National Institute on Drug Abuse

Research 24

MONOGRAPH SERIES

Synthetic Estimates For Small Areas

Statistical Workshop Papers and Discussion

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service • Alcohol, Drug Abuse, and Mental Health Administration

Synthetic Estimates for Small Areas:

Statistical Workshop Papers and Discussion

Editor:

Joseph Steinberg

NIDA Research Monograph 24

February 1979

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service
Alcohol, Drug Abuse, and Mental Health Administration

National Institute on Drug Abuse
Division of Research
5600 Fishers Lane
Rockville, Maryland 20657

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402

Stock Number 017-024-00911-3

The NIDA Research Monograph series is prepared bv the Division of Research of the National Institute on Drug Abuse. Its primary objective is to provide critical reviews of research problem areas and techniques, the content of state-of-the-art conferences, integrative research reviews and significant original research. Its dual publication emphasis is rapid and targeted dissemination to the scientific and professional community.

Editorial Advisory Board

NIDA Research Monograph series

William Pollin, M.D.
DIRECTOR, NIDA
Marvin Snyder, Ph.D.
ACTING DIRECTOR, DIVISION OF RESEARCH, NIDA
Robert C. Petersen, Ph.D.
EDITOR-IN-CHIEF
Eleanor W. Waldrop
MANAGING EDITOR

Parklawn Building, 5600 Fishers Lane, Rockville, Maryland 20857

Synthetic Estimates for Small Areas:

Statistical Workshop Papers and Discussion

ACKNOWLEDGMENT

This monograph is based on papers presented at a workshop conducted by Response Analysis, Princeton, New Jersey, under NIDA Contract No. 271-77-3425. The workshop took place on April 13 and 14, 1978, in Princeton.

Library of Congress catalog card number 79-600067

DHEW publication number (ADM) 79-801
Printed 1979

NIDA Research Monographs are indexed in the *Index Medicus.* They are selectively included in the coverage of *BioSciences Information Service, Chemica1 Abstracts, Psychological Abstracts, and Psychophamacology Abstracts.*

Foreword

The Workshop on Synthetic Estimates was cosponsored by the National Institute on Drug Abuse (NIDA) and the National Center for Health Statistics (NCHS). The collaboration came about as follows: In 1974, an inquiry was made of NCHS by NIDA about possible methods of "triangulating" national survey data and census data to produce estimates of incidence or prevalence of drug abuse in states and local areas. Indeed, according to NCHS, there were such methods, called "synthetic estimation," and they had been explored and discussed over a span of about ten years.

A short report, Synthetic State Estimates of Disability, published by NCHS in 1968, was one of the few pieces available for the nontechnician to consult. A sparse literature in the statistical journals was available but not easy to collect or disseminate.

The two agencies felt there was need for a "consumer report" on the methods. They knew that the methods have an immediate appeal to planners, demographers, program officials, and epidemiologists charged with the task of describing conditions or estimating need in small areas. Yet neither agency was ready to recommend the methods outright because little is known about the quality of synthetic estimates. They wanted to air the strengths and weaknesses of the methods in a group of statisticians and scientists who had thought about them carefully or applied them to real situations of need. Thus the idea of holding a workshop was born.

NCHS is the agency in the Federal Statistical System that has major responsibility for compiling, analyzing, and disseminating general purpose national health and vital statistics. In recent years, the demand for health statistics for small areas has greatly increased, and producing local area statistics has emerged as one of the Center's most difficult and pressing statistical problems. NIDA has responsibility for providing national statistics on nonmedical drug use and its consequences. Its support of State programs in treatment and prevention has created the need for data reflecting conditions at that level.

Most of NCHS's data systems are incapable of producing local area statistics. The exceptions, those based on complete counts of the population, include the birth and death registration systems, and the data systems for producing health establishment and health manpower statistics. On the other hand, the capabilities of NCHS's sample data systems are limited to producing national estimates, and estimates for the geographic regions and divisions and the larger standard metropolitan statistical areas (SMSA's). Priority was not given to local area statistics when the sample data systems were originally designed. In most instances, the cost effects would have been prohibitive.

Similarly, NIDA has found it prohibitively expensive to require States to conduct their own surveys to establish need. The Client Oriented Data Acquisition Process (CODAP) produces information at the State and SMSA level on treatment admissions and discharges, but other systems provide only national estimates or data on a limited set of local areas.

Local area health data are increasingly needed to implement the programs legislated by Congress. However, changes in the appropriations for health statistics programs have not kept pace with the needs for new data and new data priorities. Therefore, agencies are looking for more cost-effective methods for producing them.

Neither NIDA nor NCHS is committed to synthetic estimation as the keystone of its policy for producing small area statistics. At present, NCHS is investigating two other strategies in addition to synthetic estimation. One of these is the Cooperative Health Statistics System. In this approach, State data systems serve as building blocks for national sample designs and methods for producing local area data. Currently NCHS is exploring the cost and error effects of network surveys, and of computerized telephone surveys on random digit dialing.

It is our belief that we have assembled the outstanding workers in the field of synthetic estimation for this workshop. We feel that the papers, and the editing by Joseph Steinberg, have resulted in a landmark publication. We hope that future users or potential users of the methods will find this volume a solid foundation for their efforts.

Louise G. Richards, Ph.D.
National Institute on Drug Abuse

Monroe G. Sirken, Ph.D.
National Center for Health Statistics

Contents

PART III

PART IV

Introduction

Joseph Steinberg

There are many and varied needs for small area data. Traditionally, this has led to consideration of large-scale data collection as the basis for satisfying the need. On occasion, a method has been tried that provided estimates for a number of individual areas on the basis of a direct collection of data for the desired characteristic for only a sample of areas and data on a related characteristic for each area. The Radio Listening Survey, discussed in Hansen, Hurwitz and Madow (1953) is an illustration of this approach used in the early 1940's. Similarly, Lillian Madow used a derived method for providing small area data in a report of the Advertising Research Foundation (1956). There has been an increase in the use of a variety of procedures for small area estimation since the National Center for Health Statistics (1968) published derived "synthetic estimates."

"'Synthetic estimate' is a label that has been given to the product of a class of devices that yield estimates of a target statistic for specific subnational areas, using descriptive data for the specific area in combination with average values of the target statistic for national or regional territory." This is the way Simmons (1977)) who coined the term, described the technique which is the focus of this WORKSHOP ON SYNTHETIC ESTIMATES FOR SMALL AREAS.

Discussion of synthetic estimates evokes a great deal of enthusiasm by some and skepticism by others. The Workshop provided a forum for sharing experiences of what is the current state of the art in methodology and in application. An additional purpose of the Workshop was to suggest refinements of estimating procedure beyond what is currently known.

Invited papers and remarks of invited discussants were the Workshop framework. Extensive informal discussion also helped to serve the purposes of the conference. The papers, invited discussion, and abstracts of the informal discussion constitute the body of this volume. Papers and associated discussion have been grouped into four parts. A historical overview is the core of Part I. Part II consists of papers on methodological contributions. Groupings of applications constitute Parts III and IV.

Different types of strategies for providing local area estimates were discussed in Levy's paper, which presents a historical perspective of efforts in the past decade. The papers by Schaible and Royall deal with refinements in estimation procedures and use of models. The possibilities in the use of composite estimators were also indicated in some of the work presented by Fay and received consideration during the informal discussion of Froland's paper.

How to devise useful subsets of a population to permit the best application of synthetic estimates received attention. The degree of homogeneity within classes across areas was identified as a primary interest in producing synthetic estimates. Partitioning areas into subareas as one way to help decrease the within variance is a facet of Steven Cohen's paper. The use of AID for determining the demographic categories for synthetic estimation is a methodological aspect of Promisel's paper.

The need for the producer to supply information about the quality of the synthetic estimates came up a number of times during the conference. Some possible ways of accomplishing this are described in Fay's and Gonzalez's papers.

Several types of applications of synthetic estimates in the work of the Census Bureau are described in the papers by Gonzalez and Fay. Applications in the drug and alcohol abuse fields are discussed in the papers by Reuben Cohen, Froland and Promisel. Reuben Cohen's paper illustrates use of a multiple regression model.

Publication of the papers and discussion should permit a wider audience of users to understand the characteristics, strengths, and limitations of the current types of Synthetic Estimators. Producers of subnational data will be able to review the current state of the art as viewed by the Workshop participants.

The desirability of additional research was identified at a number of points in the Workshop. What is known to date is represented by the contributions in these proceedings. It is reasonable to expect that this compilation will help stimulate additional productive ideas and results.

REFERENCES

Advertising Research Foundation U.S. Television Households, by Region, States, and County, New York, March 1956.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. Sample Survey Methods and Theory, Vol. I. New York: John Wiley and Sons, 1953.

National Center for Health Statistics Synthetic State Estimates of Disability, Public Health Service, PHS Publication No. 1759, Washington: U.S. Government Printing Office, 1968.

Simmons, W.R. Subnational Statistics and Federal-State Cooperative Systems, Committee on National Statistics, Assembly of Behavioral and Social Sciences, National Research Council. Washington: National Academy of Sciences, 1977.

