nevada, the proposed site of a geologic nuclear waste repository (U.S. Department of Energy 1988). The data will comprise part of the set used in a comprehensive impact analysis of the proposed repository.

Yucca Mountain is located in Nye county, approximately 90 miles northwest of Las Vegas in a sparsely populated, desert area. The impact analysis is focused on the communities that are within a fifty mile radius of the Yucca Mountain site. The study areas includes the unincorporated communities of Amargosa Valley, Beatty, and Pahrump in southern Nye county and Indian Springs in Clark county. Tax boundaries specified by the county commissioners are used to deliniate community boundaries for purposes of the impact analysis.

6. DATA AND METHODS

During a preliminary phase of the research, contacts were made with community leaders and residents. These contacts resulted in a network that later facilitated the collection of data. Field notes were taken describing the general layout of each community in the study area. These included the types and locations of businesses and residential areas. Four separate housing types were defined using the guidelines developed by the U.S. Bureau of the Census.

Following the preliminary investigation, the road system and other features were mapped for each community. Using these maps and utility records, representatives of the electrical company servicing southern Nye county identified the location and type of housing, if any, associated with all current electrical connections. This information was added to the housing unit file constructed from the utility records for each community. Because of the lack of adequate utility records for Indian Springs, housing information for this area was collected by a "windshield survey," a systematic, block-by-block canvassing of housing units by teams operating from automobiles (Lowe, Pittenger and Walker 1977). As a consequence, Indian Springs is not included in the test results reported in this paper.

The preliminary fieldwork indicated that substantial differences in *PPH* could be expected across the communities in the study area. Thus, a random selection of units from the housing unit file was drawn separately for each community, based on the number of housing units in each community. A conservative approach was used to determine the size of each community's sample. It assumed a 5% margin of error, a significance level of .05 and interest in a dichotomous variable with a 50-50 distribution (Cochran 1977). Once the initial size was determined, an additional 15% was added to allow for missing cases. The final sample size for Amargosa Valley was 175 housing units, for Beatty it was 222, and for Pahrump, 355.

Local experts were initially identified through the contact network on the basis of their experience in community activities and their familiarity with local residents. Each potential expert was interviewed and asked to complete a form designed to assess their qualifications. A written explanation of the project and specific instructions regarding the data collection process were provided and discussed. The persons selected as local experts were given instructions regarding confidentiality. For this project, we found that the "meter readers" employed by the local utilities constituted a good source of local experts. The local experts were provided with the sample set of housing units for their respective communities. In most cases, two local experts worked together, which made it possible to verify the accuracy of information as it was recorded. For each unit, the local experts communicated to the researcher only the number of persons in the household as of July 15, 1990, the age (using eight age groups) and gender of each household member, and the retirement status of any member fifty years of age and over. If either of the two local experts was unsure about the composition of a given household, another member of the community was contacted to confirm the data. In the case where the composition of any part of the household could not be confirmed, "data unknown" was recorded for the entire unit. The data were recorded on a form that listed and identified the sample units by an attribute number (designated according to location on the housing unit map), the electrical meter number assigned to the unit, and the type of housing unit. All residential units, including those identified as "burned down" or otherwise destroyed, unoccupied or "removed from pad" (in the case of mobile homes and trailers) are considered part of the final sample. Units identified as "not a residence" were eliminated from the frame and not included in the sample. There were a few units for which data were unknown. These units are not included in the final sample, which may cause some slight bias.

7. RESULTS

The first data product is the number of households, which is derived directly from the active meter records, screened and classified by utility company personnel. Table 1 displays these figures by community along with other results that are discussed later.

Table 1 also provides the estimated *PPH*, which is taken from aggregate number of persons identified in the occupied sample units by the local experts. Also found in this table is the estimated household population of each community, which was found by applying the HUM formula to the household and *PPH* components. There were no group quarters identified in any of the communities.

Table 1

Sample Characteristics and Results of the Accuracy Test*

Community	Households	Estimated 1990			1990 Census Count	95% Confidence Interval	
		PPH	*SE*	Population		Low	High
Amargosa Valley	326	2.58	.11	841	853	771	911
Beatty	672	2.43	.10	1,633	1,623	1,501	1,765
Pahrump	3,224	2.23	.06	7,190	7,425	6,810	7,569

* The Estimated data and confidence intervals are produced by the procedures described in the text. The 1990 census counts are taken from Table 3 in the "1990 Census Extract, Nevada, Public Law 94-171 Data," dated February 11, 1991 and distributed by Betty McNeal, Nevada State Data Center Librarian, Nevada State Library and Archives, Capitol Complex, Carson City, Nevada 89710. The count for the area "Amargosa Valley is made up of the 1990 population reported for Nye county's Amargosa Valley Division (761) and Crystal Division (92). The count for the area "Beatty" is taken from the Beatty Census Designated Place and the count for the area "Pahrump" is taken from the Pahrump Division of Nye county.

8. MEASURING UNCERTAINTY IN THE ESTIMATES

One major advantage of estimates based on random sampling is that confidence intervals can be generated. Rives (1982) advocates this approach. However, he did not consider the use of local experts and believed that his suggestion would only be followed in exceptional circumstances because of the high expense associated with traditional surveys. This was also noted by Morrison (1982) and Lee and Goldsmith (1982) in their critical review of Rives' suggestion.

In the case of the local expert procedure, the "statistic" is the *PPH* value, which in practice would vary from sample to sample depending on the variation in *PPH* values. Our interest is less in the *PPH* values than in the estimate of population, however, so we use a simple transformation introduced by Espenshade and Tayman (1982) and used more recently by Swanson (1989) to place the confidence interval originally generated for a given community's *PPH* value around each of the community population estimates.

Let

$$P = \text{estimated household population},$$
$$N = \text{number of households},$$
$$PPH = \text{estimated persons per household}.$$

Then

$$\text{lower limit } (P) = (N) * (PPH - (t_{n-2}, \alpha/2) * (se)),$$
$$\text{upper limit } (P) = (N) * (PPH + (t_{n-2}, \alpha/2) * (se)),$$

where

n = number of households sampled,
α = level of significance desired,
se = standard error of the estimated *PPH*,
t_{n-2} = $(\alpha/2)$100th percentile of the t distribution,with $(n - 2)$ degrees of freedom.

As an example, using a significance level of .05, the corresponding 95% confidence interval for the estimated 1990 population of Pahrump (7,190) is

$$\text{lower limit} = 6{,}810 = (3{,}224) * (2.23 - (1.96 * .06)),$$
$$\text{upper limit} = 7{,}569 = (3{,}224) * (2.23 + (1.96 * .06)).$$

9. TEST OF ACCURACY

Before turning to the test results, which are also included in Table 1, some data qualifications require discussion. The single most problematic issue in terms of comparing the estimates with the 1990 census results lies in the fact that the Bureau of the Census does not recognize the "tax districts" as administrative boundaries for the communities in the study area. This means that the Bureau's "statistical" geography must be used, which requires some adjustments so that the geography used for purposes of the impact analysis matches that used by the Bureau.

In terms of these adjustments, the area identified as Amargosa Valley for purposes of the impact study is known to vary from the Amargosa Valley Census Division of Nye county used by the Bureau in that the study's definition includes the Crystal Census Division of Nye county. Fortunately, this is a case where two pieces of statistical geography used by the Bureau can be combined to virtually match that used in the impact study. Thus, the 1990 census population counts shown in Table 1 for the Amargosa Valley include the Crystal Division along with the Amargosa Valley Division. "Beatty" is another area that is known to vary in terms of geography. It is identified as both a Census Designated Place and as the Beatty Census Division of Nye county by the Bureau. In this situation, it is the Census Designated Place that corresponds very closely to the definition of Beatty used in the impact study. Thus, the 1990 census population count for Beatty shown in Table 1 is for the Beatty CDP.

The third community, Pahrump, is identified as a Census Division of Nye county. This piece of statistical geography used by the Bureau is virtually identical to that used in the impact study. Consequently, the 1990 census population found in Table 1 for Pahrump is that given for this division of Nye county.

There are other differences between the estimates and the 1990 census figures. The official date of the census count is April 1st; the estimates are for July 15th. In terms of this difference, seasonal effects are believed to be very slight for the communities in question. With the exception of the outflow of some "snowbirds," who may have been counted in the study area because they had no usual residence elsewhere, there were no known migration streams of any consequence between April and July. Similarly, the other components of population change were slight.

Had the Bureau found transient persons with no usual residence elsewhere, the estimation procedure is likely to have missed them. These differences would also impact housing unit counts. If a transient person, identified as a resident for purposes of the decennial census, is found in a recreational vehicle it would be included in the community's "other" housing stock by the Bureau. Such accommodations would not be included in the data derived from the residential electrical meter records.

We believe, however, that such instances are rare and, further, that the test results are not confounded by comparing a household population with a population that resides mainly in households but also, to some extent, in group quarters.

The results of the test of accuracy are also summarized in Table 1, along with the "low" and "high" estimates corresponding to the 95 percent confidence interval placed around each community's estimated population. The estimated population is very close to the population reported by the Bureau. Overall, the mean absolute difference is 86 persons and the mean absolute percent difference is 1.7.

The three confidence intervals contain the 1990 census population in each of the three communities, respectively. This finding is of special interest because the intervals are relatively narrow for a 95 percent level of confidence. On average, the width, as measured from the estimated population to either boundary is 7.2 percent of the estimated population. This suggests that confidence intervals constructed around the estimates derived from this variation of the HUM are meaningful, even in the presence of some unknown level of nonsampling error.

Two of the three communities are underestimated. In the case of Pahrump, it appears that the estimation technique was not able to capture all of the recent growth that appears to be spilling from the Las Vegas Valley into the Pahrump area. It is not known to what extent this was due to missing households on the frame and what was due to underestimating Pahrump's *PPH* value.

10. SUMMARY

While the local expert procedure may not provide satisfactory population estimates in all small, rural areas (*e.g.*, vacation spots, with a high incidence of seasonal housing units and privately owned rental units), it appears to hold promise based on the data for the area included in this study. As with any estimation technique, the key criteria for determining if it could be implemented elsewhere revolve around the possibility of obtaining the required data and implementing the procedure within available funding. In the case of the local expert procedure, this would mean that utility data can provide the number of households and be used as a sample frame. Once a sample was selected, the procedure's effectiveness would depend on the recruitment and knowledge of local experts. If these criteria can be met, the procedure would seem to be feasible. The next step would be to determine how accurate it is in a given application.

We were not able to evaluate the accuracy of the age and other composition data estimated through the procedure at the time of this writing because these data were not yet available from the 1990 decennial census. However, we are encouraged by the test results for the total population, which indicate that the procedure has the potential for highly accurate estimates, even in small, rural areas experiencing rapid change.

ACKNOWLEDGMENT

The authors are grateful for comments made by Mark Flotow and anonymous reviewers on earlier drafts of this paper.