Part I

Small Area Estimation-Synthetic and Other Procedures, 1968-1978

Paul S. Levy

ABSTRACT

Methods for obtaining small area estimates which have emerged over the past decade are reviewed with particular emphasis given to synthetic estimation, a procedure originally developed at the National Center for Health Statistics which has found wide acceptance because of its simplicity and intuitive appeal, and yet has provoked much controversy because of its lack of good demonstrable statistical properties and its equivocal results when subjected to empirical evaluation. The various methods of obtaining small area estimates are discussed in terms of their statistical properties, the feasibility of using them and the potential scope of their application. Finally, some recommendations are made concerning possible avenues of future research in small area estimation, and some tentative guidelines are given for choosing between alternative existing methods.

INTRODUCTION

It has now been ten years since the National Center for Health Statistics (NCHS) published estimates for each State in the United States of restricted activity days, bed disability days and other selected variables from the Health Interview Survey (HIS) and, in so doing, introduced in published form the concept of synthetic estimation [National Center for Health Statistics. 1968). At the time, this represented a radical departure from NCHS policy of publishing only estimates known to be for all practical purposes unbiased and for which sampling errors can be estimated. It was immediately recognized that the importance of this publication lay not in its HIS subject matter, but in its presentation at a period of time in which local, State, and regional planning were emerging as important issues, of an easily usable, inexpensive and intuitively appealing method of obtaining exactly the kind of small area estimates that were so sorely needed. At the same time, it was recognized that synthetic estimation is a crude method and that much further work was needed, especially in evaluation of this

method. Although the publication listed no individual authors, the project was initiated and carried out under the leadership of Walt R. Simmons, who should be considered the "father" of synthetic estimation if not its inventor.

Since the introduction of synthetic estimation ten years ago, there has been a moderate amount of activity in development of further methodology for small area estimation, especially at the U. S. Bureau of the Census and at the National Center for Health Statistics. Some of this activity was a direct outgrowth of the early NCHS work on synthetic estimation while other activity, particularly that of Ericksen (1975) had antecedents not in synthetic estimation but in demographic techniques of estimating population changes for small areas. Most of the activity in small area estimation, however, has centered around a relatively small group of statisticians (many of whom are at this conference) who represent either as staff members or as contractors the agencies responsible for producing such estimates. Although it is a potentially fertile field for research, it has not as yet attracted the interest of the statistical community at large.

In this paper, I will review the major work of the past decade in small area estimation and will comment on what I feel is needed in the way of future research.

2. METHODS OF PRODUCING ESTIMATES FOR SMALL AREAS

The various methods of producing estimates for small areas that have been given some attention over the past decade are discussed in order of decreasing dependency on actual direct measurement of individuals from the local area. The list is not intended to be exhaustive but represents the types of procedures that are currently being used. Undoubtedly, new procedures will emerge from the presentations at this conference.

2.1 Direct Estimation by Means of Sample Survey or Census

If one wants to estimate some parameter (e.g., mean, total, proportion) of the distribution of a variable, X, in a small area, the most direct method would be to take a sample survey or census of the individuals in the area and measure them with respect to the variable. X. If the sampline vlan were that of a vrobabilitv sample, if the survey were well planned and executed, and if a reasonable algorithm for estimation were used, unbiased estimates would be produced. The disadvantages of this approach are well known, namely the immense amount of resources needed in the way of time, money, and technical expertise for the successful completion of a sample survey that would produce estimates meeting reasonable specifications in the way of reliability or validity.

In spite of the expense involved, it should be recognized that estimates obtained from direct surveys of local areas have tremendous appeal to those individuals responsible for regional, State and local planning, and the consultant who proposes synthetic

estimation or some other method of estimation in lieu of a survey is apt to meet some resistance. In order to be effective, the consulting statistician must be able to evaluate the level of accuracy of estimates that can be produced from a sample survey conducted in accordance with the client's limitations in resources, to compare this with the level of accuracy that can be produced by synthetic estimation or some other method of indirect estimation, and to communicate these findings to the client. It is especially important to avoid amateurish, poorly planned and executed surveys, which can only result in inaccurate estimates.

2.2 Methods Using A Combination of Direct Estimation and Imputation

It will generally not be feasible for an independent survey to be conducted in a particular local area for purposes of obtaining local estimates. The only alternative then is to use data from other sources such as surveys that have been conducted in larger areas, and by some method to relate these estimates from other surveys to estimates for the small area of interest. In this section, we will discuss a method of producing small area estimates from larger area surveys which has the capacity of making extensive and direct use of whatever data is available from the survey specific to the small area.

This method, known as the nearly unbiased estimate, was discussed in the original NCHS publication on synthetic estimation (NCHS 1968). It is based on the fact that for many National Surveys such as the Current Population Survey (CPS) and the Health Interview Survey, the United States is grouped into a large number of primary sampling units and the PSU's are grouped into strata on the basis of similar geographic, economic or demographic characteristics. The PSU's are generally one or more counties or SMSA's and each stratum contains one or more PSU's. From each stratum, one PSU is sampled and estimates from the PSU's are inflated to stratum levels and aggregated to produce national estimates. From sample surveys having such designs, nearly unbiased estimates can be obtained for small areas by use of these stratum estimates. In particular, the nearly unbiased estimator, $\bar{x}'_a$, of the mean level of a variable, X, for a small area, a, is given by:

$$\bar{x}'_a = \left(\sum_{j=1}^{J} \left(\frac{n_{aj}}{n_{.j}}\right) x'_j\right) / n_{a.} \qquad (1)$$

where

x'_j = the survey unbiased estimate of the total or aggregate level of X in stratum j.

n_{aj} = the number of persons in stratum j that belong to area a.

$n_{.j}$ = the total number of persons in stratum j.

$n_{a.}$ = the total number of persons in area a.

and

J = the total number of strata in the survey.

To illustrate how this estimator is constructed, let us suppose that a population is grouped into three strata as illustrated below in Table 1:

TABLE 1

Number of Persons by Stratum and Estimated Total Level of X for Total Population and Number of Persons by Stratum for Area a

Stratum	Total Population	Estimated Total Level of X	Total Population in Area a
1	50,000	295	10,000
2	20,000	327	20,000
3	25,000	132	0
			30,000

The nearly unbiased estimate of the mean level of X in area a is given by:

$\bar{x}'_a$ = [(10,000/50,000)(295) + (20,000/20,000)(327) + (0/25,000)(132)]/ 30,000 = 386/30,000 = .0129

Conceptually, this method imputes the estimate for an entire stratum of the mean level of a characteristic to that part of the stratum that is in the small area of interest. The nearly unbiased estimate either uses local data directly or else imputes on the basis of data from similar small areas. For example, let us suppose that Stratum 1 consists of PSU's 1, 2, and 3 from which PSU 1 has been selected in the sample and that Stratum 2 consists of PSU's 4, 5, 6 and 7 of which PSU 6 is the sample representative. Let small area a consist of PSU's 1 and 6, small area b consist PSU's 1 and 5 and small area c consist of PSU's 3, 4 and 5. Then estimates for area a will be obtained completely from local data, estimates for area b partly from local data and partly by imputation and estimates for area c entirely by imputation.

The bias, $B(\bar{x}'_a)$ of the nearly unbiased estimator, $\bar{x}'_a$, is given by:

$$B(\bar{x}'_a) = \sum_{j=1}^{J} \frac{n_{aj}}{n_{a.}} (\bar{x}_j - \bar{x}_{aj}) \qquad (2)$$

where

$\bar{x}_j$ = the average level of characteristic, X, in stratum j.

and

$\bar{x}_{aj}$ = the average level of characteristic, X, in that part of stratum j that is in area a.

It follows from relation (2) that if there is little diversity within strata with respect to the characteristic being measured, the bias in the nearly unbiased estimate is likely to be small. An empirical study performed at the National Center for Health Statistics used HIS PSU's and stratification to construct nearly unbiased estimates for 42 States of 1960 deaths from all causes, major cardiovascular-renal diseases and deaths from motor vehicles (Levy and French, 1977). Since there was no sampling involved, differences between the nearly unbiased estimates and the true values are due entirely to bias, and the study showed for each of the three variables, the biases were, in general, quite small.

The problem in the nearly unbiased estimator is likely to lie not in its bias but in its variance, $\sigma^2_{\bar{x}'_a}$, given by:

$$\sigma^2_{\bar{x}'_a} = \sum_{j=1}^{J} \left(\frac{n_{aj}}{n_{a.}}\right)^2 \sigma^2_{\bar{x}'_j} \qquad (3)$$

where $\sigma^2_{\bar{x}'_j}$ is the variance of the survey estimate, $\bar{x}'_j$, of the mean level of X in stratum j. For most data systems, the $\sigma^2_{\bar{x}'_j}$ are likely to be quite large since the sample size in any one stratum is likely to be relatively small. In addition, the $\sigma^2_{\bar{x}'_j}$ might be difficult to estimate, from the data if the $\bar{x}'_j$ are based on complex sample designs.

The approach taken in constructing the nearly unbiased estimate for a small area is to use directly as much actual data from the small area as can be taken from the larger survey, and it is likely that such an approach would yield estimates having small bias but possibly large variance. This same approach was taken by Woodruff (1966) in attempting to obtain small area estimates of retail trade although his estimation procedure is quite different from that of the nearly unbiased estimator. Theoretical properties of the nearly unbiased estimator have been demonstrated by Levy and French (1977).