REFERENCES

BROWN, R., GEERTSEN, H., and KRANNICH, R. (1989). Community satisfaction and social integration in a boomtown: A longitudinal analysis. *Rural Sociology*, 54, 568-586.

BYERLY, E. (1990). State and local agencies preparing population and housing estimates. *Current Population Reports* Series P-25, 1063. Washington, DC: U.S. Bureau of the Census.

CARLSON, J., SWANSON, D., and WILLIAMS, C. (1990). The development of small area socioeconomic data to be utilized for impact analysis: Rural, Southern Nevada. In *High Level Radioactive Waste Management Proceedings of the 1990 International Conference.* LaGrange Park, Illinois: American Nuclear Society, 985-990.

COCHRAN, W. (1977). *Sampling Techniques*, (3rd Edition). New York: Wiley.

ESPENSHADE, T., and TAYMAN, J. (1982). Confidence intervals for postcensal state population estimates. *Demography*, 19, 191-210.

FREUDENBURG, W. (1982). Social impact assessment. In *Rural Society in the U.S.: Issues for the 1980s*, (Eds. D. Dillman and D. Hobbs). Boulder, Colorado: Westview Press, 296-303.

KRANNICH, R., BERRY, E., and GREDIER, T. (1989). Fear of crime in rapidly changing rural communities: A longitudinal analysis. *Rural Sociology*, 54, 195-212.

KRANNICH, R., and GREDIER, T. (1984). Personal well-being in rapid growth and stable communities: Multiple indicators and contrasting results. *Rural Sociology*, 49, 541-552.

LEE, E., and GOLDSMITH, H. (1982). Evaluation and synopsis. In *Population Estimates: Methods for Small Area Analysis*, (Eds. E. Lee and H. Goldsmith). Beverly Hills, California: Sage, 113-118.

LOWE, T., PITTENGER, D., and WALKER, J. (1977). Making the housing unit method work: A progress report. Presented at the Annual Meeting of the Population Association of America.

LOWE, T., WEISSER, L., and MYERS, B. (1984). A special consideration in improving housing unit method estimates: The interaction effect. Presented at the Annual Meeting of the Population Association of America.

MARTIN, J., and SEROW, W. (1979). Virginia's state-local cooperative program: conflict and cooperation in producing population estimates. State Government, 182-186.

MORRISON, P. (1982). Alternatives for monitoring local demographic change. In *Population Estimates: Methods for Small Area Analysis*, (Eds. E. Lee and H. Goldsmith). Beverly Hills, California: Sage, 97-111.

MUNSELL, K. (1988). Towns cope with hazardous wastes. *Small Town*, 18, 30.

PITTENGER, D. (1976). *Projecting State and Local Populations*. Cambridge, Massachusetts: Ballinger Press.

POGGIE, J. (1972). Toward quality control in key informant data. *Human Organization*, 31, 23-30.

RIVES, N. (1982). Assessment of a survey approach. In *Population Estimates: Methods for Small Area Analysis*, (Eds. E. Lee and H. Goldsmith). Beverly Hills, California: Sage, 79-96.

RIVES, N., and SEROW, W. (1984). *Introduction to Applied Demography Data Sources and Estimation Techniques*. Beverly Hills, California: Sage.

SCHLEIFER, S. (1986). Trends in attitudes toward and participation in survey research. *Public Opinion Quarterly*, 50, 17-26.

SEIDLER, J. (1974). On using informants: A technique for collecting quantitative data and controlling measurement error in organizational analysis. *American Sociological Review*, 39, 816-831.

SMITH, S. (1986). A review and evaluation of the housing unit method of population estimation. *Journal of the American Statistical Association*, 81, 287-296.

SMITH, S., and LEWIS, B. (1983). Some new techniques for applying the housing unit method of local population estimation: Further evidence. *Demography*, 20, 407-413.

SMITH, S., and LEWIS, B. (1980). Some new techniques for applying the housing unit method of local population estimation. *Demography*, 17, 323-339.

SMITH, S., and MANDELL, M. (1984). A comparison of local population estimates: The housing unit method versus component II, regression, and administrative records. *Journal of the American Statistical Association*, 79, 282-289.

STARSINIC, D., and ZITTER, M. (1968). Accuracy of the housing unit method in preparing population estimates for cities. *Demography*, 5, 475-484.

SUMMERS, G. (1982). *Industrialization. In Rural Society in the U.S.: Issues for the 1980s*, (Eds. D.Dillman and D. Hobbs). Boulder, Colorado: Westview Press, 164-174.

SWANSON, D. (1989). Confidence intervals for postcensal population estimates: A case study for local areas. *Survey Methodology*, 15, 271-280.

SWANSON, D., BAKER, B., and VAN PATTEN, J. (1983). Conceptual and practical features of the housing unit method. Presented at the Annual Meeting of the Population Association of America.

U.S. BUREAU OF THE CENSUS (1978). State and local agencies preparing population estimates and projections: Survey of 1975-76. *Current Population Reports* P-25, 725. Washington, DC: U.S. Bureau of the Census.

U.S. DEPARTMENT OF ENERGY (1988). Section 175 report. *Secretary of Energy's Report to the Congress Pursuant to Section 175 of the Nuclear Waste Policy Act, As Amended*. Washington, DC: Office of Civilian Radioactive Waste Management, U.S. Department of Energy.