2.3 Methods Based on Regression Relationships

A third class of procedures used to obtain small area estimates assumes a relationship between a dependent variable, X, and a set of independent variables, $Z_1, \ldots, Z_k$. Estimates of X

for small areas are obtained not from direct measurement of X in the small area as in a sample survey nor from a combination of direct measurement of X in the small area and imputation based on direct measurement of X in an area similar to that of the small area as is done in constructing the nearly unbiased estimate, but on measurement of the independent variables $Z_1, \ldots, Z_k$ in the small area, and use of the relationship between X and $Z_1, \ldots, Z_k$. The motivation for use of this type of methodology is that if the set of independent variables, $\{Z_i\}$ are easily obtainable for the small area and if the relationship between X and the Z_i is strong, then estimates of good quality might be produced at relatively low cost. The major disadvantage of this type of approach is that the resulting estimates are likely to be biased since they are not based on direct measurement of the variable of interest in the small area of interest.

This class of methods includes synthetic estimation which has thus far dominated the field of small area estimation in addition to other methods that have recently emerged.

2.3.1 Synthetic Estimation

Let us suppose that estimates $\bar{X}'_1, \bar{X}'_2, \ldots, \bar{X}'_K$ are available from a survey conducted in a large area (e.g., nationwide) of the mean levels $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_K$ of a variable X in a set of K mutually exclusive and exhaustive classes (e.g., age, sex, race, family income, etc.). Let us suppose that estimates $Z'_{a1}, Z'_{a2}, \ldots, Z'_{aK}$ are available of the proportion of individuals in a small area, a, belonging to each of the K classes. Then the synthetic estimator, $\tilde{\bar{X}}_a$, of the mean level of X in area a, is defined by the relation:

$$\tilde{\bar{X}}_a = \sum_{k=1}^{K} \bar{X}'_k \, Z'_{ak} \tag{4}$$

We see from relation (4), that the synthetic estimator, $\tilde{\bar{X}}_a$ is a regression estimator in which the $\bar{X}'_k$ are the estimated regression coefficients and the Z'_{ak} are the independent variables obtained from the small area. In other words, a synthetic estimate is an estimate obtained from a multiple regression equation in which the independent variables are the small area population pro-

portions falling into mutually exclusive and exhaustive classes (obtained generally on the basis of demographic variables) and the estimated regression coefficients are estimates of the mean level of the dependent variables for the classes based on a survey or census conducted nationwide or at least in an area much larger than that for which estimates are desired.

There are several reasons why synthetic estimation is very appealing. First and foremost is its intuitive appeal. It seems likely that the mean level of many variables in a population is likely to be highly related to the distribution of the population by such demographic variables as age, sex, race, income, residence, etc., which are the independent variables generally used in obtaining synthetic estimates. In addition to its intuitive appeal, synthetic estimates are generally easy and inexpensive to obtain since the independent variables, Z'_{ak}, are easily available from census or other population data and the regression coefficients, $\bar{X}'_k$, are obtainable from National Surveys.

Some important instances in which synthetic estimates have been used over the past decade are listed in Table 2.

In addition to the six studies mentioned in Table 2 (plus others not mentioned) it should be noted that biostatisticians and epidemiologists have been using for many years a process very much akin to synthetic estimation in constructing rates and ratios by the indirect method of standardization. According to this method, class specific rates found in a "standard" population are combined in an equation similar to equation (4) with data from a population of interest relating to its proportionate distribution into these classes to obtain the expected rate that would be obtained in the population of interest on the basis of the standard population's class specific rates. The expected rate is then compared with the observed rate in the population of interest, and the ratio of the observed to expected rates is called a standard ratio.

Statistical properties of the synthetic estimator such as its variance, bias and mean square error have been developed in papers by Gonzalez and Waksberg (1978) and by Levy and French (1977) along with methods of estimating these parameters from the data. In particular, the variance, $\sigma^2_{\tilde{\bar{x}}_a}$ and bias, $B_{\tilde{\bar{x}}_a}$ of a synthetic estimator, $\tilde{\bar{X}}_a$, are given by:

$$\sigma^2_{\tilde{\bar{x}}_a} = \sum_{k=1}^{K} Z_{ak}^2 \, \sigma^2_{\tilde{x}'_k} + \sum_{k=1}^{K} \bar{X}_k^2 (1-Z_{ak}) Z_{ak}/n_a \qquad (5)$$

$$+ 2 \sum_{k<r} Z_{ak} \, Z_{ar} \, \mathrm{cov} \, (\bar{X}'_k, \, \tilde{X}'_r)$$

TABLE 2

Recent Studies Using Synthetic Estimation

Organization or Individual Investigators	Small Area	Variables Being Estimated (Dependent Variables)	Independent Variables	Regression Coefficients
1. NCHS, 1968	States	5 HIS variables relating to short and long term disability.	Population proportions falling into 78 classes on the basis of age, sex, race, residence, family income, family size, industry of head of family.	1963-1964 HIS estimates of mean level of dependent variables for each class based on national data.
2. U.S. Bureau of Census - Gonzalez and Hoza, 1978	Counties, SMSA's	Unemployment rates.	Population proportions falling into classes on the basis of occupation, sex, race, or on the basis of age-sex-race-marital status.	Current Population Survey (CPS) or census estimates of unemployment based on the geographic division in which the small area is located.
3. Namekata, Levy and O'Rourke, 1975	States	Complete and partial work loss disability.	Proportion of population falling into 60 age-race-sex-residence classes.	1970 census estimates of mean levels of complete and partial work loss disability for each of 60 classes for U.S., as a whole.

TABLE 2: Recent Studies Using Synthetic Estimation (Cont'd.)

Organization or Individual Investigators	Small Area	Variables Being Estimated (Dependent Variables)	Independent Variables	Regression Coefficients
4. NCHS, 1977	States	15 HIS variables relating to long and short term disability and to utilization of health services.	Proportion of population falling into 60 age-sex-race-family size-family income-industrv of household head class.	1969-1971 HIS estimates of mean level of dependent variables for each class based on national data.
5. Schaible , Brock and Schanck, 1977	Groups of Counties, States	Unemployment rates, percent of population having completed college.	Proportion of population falling into 64 age-sex-race-family size-industry of household head classes	HIS estimates of mean level of dependent variables for each class based on national data.
6. Levy, 1971	States	1960 U.S. deaths from four different causes.	Proportion of population falling into 40 age-sex-race classes.	1960 U.S. estimates of death rates for each class and for each cause.

and

$$(\tilde{\bar{X}}_a) = \sum_{k=1}^{K} Z_{ak} (\bar{X}_k - \bar{X}_{ak}) \qquad (6)$$

where

$\{Z_{ak}, k=1, \ldots, K\}$ are the true proportions of the population of area a falling into each class,

$\sigma^2_{\bar{x}'_k}$ = the variance of $\bar{X}'_k$, $k=1, \ldots, K$

n_a = the size of sample upon which the Z'_{ak} are based

and

$\bar{X}_{ak}$ = the mean level of X, in classes k of area a.

In most applications of synthetic estimation, both the estimated regression coefficients, $\bar{X}'_k$ and the estimated population proportions, Z'_{ak} are obtained from very large data systems and are likely to have very small sampling variances, so that one would anticipate that the sampling variances of synthetic estimates would be quite small. Estimates of the sampling variances of the 1969-1971 HIS synthetic estimates for States based on equation (5) seem to confirm this since the coefficients of variation of almost all the synthetic estimates were estimated to be less than 5% (NCHS, 1977).

Examination of equation (6) shows that the bias in a synthetic estimate is a weighted average of the difference between the expected value, $\bar{X}_k$, of the estimated regression coefficients and true regression coefficients, $\bar{X}_{ak}$ approprate for the particular class and area. In other words, the bias in a synthetic estimate depends on differences between the class specific mean levels, $\bar{X}_k$, for the large area used in obtaining the estimated-regression coefficients and the class specific mean levels, $\bar{X}_{ak}$, for the small area. Examination, a priori, of equation (6) cannot lead us to surmise, as we have done for the variance, that the bias of a synthetic estimate is likely to be small. It may in fact be large if the level of a variable X in an individual is less dependent on the individual's being in a particular class than on other factors and if the distribution of these other factors differs among areas. This might be seen in the following simplified linear model:

$$X_{ak\ell} = \mu + a_k + \sum_{j=1}^{J} \beta_j Y_{jak\ell} \qquad (7)$$

where

μ = an overall mean.

$X_{ak\ell}$ = the level of X for individual ℓ in class k of area a.

a_k = the effect due to being in class k.

$\{\beta_j, j = 1, \ldots, J\}$ = the effects due to a set of other variables, $y_1, \ldots, y_J$.

and

$Y_{jak\ell}$ = the level of variable y_j, for individual ℓ in class k of area a; $j = 1, \ldots, J$.

Under model (7), the mean level, $\bar{X}_{ak}$, for class k area a would be given by:

$$\bar{X}_{ak} = \mu + a_k + \sum_{j=1}^{J} \beta_j \bar{Y}_{jak} \qquad (8)$$

If the class mean levels, $\bar{Y}_{jak}$ of the variables, y_j do not differ appreciably among the areas, then the $\bar{X}_{ak}$ will be approximately the same among areas, which would imply that the bias in the synthetic estimate is likely to be small, even if the β_j are large. On the other hand, differences among areas with respect to those $\bar{Y}_{jak}$ which are associated with sizeable β_j would indicate the possibility of a large bias in a synthetic estimate.

Evaluation of synthetic estimates has been difficult in situations where the true value of the characteristic being estimated is not known. The difficulty lies primarily in the fact that the bias of the synthetic estimator cannot be estimated from the data used to construct it. Gonzalez and Waksberg (1973) have used a method of evaluation of a set of synthetic estimates based on the fact that if an unbiased estimate, $\bar{X}'_a$, exists of the mean level, $\bar{X}_a$, of variable X in area a, and if $\bar{X}'_a$ is uncorrelated with the synthetic estimator, $\tilde{\bar{X}}_a$, then an unbiased estimator, $\hat{MSE}_{\tilde{\bar{x}}_a}$ of the mean square error of $\tilde{\bar{X}}_a$ is given by:

$$\hat{MSE}_{\tilde{\bar{x}}_a} = (\bar{X}'_a - \tilde{\bar{X}}_a)^2 - \hat{\sigma}^2_{\bar{x}'_a} \qquad (8)$$

$\hat{\sigma}^2_{\bar{x}'_a}$ is an unbiased estimate of the variance of $\bar{X}'_a$

Since the $\tilde{X}'_a$ are likely to have high variances (or else they would be competitive with synthetic estimates) it is likely that the estimated mean square errors given in equation (9) are unstable. Realizing this, Gonzalez and Waksberg concentrated on estimating the average mean square error (denoted AMSE) of a set of M synthetic estimates by the more stable estimator:

$$\hat{AMSE} = \frac{1}{M}\sum_{a=1}^{M}(\bar{X}'_a - \tilde{X}_a)^2 - \frac{1}{M}\sum_{a=1}^{M}\hat{\sigma}^2_{\bar{x}'_a} \qquad (10)$$

Using this criterion, Gonzalez and Waksberg (1973) evaluated synthetic estimates of unemployment for SMSA's against competing unbiased estimates, and found that synthetic estimates were superior to unbiased estimates for monthly rates, but that the reverse was true for annual unemployment rates.

Some studies have been designed to evaluate synthetic estimates by comparing them with known true values of the parameter being estimated. Such studies have been performed for such variables as death rates from selected causes (Levy 1971), complete and partial work disability (Namekata, Levy, and O'Rourke 1975)) unemployment rates and percent completing college (Schaible, Brock, and Schnack 1977). The overall conclusion emerging from these empirical evaluation studies concerning the accuracy of synthetic estimates is at best equivocal. For some variables, synthetic estimates were quite accurate, whereas for others they were not good at all.

Two interesting findings have emerged from these and other evaluation studies. It has been found in most instances that there is not much variability in the Z'_{ak} among small areas, and that as a result, there is generally not much variability among small areas, with respect to actual values of synthetic estimates. For this reason there is often low correlation, over a set of small areas, between synthetic estimates and true values of the parameter being estimated, and this is a serious deficiency if the synthetic estimates are being used to order a set of small areas on the basis of the variable being estimated. A second finding is that the large number of classes used to construct synthetic estimates is probably not needed since the values of synthetic estimates based on relatively small numbers of classes correlate very well with values of synthetic estimates based on a much larger number of cells.

2.3.2 Other Methods Based on Regression Relationships

Perhaps the most successful use of small area estimation has been in the estimation of population changes for small areas.

In particular, Ericksen (1974 and 1975) has built a regression equation using as independent variables data on births, deaths and school enrollment for CPS PSU's and as the dependent variable, data on population size for these PSU's as estimated from CPS. This regression equation was then used to estimate population changes from 1960 to 1970 for 2,586 counties, and the agreement between the predicted values and the actual census values was, in general, quite good. Perhaps the main reason that a regression method worked so well in this application lies in the fact that the independent variables births, deaths, and school enrollment are known to be very highly correlated with population change.

Two methods have been developed in which synthetic estimates are constructed, and then used essentially as independent variables in a regression equation which includes other variables characterizing the small area of interest. One such method, proposed by Levy (1971) assumes the following model:

$$Y_a = \beta_0 + \beta_1 W_{a1} + \ldots + \beta_h W_{ah} \qquad (11)$$

where

$Y_a = 100\ (\bar{X}_a - \tilde{\bar{X}}_a)/\tilde{\bar{X}}_a$

β_i ; i = 0 , . . . , h are a set of regression coefficients.

and

W_{ai} , i = 1, . . . , h are values for area a of a set of independent variables.

In other words, Y_a, the percentage difference between a synthetic estimate, $\tilde{\bar{X}}_a$ and the true mean level, $\bar{X}_a$, of a variable X in a small area a is assumed to be a linear function of a set of independent variables, $W_{a1}, \ldots, W_{ah}$. If enough larger areas are available for which $\bar{X}_a$, $\tilde{\bar{X}}_a$ and the set of W's are known, then the regression coefficients, β_i, can be estimated and by use of these estimated regression coefficients $\hat{\beta}_i$, an estimator, $\hat{\tilde{\bar{X}}}_a$ can be derived from equation (11) and can be used for small area estimation as an "improved" synthetic estimator. This estimator is given by:

$$\hat{\tilde{\bar{X}}}_a = \tilde{\bar{X}}_a(1 + .01(\hat{\beta}_0 + \hat{\beta}_1 W_{a1} + \ldots + \hat{\beta}_h W_{ah})) \qquad (12)$$

This estimator, when evaluated on mortality data, showed a considerable improvement over the synthetic estimator (Levy 1971).

A similar approach was taken by Gonzalez and Hoza (1978) who used synthetic estimates of unemployment as an independent variable

along with other independent variables and built a regression equation to produce small area estimates of unemployment.

The approach taken by these two regression procedures is based on the realization that some kind of regression estimator is likely to be an improvement over a direct estimate for a small area even when such an estimate is obtainable, and that the synthetic estimate, while useful, does not tell enough of the story to accurately estimate a population parameter.

2.4 Methods Based on a Combination of Regression Methods and Direct Estimation

Very recently, Schaible, Brock and Schnack (1977) have proposed an estimator based on a linear combination of a direct unbiased estimator and a synthetic estimator. The rationale for their estimator is that often the same data upon which the regression coefficents, $\bar{X}'_k$, are obtained for the synthetic estimator, contain sample units from the small areas for which estimates are desired, and that often these sample data can be used by themselves to obtain direct estimates for the local data. In particular, they speculated, that the mean square error, denoted b', of synthetic estimate, $\bar{X}'_a$, is relatively independent of n_a, the number of units sampled in area a, whereas the mean square error of a direct estimate, $\bar{X}'_a$, is dominated by its variance rather than its bias and is of the form, b/n_a. Then the linear combination of $\bar{X}'_a$ and $\tilde{\bar{X}}_a$ which has the minimum variance over all such linear combinations is given by:

$$C\bar{X}'_a + (1 - C)\,\tilde{\bar{X}}_a \tag{13}$$

where

$$C = n_a/(n_a + (b/b')) \tag{14}$$

If $\tilde{\bar{X}}_a$ and $\bar{X}'_a$ had equal mean square errors, then C = ½ and:

$$(b/b') = n_a \tag{15}$$

Thus, from relation (15) b/b' is equal to the sample size, n_a, at which synthetic and direct estimates have equal error. From available data, Schaible, Brock and Schnack were able to estimate b/b' and hence C for two HIS variables, and demonstrated that their composite estimator had considerably lower average MSE than either $\bar{X}'_a$ or $\tilde{\bar{X}}_a$ used alone.

3. WHERE SMALL AREA ESTIMATION STANDS NOW AND WHERE IT SHOULD GO

When demographics tell most of the story concerning the expected level of a characteristic, the synthetic estimator is likely to be the estimator of choice. However, the empirical studies of the synthetic estimator have accumulated sufficient evidence to indicate that for most variables of interest, demographics do not tell most of the story. As a consequence, there is a general feeling of dissatisfaction with synthetic estimation. However, there seems to be no clarion call for allocating the huge amount of resources needed to obtain good small area estimates by direct estimation.

It seems that the most productive approach would be to develop an estimator based on demographics, on whatever direct information is available for the small area with respect to the dependent variable being estimated, and on independent variables other than demographics. The statistical properties of any such estimation procedure should be established, and by that I mean not only variance and bias, but such characteristics as optimality, cost efficiency and admissibility. To investigate these properties and gain some insight, it might be necessary to go beyond conventional finite population sampling and estimation theory.

Good local planning requires good local estimates. At present, we cannot deliver these for most variables. However, if we make this a high priority item for statistical research and build upon what has been developed over the past decade, it is likely that much progress will be made in the next decade.

REFERENCES

Ericksen, E.P. A regression method for estimating population changes of local areas. Journal of the American Statistical Association, 69: 867 - 875, 1974.

Ericksen, E.P. A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. Demography, 10: 137 - 160, 1975.

Gonzalez, M.E., and Waksberg, J.E. Estimation of the error of synthetic estimates. Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria, 1973.

Gonzalez, M.E., and Hoza, C. Small area estimation with applications to unemployment and housing estimates. Journal of the American Statistical Association, 73: 1978.

Levy, P.S. The use of mortality data in evaluating synthetic estimates. Proceedings of the American Statistical Association, Social Statistics Section: 328, 1971.

Levy, P.S., and French, D.K. Synthetic Estimation of State Health Characteristics Based on the Health Interview Survey. Vital and Health Statistics: Series 2, No. 75, DHEW Publicaion (PHS) 78 - 1349. Washington: U.S. Government Printing Office, 1977.