Survey Methodology, June 1992
Vol. 18, No. 1, pp. 165–174
Statistics Canada

Single Stage Cluster Sampling in Prevalence-Incidence Surveys: Some Issues Suggested by the Shanghai Survey of Alzheimer's Disease and Dementia

ZHISEN XIA, PAUL S. LEVY, ELENA S.H. YU, ZHENGYU WANG, and MINGYUAN ZHANG[1]

ABSTRACT

The scenario considered here is that of a sample survey having the following two major objectives: (1) identification for future follow up studies of n^* subjects in each of H subdomains, and (2) estimation as of this time of conduct of the survey of the level of some characteristic in each of these subdomains. An additional constraint imposed here is that the sample design is restricted to single stage cluster sampling. A variation of single stage cluster sampling called *telescopic single stage cluster sampling* (TSSCS) had been proposed in an earlier paper (Levy *et al.* 1989) as a cost effective method of identifying n^* individuals in each sub domain and, in this article, we investigate the statistical properties of TSSCS in crossectional estimation of the level of a population characteristic. In particular, TSSCS is compared to ordinary single stage cluster sampling (OSSCS) with respect to the reliability of estimates at fixed cost. Motivation for this investigation comes from problems faced during the statistical design of the Shanghai Survey of Alzheimer's Disease and Dementia (SSADD), an epidemiological study of the prevalence and incidence of Alzheimer's disease and dementia.

KEY WORDS: Single stage cluster sampling; Prevalence estimation; Telescopic single stage cluster sampling; Alzheimer's disease; Dementia.

1. BACKGROUND AND INTRODUCTION

Many studies have both a crossectional component in which the levels of quantitative variables or prevalences of dichotomous variables are estimated by means of a sample survey, and a longitudinal component in which a cohort of individuals is identified by means of the same sample survey and followed over a defined period for subsequent events. This type of study is especially common in the field of epidemiology in which estimates of the prevalence of a disease or condition are required both for the study population as a whole as well as for defined subgroups of it, and a sufficient number of individuals initially free of the disease or condition need to be identified within each of the defined subgroups for future estimation of the incidence of the disease or condition (*cf.* Kannel 1966).

Design of a cost efficient sampling plan for such studies poses a challenge since sufficient numbers of individuals within each domain must be selected, often under some type of cluster sampling scheme, to ensure reliable estimation of both the prevalence and incidences discussed above. In this report, which has been motivated by a recent study conducted in China, we discuss these issues of sample design under a particular type of cluster sampling (single stage cluster sampling).

---

[1] Zhisen Xia, Paul S. Levy and Zhengyu Wang, School of Public Health, University of Illinois at Chicago, P.O. Box 6998, Chicago IL. 60680, U.S.A., Elena S.H. Yu, School of Public Health, San Diego State University, San Diego, California, U.S.A., Mingyuan Zhang, Shanghai Institute of Mental Health, Shanghai, Peoples Republic of China.

2. STATISTICAL FORMULATION

Let us suppose that a population consists of N individuals divided into H mutually exclusive subdomains, each containing N_h individuals ($h = 1, \ldots, H$). Suppose further that the population is grouped into M clusters which will comprise the sampling units for the survey. Let us assume that sampling of the clusters will be according to ordinary single-stage cluster sampling (*i.e.*, simple random sampling of clusters followed by selection of all individuals within each sample cluster.)

If we wish to identify with $100 \times (1 - \alpha)\%$ confidence at least n_h^* individuals in a particular domain, h, then the following number, m_h', of clusters must be selected (*cf.* Levy *et al.* 1989):

$$m_h' = \left[A_h + \left(A_h^2 + \frac{n_h^*}{\bar{N}_h}\right)^{1/2}\right]^2, \tag{1}$$

where

N_{hi} = the number of individuals in domain h, cluster i, ($i = 1, \ldots, M$),

$$\bar{N}_h = \sum_{i=1}^{M} N_{hi}/M,$$

$$V_{N_h} = \sigma_{N_h}/\bar{N}_h,$$

$$\sigma_{N_h}^2 = \sum_{i=1}^{M} (N_{hi} - \bar{N}_h)^2/(M - 1),$$

z_α = the 100α'th percentile of the normal distribution

and

$$A_h = |z_\alpha| \times V_{N_h}/2.$$

The above assumes that the N_{hi} are normally distributed over the M clusters. Also, the number, n_h^*, of individuals needed in domain h is based on statistical considerations relevant to the longitudinal component of the study. For example, it could be based on the expected occurrence rate of the event of interest in the follow up period and the precision required for the estimate of this occurrence rate.

If one also wishes to estimate with $100 \times (1 - \alpha)\%$ confidence the total or mean level of some variable $\mathcal{X}$ to within $100 \times \epsilon\%$ of its true value for each domain, h, then one would require sampling of the following number, M_h'', of clusters in domain h;

$$m_h'' = \frac{z_{1-\alpha/2}^2 \, M \, V_{hx}^2}{z_{1-\alpha/2}^2 \, V_{hx}^2 + (M - 1)\epsilon^2}, \tag{2}$$

where,

X_{hij} = the level of variable $\mathcal{X}$ for individual j within domain h of cluster i ($j = 1, \ldots, N_{hi}; i = 1, \ldots, M$),

$$X_{hi} = \sum_{j=1}^{N_{hi}} X_{hij},$$

$$\bar{X}_h = \sum_{i=1}^{M} X_{hi}/M,$$

$$\sigma_{hx}^2 = \sum_{i=1}^{M} (X_{hi} - \bar{X}_h)^2/M,$$

and

$$V_{hx}^2 = \sigma_{hx}^2/\bar{X}_h^2.$$

For both of the specifications stated above to be satisfied within each domain, it follows that we would require m_h clusters to be sampled where for $h = 1, \ldots, H$,

$$m_h = \max(m_h', m_h''). \tag{3}$$

Without loss of generality, we can relabel the domains in order of increasing required m_h (*i.e.*, $m_1 \leq m_2 \leq \ldots \leq m_H$).