Namekata, T.; Levy, P.S.; and O'Rourke, T.W. Synthetic estimates of work loss disability for each State and the District of Columbia. Public Health Reports, 90: 532 - 538, 1975.

National Center for Health Statistics. Synthetic State Estimates of Disability. PHS Publication No. 1759. Public Health Service, Washington: U.S. Government Printing Office, 1968.

National Center for Health Statistics. State Estimates of Disability and Utilization of Medical Services, United States, 1969 - 1971. DHEW Publication No. (HRA) 77 - 1241. Health Resources Administration. Washington: U.S. Government Printing Office, Jan., 1977.

Schaible, W.L.; Brock, D.B.; and Schnack, G.A. An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics. Proceedings of the American Statistical Association, Social Statistics Section: 1017 - 1021, 1977.

Woodruff, R.A. Use of **a** regression technique to produce area breakdowns of the Monthly National Estimates of Retail Trade. Journal of the American Statistical Association, 61: 497 - 504, 1966.

Discussion

Walt R. Simmons

INTRODUCTION

Let me say first that Paul Levy's paper is an excellent introduction to our workshop on synthetic estimates, and an opening review of efforts to produce useful estimates for subnational areas.

I should like to offer my general perspective of these issues. You will discover that Paul already has touched on several facets that I consider particularly important, while a scanning of the agenda suggests that other elements of my position will be treated by other speakers.

A CENTRAL CONCEPT

I start with a central concept, or model, or proposition. Let us say that the primary objective is to estimate a parameter z for a defined universe. Consider a very general estimator

$$\bar{z}' = \sum_a w_a \bar{x}'_a$$

in which $\bar{x}_a$ is an estimator for the a-th component of the $\bar{z}$-value, and w_a is a weight applied to $\bar{x}'_a$-value -- all terms to be defined later -- so that the estimator $\bar{z}'$ is a linear combination of the weighted $\bar{x}'_a$ estimates.

This estimator encompasses a very wide range of possible processes; its descriptive characteristics depend upon the definitions given to the x, w, and a-values.

A. One class of definitions makes $\bar{z}'$ the basic estimator of stratified probability sampling, for either a simple or more complex design involving differential sampling rates, multi-stage procedure, ratio controls or other elaborations.

B. Slightly different specifications make $\bar{z}'$ a post-stratification estimator.

C. With another orientation, $\bar{z}'$ is the result of a standard multivariate regression analysis.

D. The estimator can also be considered a formal statement of an a-standardized estimate, although for this model one needs also a particular definition of the target parameter $\bar{z}$.

E. And $\bar{z}'$ can represent a Synthetic Estimate of the type Paul Levy has discussed, or allied types, some of which are composites of two or more primary estimates.

Our task is to select a specific model and associated definitions and procedure that in some sense will produce a "best" estimate of the target parameter. This best estimate will be evaluated most likely in terms of particular objective, variance, bias, cost and feasibility. It helps me to think about the problem within the framework of the general linear equation I first mentioned.

WHY NOT USE ALWAYS AN UNBIASED PROBABILITY PROCEDURE?

Think of the common problem of securing estimates for subnational geographic areas, as Levy does. These areas nay be states, counties, metro areas, or the 38,000 different political jurisdictions designated for federal revenue-sharing. The topic may be unemployment, disability, crop production, price level, or something else. If good administrative data collected on a 100-percent basis for some operational purpose exist, they should be used.

If universe figures do not exist, the need for small area data is sufficiently great, the number of areas not too large, the cost not too high, and technical resources adequate, direct measurement by probability surveys is in order.

Too often these conditions are not met. Then we must adopt some model and one of the other strategies mentioned. We need not be entirely apologetic about such action. Quite aside from cost and feasibility, direct measurement of each of many small areas does not always yield the best possible set of estimates. The measure-nent process itself nay be biased for some and not for other areas. More commonly, the measurement process -- especially if it involves interviewers or other local agents -- is very likely to be subject to considerable between-agent variance, and lead to questionable between-area comparisons. Good estimates of variance and bias for such local estimates are difficult to secure. I wish Paul had put some emphasis on the weaknesses of criterion measures.

On the positive side, I would note that analysts have not hesitated to adopt model approaches to solution of a great variety of problems. Whether we utilize a simple model such as "distance equals rate times time," or a more complex model such as the actuary's "life expectancy," in the great majority of analysis some

hypothetical approximation to real world transactions is adopted. Indeed much can be said for acceptance of the "convention" of the product of a defined measurement process as an official value, instead of the unobtainable "true" value.

Sometimes we don't really need a correct measure of level of a statistic specific to each small area. All that is needed is a set of relative indicators -- perhaps rank order, or knowledge that Area A is a member of one class and Area B a member of another class. I was impressed by a remark I heard recently that this principle should be adopted by declaring that the count of the population obtained by the Census Bureau in the decennial census is the basis of congressional apportionment.

PLAUSIBILITY

The desire for estimates specific to small areas is often, perhaps usually, based on the notion that geography is some amalgamated proxy for other factors. Much of the reason for interest in the unemployment rate in Detroit is not because of its latitude and longitude, but a consequence of the industry and occupational distribution of the people who live there. Similarly for health phenomena: we believe that most health characteristics are functions of age, sex, marital status, education, income, occupation . . . as Paul Levy says, the principal attribute of the synthetic estimate is its intuitive appeal, its plausibility.

TWO WEAKNESSES OF SYNTHETIC ESTIMATES

First is the fact that the synthetic estimate takes account of only some of the causal or even correlated components of a dependent variable. This is indeed a fact and a weakness. How to minimize its impact is one of our central tasks -- albeit a task not unique to the synthetic technique.

Second is that we cannot estimate the precision of the synthetic estimate. Gonzalez and Waksberg (1973), among others, have tackled this problem. They have developed a scheme, applicable in some situations, in which an average variance and average mean square error can be calculated for the small area estimate. This approach has been criticized on the ground that it yields only an average value which is not specific to any particular area. I agree that this is an imperfect situation. Yet it is not as radically different from more conventional survey practice as it may appear. In the usual operational probability survey, we almost never know the true variance of the estimate. What we have is an estimate of variance, which is itself subject to variance, and is a "good" measure of the precision of a specific primary statistic only in an average sense. Most estimates of variance omit certain components of measurement error, and only rarely is one able to incorporate a decent measure of bias in estimating mean square error.

INDEPENDENT VARIABLES AND REGRESSION COEFFICIENT

I would appreciate a little amplification from Paul of the principles behind his view that a synthetic estimator is a regression estimator in which the independent variables are the population proportions, and the regression coefficients are mean values of the statistics for the various population classes. I have no quarrel with this view, and have myself spoken of the close relationships between regression and synthetic estimates. But I suspect some observers would have expected the "independent variables" and the "coefficients" to have been interchanged.

AN EXPLANATORY NOTE

Reasons for the initial choice of the label "synthetic" at the National Center for Health Statistics may be of interest. These reasons were a merger of two distinct avenues of thinking. One was a recognition that there is widespread use of the term "analysis" in drawing conclusions from a body of data, whereas our objective was to "synthesize" the evidence from more than one source. The other was an effort to distinguish this contrived estimate, which lacked some of the desirable attributes of an unbiased probability estimate, from results of the classical probability survey. Despite some criticisms, the term seems to have caught on, and I continue to like it.

CLOSING REMARK

Let me close with the same remark with which I ended a paper given at the International Statistical Institute a few years ago, paraphrasing Alexander Pope:

When first one casts his eye upon the synthetic estimate, he shrinks away in horror; with a second and then a third look, the aversion begins to fade, until finally one clasps the estimator to his bosom, and embraces it with affection. As a probability sampler, and an experimenter with the technique, this statement tends to reflect my current position. The synthetic estimator is a dangerous tool, but with careful further development, it has an attractive potential.

REFERENCE

Gonzalez, M. E., and Waksberg, J. Estimation of the error of synthetic estimates. Unpublished paper, presented at meeting of International Association of Survey Statisticians, Vienna, Austria, 1973.

Discussion

Gary G. Koch

This paper by levy represents an excellent discussion of the current status of statistical methodology for the estimation of various parameters for local areas (like states or counties) of a national population. For this purpose, four basic types of strategies are identified. These are as follows:

1. Direct estimators
2. Covering (or nearly unbiased estimators)
3. Prediction (or regression) estimators
4. Composite estimators involving various types of combinations of (1), (2), (3)

Each of these procedures has certain advantages and certain disadvantages whose relevance to their practical usefulness (or sensibility with respect to validity and reliability) inherently depends on the specific nature of the situation where they are to be used as reflected by cost considerations, on the one hand, and the plausibility of their underlying technical assumptions. on the other. These issues are clearly presented here by Levy in a manner which indicates the extensive work by both statisticians and other interested persons concerning the theoretical statistical properties and empirical performance of different types of local estimation methods. From this discussion, the following general conclusions seem to emerge:

1. Direct estimators are the most desirable in principle because they are based solely on data from the corresponding local areas for which they are produced. However, for many existing sample survey designs, their computation may not be straightforward. In addition, they may also fail to satisfy variance specifications. Cost considerations also represent a major limitation for the feasibility of designs for which local estimation is a primary' objective.
2. Covering estimators are intuitively appealing since they are based on the relatively reasonable assumption that small areas are approximately similar to larger

areas which contain them. For this reason, they are nearly unbiased. On the other hand, their variance may be rather large and thereby restrict the scope of their applicability.

3. Prediction estimators are the most well-known method for small area estimation because of their computational convenience; yet they are the most controversial because their validity inherently depends on rather strong assumptions whose appropriateness for any specific situation is difficult to evaluate. In this regard, the basic assumption is that the variation of the parameter of interest among local areas (or certain sub-units which comprise them) can be entirely characterized by a statistical prediction model which involves an available set of independent (or symptomatic) variables. In the simplest cases, such models are based on weighted (with respect to local area composition) linear combinations of domain means. More complex extensions include a regional level ratio adjustment and/or a regression adjustment for potential bias. Other related methods are based directly on multiple regression models. In all of these cases, the critical issue is whether or not the prediction model does indeed include all of the independent variables which may be related to the variation of the parameter of interest and that its specific structure is formally correct with respect to their separate and simultaneous roles. For those cases where this type of assumption is reasonable, prediction estimators are probably useful. Otherwise, their potentially large bias may cause them to be misleading.

4. Composite estimators are of interest because they permit trade-offs among the advantages and disadvantages of the estimators (1), (2), and (3) through their weighted combination. Thus, each type of estimator is emphasized (by receiving the greatest weight) for those local areas for which it performs the best in the sense of the smallest mean square error.

Given this summary of the current methods for local area estimation, Levy concludes his discussion with the recommendation that total survey design concepts be a principal focus of future research. In other words, attention should be given to the formulation of a unified framework for evaluating alternative estimation strategies in terms of their overall cost in a manner which takes into account the combined use of:

a. direct information pertaining to the parameters of interest for the respective local areas through modification of the sample survey design

b. indirect information on both readily available demographic variables and other potentially important independent variables for which special purpose data collection or data management efforts may be required

c. straightforward vs. complex computational algorithms for both the local area estimates themselves and corresponding estimates of their standard errors

Thus the most appropriate method of local area estimation for a specific situation could be based on either cost efficiency considerations, given the satisfaction of quality control specifications with respect to bias and variance, or accuracy considerations, given cost constraints. Since this type of approach permits the statistical issues concerning alternative procedures to be resolved in terms of sample survey design, data management, and data analysis considerations simultanebusly, it should indeed be a high priority item for future statistical research.

All of the previous remarks were specifically concerned with the material presented by Levy. In the remainder of this discussion, attention will be focused on certain philosophical and methodological principles which pertain to the field of statistics in general and their relevance to the topic of local area estimation. First of all, it is necessary to recognize that the problem of local area estimation is really not different from any other statistical estimation problem. To be specific, a sample is selected from a particular population and estimates for some parameters of interest are sought for a particular partition of it into subpopulations (or domains). In addition, for certain independent variables which are potentially related to the parameter of interest, data are available either for the individual elements which comprise the population and/or certain clusters of such elements. Thus, such information can be used to obtain improved estimates (in the sense of variance reduction) for the respective subpopulations via regression methods, provided such adjustments are considered to be philosophically acceptable from the points of view of both the statistician who is responsible for producing subpopulation estimates and the investigator or policymaker who intends to use them. Here the fundamental issue is whether or not the subpopulations under consideration are individually unique and thereby require estimates based solely on their own separate data. If this is the case, then only direct estimates are appropriate, and the sample survey should be designed accordingly. For extensive surveys like the Health Interview Survey, which involve approximately 40,000 households per year, Schaible, Brock, and Schnack (1977) have observed that direct estimators are already potentially feasible for larger (with respect to population) areas like California. Thus, if they were required for all States, sample survey design modifications or supplements would seem to be needed only for the smaller States, which should not necessarily be prohibitively costly.

Alternatively, if the sub-populations under consideration are entirely homogeneous within the respective cells of the independent variable cross-classification (i.e., across the subpopulation dimension of the independent variable x subpopulation two-way partition), then prediction estimators are both reasonable and practical. For example, Levy (1971) found that synthetic State estimates based on age x sex x race cells for cardiovascular renal disease death rates in 1960 were in good agreement with the corresponding true death rates, but that those for motor vehicle accidents were in poor agreement with their

counterparts. Although this finding seems equivocal, it actually is expected because age, race, and sex are considered to be relatively important risk variables for cardiovascular renal death but relatively unimportant risk variables with respect to motor vehicle accident death. Similarly, Levy and French (1977) report that synthetic estimates based on age alone for disability and medical service utilization parameters given in NCHS (1977) agreed as well with the estimates from a more extensive seven variable cross-classification as those based on age x sex, age x sex x race, and age x sex x income. This finding is also more or less expected because age tends to be the most important of these variables with respect to the risk of disability and the potential use of medical services. With these comments in mind, it becomes apparent that the appropriateness of prediction estimators inherently depends upon the extent to which the corresponding independent variable cross-classification contains all variables which are related to the parameter of interest. For this purpose, the current literature on local area estimation gives no specific guidelines. However, the basic question which is involved is essentially the same as that which is addressed in the development of statistical prediction models for observational and experimental data. Thus, given that all potentially relevant independent variables are available, screening methods like that described in Higgins and Koch (1977) can be used to identify those which have statistically important relationships with the response (or dependent) variable which is under consideration. The cross-classification of these variables then represents the basic information for prediction purposes. However, if the number of cells which are involved here is very large, some of the corresponding estimates may not be reliable. Currently, this source of difficulty is handled bv combining various cells together (collapsing). Alternatively, linear or log-linear regression models could be fitted to the full cross-classification in order to identify whether or not it could be characterized in terms of certain main effects and lower order interactions. Fitted (or smoothed) estimates based on such models would then be obtained for the complete cross-classification and then used to obtain prediction estimators for local areas. Moreover, as long as this cross-classification requires only lower order interactions as opposed to higher order ones, the reliability of the respective fitted (or smoothed) estimates should be satisfactory (since their statistical properties are typically linked to the statistical properties of the set of lower order cross-classifications that correspond to the network of interactions which are included in the regression model). Thus, with these considerations in mind, it should be possible to make the independent variable framework for prediction (or synthetic) estimators more valid and efficient. However, the use of larger cross-classifications may not be consistent with the information concerning the overall cell distributions which is available at the local level. For this purpose, log-linear model and raking methods for contingency tables as described by Bishop, Fienberg, and Holland (1975) and Freeman and Koch (1976) become of interest provided that information is available for partially overlapping iower order cross-classifications that include all indeuendent variables. and higher order interactions which are outside these can be assumed negligible.

In summary, the use of prediction estimators can be put on a stronger statistical basis if the required supplementary data collection and computational efforts are considered worthwhile from the point of view of an overall cost model like that described previously. Otherwise, either direct or some other alternative strategy should be considered. In this regard, the method proposed by Kalsbeek (1973) and further discussed by Cohen and Kalsbeek (1977) and Cohen (1978) is of potential interest. It involves the partition of the overall population and hence all local areas into subunits which are then clustered together on the basis of their similarity with respect to an appropriate set of independent variables and/or the response variable. Local area estimates are then formed by combining national estimates for these clusters together in accordance with the corresponding internal distribution of subunits among them. Thus, such estimates involve both covering and prediction concepts. Their principal advantage is that they do not specifically require the use of a formal regression model. However, their use is not straightforward because it inherently depends on the development of algorithms for forming the clusters.

As stated previously, local area estimation does not really involve statistical problems which are unique to it. The basic issue is to produce the most reasonable estimates which are possible within a specified set of ground rules. Unfortunately, the nature of these ground rules tends to put certain limitations on the quality of these estimates. Thus, the most straightforward approach to obtain better estimates is to adopt a new set of ground rules. This discussion has attempted to suggest some types of considerations which may be of potential future interest for this purpose.

REFERENCES

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

Cohen, S.B. A modified approach to small area estimation. Unpublished doctoral dissertation. Chapel Hill: University of North Carolina, School of Public Health, 1978.

Cohen, S.B. and Kalsbeek, W.D. An alternative strategy for estimating the parameters of local areas. Proceedings of the ASA, Social Statistics Section, 781-786, 1977.

Freeman, D.H. and Koch, G.G. An asymptotic covariance structure for testing hypotheses on raked contingency tables from complex sample surveys. Proceedings of the ASA, Social Statistics Section, 330-334, 1976.

Higgins, D.H. and Koch, G.G. Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey. International Statistical Review, 45:51-62, 1977,

Kalsbeek, W.D. A method for obtaining local postcensal estimates for several types of variables. Unpublished doctoral dissertation, Ann Arbor: University of Michigan, 1973.

Levy, P.S. The use of mortality data in evaluating synthetic estimates. Proceedings of the ASA, Social Statistics Section, 328-331, 1971.

Levy, P.S. and French, D.K. Synthetic estimation of state health characteristics based on the health interview survey. Vital and Health Statistics: Series 2, No. 75, DHEW Publication (PHS) 78-1349. Washington, D.C.: U.S. Government Printing Office, 1977.

National Center for Health Statistics. State Estimates of Disability and Utilization of Medical Services, United States, 1969-71. DHEW Publlcation No. (HRA) 77 1241. Health Resources Administration. Washington, D.C.: U.S. Government Printing Office, 1977.

Schaible, W.L.; Brock, D.B.; and Schnack, G.A. An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics. Proceedings of the ASA, Social Statistics Section, 1017-1021, 1977.