Finally, in order for both of the specifications to be satisfied in each of the H domains under an ordinary single stage cluster sampling design, the number, m, of clusters required to be sampled would be m_H. We note again that in ordinary single stage cluster sampling, every individual in every sample cluster is sampled. Thus, while the specifications of sample size are met minimally in domain H, the domain requiring the largest number of sample clusters, they are more than met in the other domains: $1, \ldots, H - 1$. This inclusion in domains other than H of more individuals than are actually required could result in a survey that has overly expensive field costs.

The alternative to ordinary single-stage cluster sampling that is generally used to avoid this needless expense would be a two-stage cluster sampling design with different second stage sampling fractions (*i.e.*, over sampling) in each domain. Given, however, a scenario in which it is not feasible to subsample at all within clusters, a methodology called *single stage telescopic cluster sampling* (SSTCS) was proposed in an earlier publication (Levy *et al.* 1989) which allowed the *eligibility rule* (*i.e.*, the rule that determines which individuals are eligible for inclusion in the sample) to vary over the sample clusters. In this design, the particular domains included in the sample would not be the same for each sample cluster. This earlier publication demonstrated the usefulness of single stage telescopic sampling in surveys which have as major objective the identification for future longitudinal follow up of a certain number of individuals in each of several domains. In this report, we will characterize the properties of estimates from this type of design and compare them to estimates from ordinary single-stage cluster sampling.

3. TELESCOPIC SINGLE-STAGE CLUSTER SAMPLING

3.1 Sampling of Clusters

As mentioned above, single-stage telescopic cluster sampling is proposed as a cost saving alternative to ordinary single-stage cluster sampling in situations where it is not feasible to subsample within sample clusters, and is performed as follows. If there are H mutually exclusive and exhaustive domains for which estimates are desired, and if m clusters are to be sampled, the m sample clusters are divided randomly into m_1^* *type* 1 clusters, m_2^* *type* 2 clusters, ..., and m_H^* *type* H clusters having the following properties: A type h cluster $(h = 1, \ldots, H)$ as illustrated below has as eligible sample persons individuals in domains $h, h + 1, \ldots, H$, but not in domains h' where $h' < h$.

Cluster Type	Domains Sampled 1	2	h	H
1	+	+	+	+
2	−	+	+	+
h	−	−	+	+
H	−	−	−	+

"+" = domain sampled "−" = domain not sampled.

The term *telescopic* was suggested by the appearance of the above diagram.

The number, m_h^*, of type h clusters is generally determined according to the following strategies: Suppose that under single-stage cluster sampling, a sample of m_h clusters as determined by relation (3) is required for domain h, $(h = 1, \ldots, H)$; and, again supposing that $m_1 \leq m_2 \ldots \leq m_H$, we would let:

$$m_1^* = m_1; \quad \text{and} \quad m_h^* = m_h - m_{h-1} \quad \text{for} \quad h = 2, \ldots, H.$$

Clearly, this allocation results in a total of m_H sample clusters being selected, with elements in each domain, h, being sampled in m_h sample clusters, exactly the number of clusters required to achieve the specifications placed on the reliability of estimates and the identification of individuals for future follow up. As discussed above, if ordinary single-stage cluster sampling (OSSCS) were used, a sample of m_H clusters would be needed to meet specifications in domain H, but this would entail individuals in the other domains also being sampled in m_H clusters in excess of that needed to meet the stated specifications.

3.2 Characterization of Estimates

Let

$$\sigma_{hkx} = \sum_{i=1}^{M} (X_{hi} - \bar{X}_h)(X_{ki} - \bar{X}_k)/M,$$

$S_h = \{i_1, i_2, \ldots, i_{m_h}\}$ = the set of sample clusters having eligible persons in domain h.

The following results can then be obtained from combinatorial theory.

1. The estimated total, x'_{tel}, under TSSCS of a population total X is given by

$$x'_{\text{tel}} = \sum_{h=1}^{H} x'_h, \tag{4}$$

where x'_h is given by

$$x'_h = (M/m_h) \sum_{i \in S_h} X_{hi}.$$

2. The mean, $E(x'_{\text{tel}})$, and variance, $\text{Var}(x'_{\text{tel}})$, of x'_{tel} are given by

$$E(x'_{tel}) = X, \tag{5}$$

$$\text{Var}(x'_{tel}) = \sum_{h=1}^{H} \frac{M^2}{m_h} \left(\frac{M - m_h}{M - 1} \right) \left(\sigma^2_{hx} + 2 \sum_{k<h} \sigma_{hkx} \right). \tag{6}$$

These relationships follow in a straightforward way from combinatorial theory.

4. COST COMPARISONS BETWEEN OSSCS AND TSSCS

We can examine the comparative costs of OSSCS *vs.* TSSS by considering the following simple cost function that would be associated with OSSCS:

$$C_0 = C_1 m_H + C_2 m_H (\bar{N}_1 + \bar{N}_2 + \ldots + \bar{N}_H) = m_H \left(C_1 + C_2 \sum_{h=1}^{H} \bar{N}_h \right), \tag{7}$$

where C_0 is the expected cost, C_1 is the cost component associated with clusters (*e.g.*, travel to and from cluster, procurement of the list of enumeration units in the cluster, preparation of materials for field work within the cluster, *etc.*) and C_2 is the cost component associated with listing units (primarily travel between listing units and interviewing). It should also be noted that the expression, $\sum_{h=1}^{H} \bar{N}_h$, is the average number of listing units per cluster. Again, throughout this discussion the listing units are the individuals themselves. The analogous expected cost, C_t, associated with telescopic sampling would then be given by:

$$C_t = C_1 m_H + C_2 (m_1 \bar{N}_1 + m_2 \bar{N}_2 + \ldots + m_H \bar{N}_H) = m_H \left(C_1 + C_2 \sum_{h=1}^{H} \gamma_h \bar{N}_h \right), \tag{8}$$

where $\gamma_h = m_h / m_H$ (which is ≤ 1). Thus, the cost, C_t, associated with TSSCS is less than or equal to that associated with an OSSCS of the same number of clusters with the difference being equal to

$$C_2 m_H \sum_{h=1}^{H} (1 - \gamma_h) \bar{N}_h.$$

The most important comparison between the two sample designs, in many instances, would be that involving their performance at equivalent cost in estimating the overall level, X, of a characteristic, $\mathcal{X}$. An estimator, x'_{ord}, based on an OSSCS of m_H clusters (the number required to meet the specifications within each domain) would have variance given by:

$$\operatorname{Var}(x'_{\text{ord}}) = \frac{M^2}{m_H}\left(\frac{M - m_H}{M - 1}\right) \sum_{h=1}^{H} \left(\sigma_{hx}^2 + 2\sum_{k<h} \sigma_{hkx}\right). \tag{9}$$

This is not the usual form of the variance (*cf.* Levy and Lemeshow 1991, chapter 9), but is an algebraically equivalent form that can be compared directly with the variance of x'_{tel} based on a TSSCS design with m_H clusters sampled (equation (6)). The difference between these two variances is given by

$$\operatorname{Var}(x'_{\text{tel}}) - \operatorname{Var}(x'_{\text{ord}}) = \frac{M^3}{M - 1} \sum_{h=1}^{H} \left(\frac{m_H - m_h}{m_H m_h}\right) \sum_{h=1}^{H} \left(\sigma_{hx}^2 + 2\sum_{k<h} \sigma_{hkx}\right), \tag{10}$$

which is greater than or equal to zero (0). This is not surprising since an OSSCS of m_H clusters will invariably result in a larger overall sample size than a TSSCS of the same number of clusters.

Although an OSSCS of m_H clusters will result in an estimator, x'_{ord}, which has a lower variance than the estimator, x'_{tel}, resulting from a TSSCS of the same number, m_H, of clusters, it does so at a higher cost. For this reason, it is more reasonable to compare x'_{tel} based on a sample of m_H clusters to x'_{ord} based on a sample of m^* clusters where m^* is the number of clusters that can be sampled from an OSSCS design at cost equivalent to that based on a TSSCS design having m_H sample clusters. From equations (7) and (8), it follows that m^* is given by:

$$m^* = m_H \left(\frac{1 + \dfrac{C_2}{C_1} \displaystyle\sum_{h=1}^{H} \gamma_h \bar{N}_h}{1 + \dfrac{C_2}{C_1} \displaystyle\sum_{h=1}^{H} \bar{N}_h} \right). \tag{11}$$

It should be noted that

(1) $m^* \le m_H$.

(2) As $C_2/C_1 \to \infty$, then $m^* \to \bar{m}_w$

where, $$\bar{m}_w = \sum_{h=1}^{H} m_h \bar{N}_h \Big/ \sum_{h=1}^{H} \bar{N}_h.$$

(3) As $C_2 \backslash C_1 \to 0$, then $m^* \to m_H$

and

(4) m^* decreases monotonically with increase in C_2/C_1 which implies that $\bar{m}_w \le m^* \le m_H$.

From the above analysis, we note that at a cost equivalent to that of a TSSCS of m_H clusters, the variance of x'_{ord} (ignoring the finite population correction) will be inflated by at most a factor equal to $m_H/\bar{m}_w$ over that which would have been obtained from an OSSCS of m_H clusters, where $\bar{m}_w$ is a weighted mean of the m_h clusters required within each domain for the domain specific specifications to be met. The weights in this instance are the $\bar{N}_h$, which are the average number of individuals within each particular domain. It should be noted also that the reduction in effective sample size of an OSSCS equivalent in cost to a TSSCS increases with increase in C_2/C_1, which is essentially the ratio of the cost of extracting information from sample individuals to that of preparing the sample clusters for the survey. This makes sense intuitively.

The issues discussed above are illustrated in the next section with data from the Shanghai Survey of Alzheimer's Disease and Dementia.

5. THE SHANGHAI SURVEY OF ALZHEIMER'S DISEASE AND DEMENTIA

The SSADD was planned in 1986 having as major objectives: (1) estimation of the prevalence of physical and mental impairments including Alzheimer's and other dementing diseases among persons in each of three age groups (55-64 yrs/65-74 years/ and 75 yrs. and older) in the Jing-An district of Shanghai, China, and (2) identification of approximately 1,400 persons in each of these 3 age groups for future determination of the incidence of these conditions. Jing-An, is one of twelve districts comprising the city of Shanghai, and was chosen as the target area because of its relatively large and stable population of elderly and its proximity to the Shanghai Institute of Mental Health which was responsible for the field work. Findings from this study have been discussed by Zhang *et al.* (1990) and by Yu *et al.* (1989). Methodological issues have been discussed by Levy *et al.* (1988 and 1989).