Comments

Paul S. Levy

I would like to thank both Walt Simmons and Gary Koch for their very well prepared remarks and would like to address some of the issues raised by them.

First, Walt mentioned the important issue of interview bias. The biases discussed in my paper are basically sampling biases. As an illustration, the nearly unbiased estimator as applied to the Health Interview Survey is a linear combination of stratum estimates, and some of these stratum estimates might be based on data obtained from a single interviewer. Thus, it is very likely that the nearly unbiased estimator might be very sensitive to measurement error arising from the eccentricities of the interviewers.

Walt's second point is about the use of synthetic estimates or other methods to order a set of local areas with respect to the level of some variable. I would like to reemphasize what was mentioned in my presentation about this issue, namely that the synthetic estimator may not be good for this purpose since it is often based on demographics which show little variation from area to area. For example, the age, race, sex distribution of New York might not differ that much from Philadelphia, and hence synthetic estimates for the two would be very much alike. Typically one obtains synthetic estimates for a set of local areas which show little diversity and do not correlate well with the corresponding set of direct estimates.

Walt's final point concerns the formulation of the synthetic estimate as a regression estimate. In my formulation, the $\bar{X}'_1, \ldots, \bar{X}'_k$ are the class specific estimates from a large survey and serve as the regression coefficients that are used for every area, whereas the Z's are the "measurements" from the local area. In other words, the set, $\{\bar{X}'_1, \ldots, \bar{X}'_k\}$ are the "betas" in classical regression terminology. The first time I heard a synthetic estimate called a regression estimate was in a 1973 ASA invited paper session on local area estimation, and I believe that Eli Marks raised the point from the audience that the synthetic estimate is just another regression estimate.

I like Gary's term "covering estimate" instead of "nearly unbiased; and I think that we should proclaim him the "father of covering estimates."

His other point is very well taken concerning the use of modem multivariate methods, such as those developed by Gary and the North Carolina group as well as loglinear methods developed by Bishop, Fienberg, and Holland. These methods have considerable potential in exploring relationships in data obtained from complex surveys. These methods are most useful on the unweighted survey data, again as exploratory devices, and are now available in many of the standard statistical software packages.

General Discussion

* In the prior discussion both enthusiasm and skepticism were expressed concerning synthetic estimates. We should wait to see how we feel at the conclusion of the Workshop.

* One of the questions which is worth addressing is: are there biases introduced when Z'_{ak} is out of date? What are the orders of magnitude?

Is there a theoretical formulation for showing the effect of the biases of Z'_{ak} similar to the formulation showing the biases of the $\bar{x}'_k$? Most people seem to assume that the Z'_{ak} are current data, when in fact the data may be six or eight or more years old and there may have been material changes in demographic composition. Unlike direct survey estimates where both components are handled as a current estimation procedure, in the synthetic estimates there are also errors and other problems in the Z'_{ak}.

* One possibility, of course, is to create the Z'_{ak} as a set of synthetic estimates. To some extent Ericksen's work in making population estimates through synthetic procedures approaches this. Thus, this may result in having synthetic estimates of the second order.

* The need for local area statistics, of course, was the reason for the change in the census legislation to have a quinquennial census. Thus, demographic components of the synthetic estimate may have a smaller bias in the future than at present.

* Indicates a change of speaker.

* It may be useful to note that Paul Levy's paper started with 1968. However, before that time there were a number of practical applications of what since 1968 has been called synthetic estimates.

As mentioned in the introduction, the "FCC Radio Survey" and the Television Set County by County Distribution Estimates are two illustrations. A third illustration is the use of a sample survey of the insured population of the social security system in Chile in combination with census projected estimates by small areas proposed by Steinberg (1965). There are a number of other applications of synthetic estimates. The Consumer Price Index is calculated to provide not only national estimates but also estimates for a number of local areas. A paper by Marks (1978) describes how, as part of the current revision, the weights for the local areas are determined as a composite synthetic estimate, Further illustrations are to be found in the papers to be presented at this Workshop by Gonzalez and Fay. A series of papers by Ghangurda and Singh (1976, 1977a, 1977b), of Statistics Canada, have dealt with the methodological development of synthetic estimates and empirical evaluation in reference to the Canadian Labour Force Survey. The questions of bias and efficiency of synthetic estimates in household surveys are a major focus of this ongoing research. (Contributing to the general discussion during this period were: Eugene Ericksen, Robert Fay, Maria Gonzalez, Monroe Sirken, Joseph Steinberg, and Joseph Waksberg.)

REFERENCES

Ghangurda, P.D. and Singh, M.P. Synthetic Estimation in the LFS, Technical Memorandum, Household Surveys Development Staff, Ottawa: Statistics Canada, December 1976.

Ghangurda, P.D. and Singh, M.P. Evaluation of Synthetic Estimation in the LFS, Technical Report, Household Surveys Development Staff, Ottawa: Statistics Canada, July 1977.

Ghangurda, P.D. and Singh, M.P. Synthetic Estimation in Periodic Household Surveys, Survey Methodology, Vol. 3, No. 2, Ottawa: Statistics Canada, December 1977, pp. 152-181.

Marks, H.M. Estimation of Cost-Weights for the Consumer Price Index, Proceedings of the Survey Research Methods Section, Washington: American Statistical Association, 1978.

Steinberg, J. Recommendations for Sample and Survey Design, General Population Survey, . . . , Servicio de Seguro Social, CHILE, Department of Technical Cooperation. Washington: Organization of American States, 1965.

Part II

A Composite Estimator for Small Area Statistics

Wesley L. Schaible

I. ABSTRACT

Samples designed to provide estimates for large geographic areas are sometimes used to provide estimates for small areas. In such cases the sample in a small area may be "unrepresentative" or of small size. Various estimators, including a composite estimator, which is a weighted function of two component estimators, have been suggested for use in these situations. The choice of weights for the composite estimator is considered in this paper. It is shown that with appropriate weights the composite estimator has smaller mean square error than either component estimator and also that this estimator is remarkably robust against poor choices of weights. Data from the National Center for Health Statistics' Health Interview Survey and the Bureau of the Census' Public Use Tapes are used to illustrate results when direct and synthetic estimators are used as components of the composite estimator.

II. INTRODUCTION

Large samples such as those of the Current Population Survey (CPS) and Health Interview Survey (HIS) were designed to provide national and regional estimates. Although such statistics are useful, there is considerable demand for estimates for smaller geographic areas, for example, States and counties. One way to meet this demand is to redesign or supplement existing surveys, but this can be both expensive and time consuming. An alternative approach, which in some cases may be only an interim solution, is to produce biased estimates using existing data sources. Considerable attention has been devoted to the problem of producing estimates for small areas from existing sample surveys that were designed to produce national and regional estimates.

In 1968 in the publication Synthetic State Estimates of Disability (NCHS) the authors state that the sample size [and design) of the HIS was inadequate to make State estimates by conventional procedures. Several estimators were considered and a synthetic estimator

was selected to produce State estimates of disability. Since this publication, other estimators, including modifications of the synthetic estimator, have been investigated by Levy (1971) Gonzalez and Hoza (1975), Schaible (1975), and Royall (1977). However most of the research into how to make estimates for small areas has been devoted to evaluating the synthetic estimator. Levy (1971) used mortality data to evaluate average relative errors of synthetic estimates for States. Gonzalez (1973) suggested an estimated "average mean square error" as a measure for evaluating the synthetic estimator and used estimates of the number of dilapidated housing units to investigate the bias of this estimator. Gonzalez and Hoza (1975) compared synthetic estimates of county unemployment rates from the CPS to 1970 census results. Namekata, Levy and O'Rourke (1975) investigated synthetic State estimates of work loss disability in a similar manner. Levy and French (1977) discussed the properties of three small area estimators and compared several synthetic estimators which differed in the ancillary information used to produce the synthetic estimates.

III. COMPOSITE ESTIMATORS

It is evident that at some point, as the sample size in a small area increases, a direct estimator becomes more desirable than a Synthetic one. This is true whether or not the sample was designed to produce estimates for small areas. Gonzalez and Waksberg (1973) and Schaible, Brock and Schnack (1977a) compared errors of synthetic and direct estimates for Standard Megopolitan Statistical Areas and counties. The authors of both papers concluded that when small area sample sizes were relatively small the synthetic estimator outperformed the simple direct, whereas, when the sample sizes were large the direct outperformed the synthetic. These results suggest that a weighted sum of these two estimators would be an alternative to choosing one over the other.

Estimators that are weighted sums of two component estimators have been studied previously. The James-Stein estimator (James and Stein 1961) is such a weighted sum. Efron and Morris (1973, 1975) have generalized this estimator. In the 1968 publication cited above a composite estimator consisting of a synthetic estimator and an adaptation of a regression estimator was considered. Royall (1973) in a discussion of papers by Gonzalez (1973) and Ericksen (1973), suggested that a choice between direct and synthetic approaches need not be made but that"... a combination of the two is better than either taken alone." Also, as related by Gonzalez and Hoza (1975), "In a seminar given at the Bureau of the Census in March 1975, Madow suggested a combination of synthetic estimates and observed values for the primary sampling units included in the CPS." Royall (1977) has investigated optimal estimators under various population models.

Schaible, Brock, and Schnack (1977b) compared the performance of a composite estimator with that or direct and synthetic component estimators using data from the HIS and the 1970 census.