The clusters in this survey are administrative entities called *neighborhood groups* consisting of geographically contiguous households having a well defined social and political structure. The strategy was to involve the leaders of neighborhood groups selected in the sample in the identification and recruitment of eligible persons. At the time of the planning of the survey, there were 4,066 neighborhood groups within the Jing-An District. This particular population of aging and elderly Chinese generally had a low level of education and had experienced in their lifetimes repeated periods of political upheaval and repression (*e.g.*, the Warlords, the Japanese invasion, the Cultural Revolution), where being singled out or selected often had adverse consequences. For these reasons, it was felt strongly, especially by the local Chinese members of the research team who were most familiar with the target population, that any attempt to subsample persons in the target age groups within neighborhood groups that fall into the sample would compromise response rates and overall cooperation.

Restricted to single stage cluster sampling and faced with a very tight deadline for designing the sample, the member of the study team responsible for the sample design (PSL) proposed a heuristic method that would result with reasonable certainty in the identification of 1,400 individuals within each of the three target age groups. The resulting design was essentially a TSSCS in which 446 neighborhood groups were sampled. For details of this design, the reader is referred to the publications on the SSADD cited above. It should be emphasized that the resulting design was chosen purely on heuristic grounds and long before the theory behind this methodology was developed.

Of the 446 neighborhood groups sampled, 149 were designated as type 1, and 136 of these contained at least 1 person in the target age group (55 years and above). Since only the type 1 clusters have as eligible respondents all persons in each of the 3 target age groups, they can be used to estimate all of the parameters needed to evaluate the cost effectiveness of TSSCS relative to OSSCS. In the ensuing discussion, we will use the data from these 136 clusters to illustrate numerically how, on the basis of available "pilot" data, comparisons can be made between OSSCS and TSSCS with respect to cost effectiveness. From this sample of 136 clusters, we have for each domain, h, estimates of relevant parameters as shown below:

Age	$\bar{N}_h$	V_{N_h}	$\bar{X}_h$	V_{hx}	$\sum_{k<h} \sigma_{hkx}$
55-64	10.985	.485	.125	2.991	0.000
65-74	8.088	.513	.360	2.357	0.190
75+	3.478	.643	.456	1.665	0.296.

If we wish to identify with 95% confidence at least 1,400 persons in each age group, then from relation (1) and the data shown above, we would have

$$A_1 = 1.645 \times 0.485/2 = 0.3989$$

and

$$m_1' = \left[0.3989 + \left((0.3989)^2 + \frac{1{,}400}{10.985}\right)^{1/2}\right]^2 = 136.78 \approx 137.$$

Similarly, $m_2' \approx 185$, and $m_3' \approx 419$.

Let us suppose that for each of the three age groups, we wish to estimate with 80% confidence to within 30% of its true value the proportion, $\bar{X}_h$, of persons showing evidence of cognitive dysfunction as judged by a score below 18 on the Mini Mental State Examination (MMSE), which is a screening test for cognitive dysfunction. From these same data, we have the following estimates of the parameters necessary to determine the number of sample clusters required to meet this specification:

From relation (2), with M = 4,066, $\epsilon = 0.30$, and $z_{1-\alpha/2} = 1.28$, we have the following values of m_h'':

$$m_1'' = 157; \quad m_2'' = 99; \quad m_3'' = 50$$

and from relation (3), the number, m_h, of clusters required to satisfy both conditions in each domain is given by:

$$m_1 = \max(137,157) = 157; \quad m_2 = \max(185,99) = 185; \quad m_3 = \max(419,50) = 419.$$

Thus, for an OSSCS design to satisfy both specifications, the number, m, of clusters required to be sampled would be 419. Likewise, a TSSCS design having 157 type 1, 28 type 2, and 234 type 3 sample clusters would satisfy both requirements.

The cost components, C_1 and C_2, expressed in person hours, are estimated to be 20 and 2 respectively. The relatively high cost component, C_1, associated with clusters is due to the fact that once a neighborhood group is selected in the sample, many hours must be spent obtaining the list of households and persons from a central bureau and enlisting the support of the neighborhood group leaders. The cost component, C_2, of 2 person hours associated with individuals involves primarily interview and call-back activities. Thus, the field costs, C_0, associated with an OSSCS design that satisfies both specifications is (from relation (7)) 27,278 person hours as compared to a cost of 17,737 person hours (from relation (8)) associated with a TSSCS design that satisfies both specifications. This represents a 35% savings in field costs, which is substantial.

From relation (9), we calculate that the estimate, x'_{ord}, of the number of persons over all 3 age groups having evidence of a cognitive disorder based on an OSSCS of 419 sample clusters would have variance equal to 70,844, whereas x'_{tel}, the analogous estimate based on a TSSCS also with 419 clusters, would have variance equal to 122,744, which is 42% greater than the variance of the OSSCS estimate. However, an OSSCS design having the same field costs as a TSSCS design based on 419 sample clusters would permit only 208 clusters to be sampled (relation (11)). The variance of x'_{ord} based on an OSSCS design with 208 sample clusters would be estimated to be 141,733, which is 15% higher than the variance of the analogous TSSCS estimate having the same field cost. Also, the OSSCS design having 208 sample clusters would not satisfy the two specifications placed on the estimates.

6. DISCUSSION

The methodology, TSSCS, discussed here and in earlier publications, arose from a situation in which cluster sampling was clearly indicated but a definite "red light" was given to any subsampling within clusters. For the Shanghai Survey of Alzheimer's Disease and Dementia considered here, the two major objectives were to identify a certain number of individuals within each of 3 domains (age groups in this instance) and to obtain domain specific estimates meeting certain specifications pertaining to precision. Based on results presented above for this particular survey, it appears that this method could result in considerable savings in field costs without compromising objectives.

One might raise questions concerning the general applicability of this methodology. It would be of use primarily in situations where it is either not feasible or too costly to subsample clusters and the individuals do not have to be screened to determine whether they belong to one of the target domains (in the SSADD, the leadership of the sample neighborhood groups provided a list of all persons in the neighborhood group along with information on data of birth). Such scenarios may occur, for example, in surveys where data are abstracted from records by personnel sufficiently familiar with the records to abstract information, but not considered capable of sampling the records without expensive supervision. Again, in such situations, TSSCS may provide a reasonable alternative.

ACKNOWLEDGEMENTS

Research on this report was supported, in part, by grant number 1 RO1 AG10327-01 from the National Institute on Aging.

REFERENCES

KANNEL, W.B. (1966). An epidemiological study of cerebrovascular disease. In *Fifth Conference on Cerebrovascular Diseases*, (Eds. R.G. Sickert and J.P. Whisnat). New York: Grune and Stratton.

LEVY, P.S., YU, E.S.H., LIU, W.T., ZHANG, M., WANG, Z., WONG, S., and KATZMAN, R. (1988). Variation on single stage cluster sampling used in a survey of elderly people in Shanghai. *International Journal of Epidemiology*, 17, 931-933.

LEVY, P.S., YU, E.S.H., LIU, W.T., WONG, S., ZHANG, M., WANG, Z., and KATZMAN, R. (1989). Single-stage cluster sampling with a telescopic respondent rule: A variation motivated by a survey of dementia in elderly residents of Shanghai. *Statistics in Medicine*, 8, 1537-1544.

LEVY, P.S., and LEMESHOW, S. (1991). *Sampling of Populations: Methods and Applications*. New York: John Wiley and Sons.

YU, E.S.H., LIU, W.T., LEVY, P.S., ZHANG, M., KATZMAN, R., LUNG, C.T.,WONG, S., and WANG, Z. (1989). Cognitive impairment among elderly in Shanghai, China. *Journal of Gerontology Social Sciences*, 44, 97-106.

ZHANG, M., KATZMAN, R., SALMON, D., JIN, H., CAI, G., WANG, Z., QU, G.,GRANT, I., YU, E.S.H., LEVY, P.S., KLAUBER, M.R., and LIU, W.T. (1990). The prevalence of dementia and Alzheimer's disease in Shanghai China: Impact of age, gender, and education. *Annals of Neurology*, 27, 428-437.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents JOS 1991, Volume 7, Number 4

A Framework for Analyzing Categorical Survey Data with Nonresponse
David A. Binder.. 393

Inference with Survey Weights
Roderick J.A. Little.. 405

Cluster Analysis in International Comparisons
György Szilágyi.. 425

A Multivariate Time Series Analysis of Fertility, Adult Mortality, and Real Wages in Sweden 1751-1850: A Comparison of Two Different Approaches
Mats Hagnell.. 437

Census by Questionnaire - Census by Registers and Administrative Records: The Experience of Finland
Pekka Myrskylä.. 457

A Comparison of the Missing-Data Treatments in the Post-Enumeration Program
Joseph L. Schafer.. 475

Miscellanea

Natural Resource and Environmental Accounting in the National Accounts
Lindy H. Ingham.. 499

In Other Journals.. 515

Special Notes.. 519

Book Reviews.. 521

Index to Volume 7, 1991.. 531

All inquires about submissions and subscriptions should be directed to the Chief Editor:

Lars Lyberg, U/SFI, Statistics Sweden, S-115 81 Stockholm, Sweden

Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

CONTENTS

Volume 41, No. 2, 1992

		Page
Analysis of ordinal paired comparison data	*A. Agresti*	287
Robust procedures for drug combination problems with quantal responses	*T. J. Vidmar, J. W. McKean and T. P. Hettmansperger*	299
Pseudofactors: normal use to improve design and facilitate analysis	*H. Monod and R. A. Bailey*	317
Adaptive rejection sampling for Gibbs sampling	*W. R. Gilks and P. Wild*	337
A note on Wadley's problem with overdispersion	*B. J. T. Morgan and D. M. Smith*	349
Sequential application of Wilks's multivariate outlier test	*C. Caroni and P. Prescott*	355
Influence in correspondence analysis	*P. Pack and I. T. Jolliffe*	365
A reduced rank regression model for local variation in solar radiation	*C. A. Glasbey*	381
Hierarchical Bayesian analysis of changepoint problems	*B. P. Carlin, A. E. Gelfand and A. F. M. Smith*	389
General Interest Section		
Interlaboratory comparisons: round robins with random effects	*M. J. Crowder*	409
Letter to the Editors		427
Book Reviews		429
Statistical Software Reviews		
DYMO		436
SPSS/PC+		438
Statistical Algorithms		
AS 273 Comparing subsets of regressor variables	*A. J. Miller*	443
AS 274 Least squares routines to supplement those of Gentleman	*A. J. Miller*	458
AS 275 Computing the non-central χ^2 distribution function	*C. G. Ding*	478
AS 276 Normal combinatoric classification	*C. L. Dunn*	483
Remark		
AS R89 A remark on Algorithm AS 76: An integral useful in calculating non-central *t* and bivariate normal probabilities	*P. W. Goedhart and M. J. W. Jansen*	496

Printed in Great Britain at the Alden Press, Oxford

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.

1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.

1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.

1.4 Les remerciements doivent paraître à la fin du texte.

1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.

3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.

3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.

3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.

3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O, 0; 1, l).

3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).

4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.

Exemple: Cochran (1977, p. 164).

5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.