To define the composite estimator more precisely let $\bar{Y}'_d$ and $\bar{Y}''_d$ be estimators for $\bar{Y}_d$, the population value for small area d. The general form of a composite estimator may then be written as

$$\hat{\bar{Y}}_d = C_d\bar{Y}'_d + (1-C_d)\bar{Y}''_d \quad . \qquad (1)$$

The mean square error (MSE) of this estimator may be written as

$$\text{MSE } \hat{\bar{Y}}_d = C_d\text{MSE } \bar{Y}'_d + (1-C_d)\text{ MSE } \bar{Y}''_d - C_d(1-C_d)\ E(\bar{Y}'_d - \bar{Y}''_d)^2$$

minimizing this mean square error with respect to c_d, it is easily shown that the weight C^*_d that gives the composite estimator minim mean square error is

$$C^*_d = \frac{\text{MSE } \bar{Y}''_d - E(\bar{Y}'_d-\bar{Y}_d)(\bar{Y}''_d-\bar{Y}_d)}{\text{MSE } \bar{Y}'_d + \text{MSE } \bar{Y}''_d - 2E(\bar{Y}'_d-\bar{Y}_d)(\bar{Y}''_d-\bar{Y}_d)} \quad . \qquad (2)$$

In practice the individual quantities in this weight are difficult to estimate, particularly the term $E(\bar{Y}'_d-\bar{Y}_d)(\bar{Y}''_d-\bar{Y}_d)$. If both component estimators are unbiased and independent this term is zero. An alternative condition under which expression (2) becomes more manageable is when $E(\bar{Y}'_d-\bar{Y}_d)(\bar{Y}''_d-\bar{Y}_d)$ is small relative to $\text{MSE } \bar{Y}''_d$. In this case the weight (2) may be written as

$$C^*_d \approx \frac{1}{1 + R_d}, \qquad (3)$$

where $R_d = \text{MSE } \bar{Y}'_d/\text{MSE } \bar{Y}''_d$.

The heighting scheme (3) can be viewed as one in which each component estimator is first weighted by the inverse of its mean square error,

and then the two component weights normalized so that they sum to unity. This approximate weight can only range between zero and one, whereas the exact weight (2) is not necessarily so restricted. It should be noted that an estimate of the weight (3) does not require individual estimates of the component mean square errors; it requires only an estimate of their relative size.

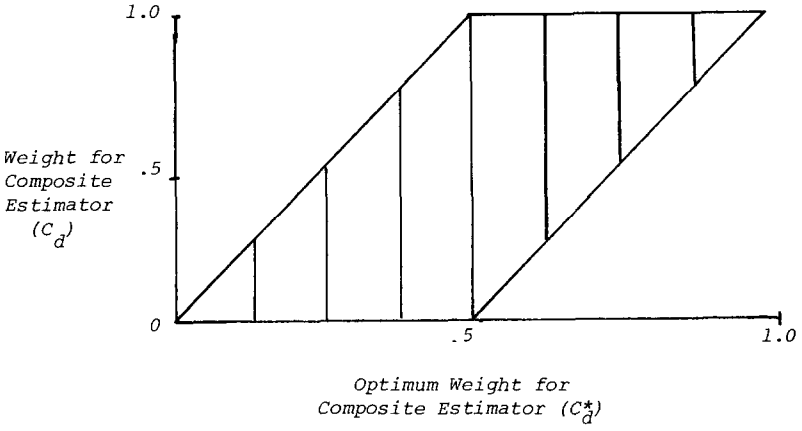
It is easily shown that if C_d : is restricted to the interval (0,1) the mean square error of the composite estimator is smaller than the larger of the two mean square errors of the component estimators regardless of the weight used.

Royall (1977) has shown that if the component estimators are unbiased, the composite estimator has smaller variance than either component estimator when $2C_d^* - 1 \leq C_d \leq 2C_d^*$.

It should be noted that if the component estimators are biased, the composite estimator has smaller mean square error than that of either component estimator under the same conditions on C_d. The width of this interval is one. However, when C_d is restricted to be between zero and one, the width of this interval varies with the size of the optimum weight, as may be seen in figure 1. When the optimum weight is close to either zero or one, there is little room for error in an estimate of the optimum weight if the composite estimator is to outperform either component estimator. The optimum weight will be close to zero or one when one of the component estimators has a much larger mean square error than the other. In this case, the estimator with large mean square error has but little information to add, and it is likely that if the relative sizes of the mean square errors of the component estimators are known, the estimator with small mean square error would be used rather than a compos ite estimator. If the mean square errors of the two component estimators are equal, then the optimum weight is one-half, and as may be seen in figure 1, the composite will outperform either component estimator regardless of the weight chosen.

FIGURE 1

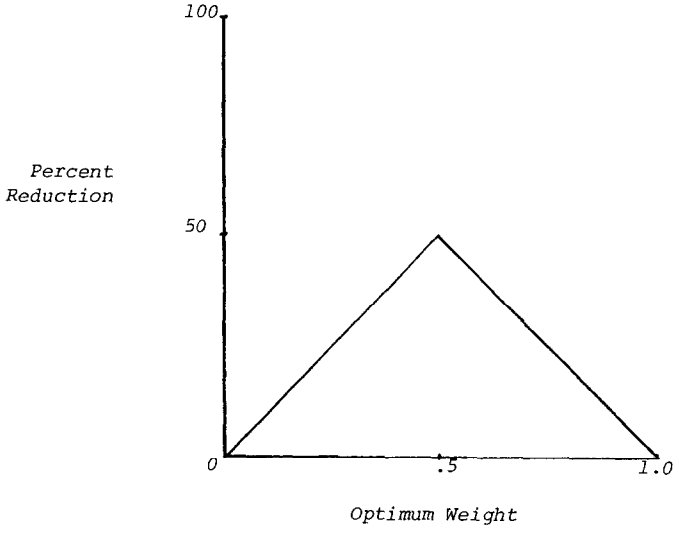
The Range of Weights (C_d) for which the Composite Estimator has Smaller MSE than either Component Estimator.



If the expected crossproduct term in equation (2) is small relative to the mean square error of the second component estimator, then the percent reduction in the mean square error of the composite estimator as compared to the smaller of the mean square errors of the two component estimators is shown in figure 2. A reduction of 50 percent can be expected when the optimum weight is one half. The percent reduction decreases to zero when the optimum weight approaches zero or one. When the mean square error of the composite estimator is compared to the larger mean square error of the two component estimators, the minimum percent reduction is 50 percent when the optimum weight is one half and approaches 100 percent as the optimum weight approaches zero or one.

FIGURE 2

Percent Reduction in the Mean Square Error of the Composite Estimator as Compared to the Mean Square Error of the Component Estimator with Smaller Mean Square Error.



IV. EMPIRICAL RESULTS

To further investigate the choice of weights for the composite estimator and to compare composite estimators with more traditional ones, estimates for the 48 contiguous States and Alaska were made from the 1969-71 data years of the Health Interview survey. The follming five variables obtained in a similar manner in the 1970 census were selected: the percent of the population less than one year of age, and the percents married, separated, having completed high school, and having completed college. Comparable values from the Bureau of Census Public Use Sample Tapes were treated as population values ($\bar{Y}_d$) for comparison with estimates from the HIS sample data. For this investigation the sample mean or simple direct estimator ($\bar{Y}'_d$) and the synthetic estimator ($\bar{Y}''_d$) were chosen as the two component estimators. Both estimators are defined in appendix I.

Weighting schemes for the composite estimator were defined under three different models or sets of assumptions. The first model allows the mean square error of each component estimator to vary across small areas, i.e. $MSE\ \bar{Y}'_d = b'_d$, $MSE\ \bar{Y}''_d = b''_d$. The second model assumes that the mean square error of each component estimator is constant across small areas, i.e. $MSE\ \bar{Y}'_d = b'$, $MSE\ \bar{Y}''_d = b''$. Finally, the third model assumes that the error function of the simple direct estimator varies across small areas but that the error function of the synthetic estimator does not; more specifically, $MSE\ \bar{Y}'_d = b'/n_d$, $MSE\ \bar{Y}''_d = b''$. Although Model 1 is perhaps the most realistic, the estimation of component estimator mean square errors for each small area is generally impractical. Under Model 2 the assumption that the mean square error of the simple direct estimator is constant over all small areas is not valid in many applications. When small area estimates are being made from large national surveys, the sample sizes in small areas vary considerably, and estimates for areas with large sample sizes generally have smaller errors than those with small sample sizes. Model 3 has the advantage that the individual Quantities to be estimated in the weight are constant across small areas, but the use of b'/n_d to represent the $MSE\ \bar{Y}'_d$ is more realistic than the constant used in Model 2.

Nine composite estimates were made for each State and for each variable by estimating the weight, C^*_d, by three different methods

for each of the three models. The first method used the estimate of the minimum mean square error weight specified in equation (2). The second method used the same approach but restricted this weight to the interval [0,1]. The third method used an estimate of the approdmate minimum mean square error weight specified in equation (3). The particular estimators used to estimate these weights under each model are given in appendix II.

The nine composite estimates and two component estimates were compared to census population values and squared errors were computed. For the five variables investigated, table 1 shows average squared errors and correlation coefficients of estimate with population value for each of these estimators. The zero average squared errors and perfect correlation coefficients shown in the first column under Model 1 reflect the fact that the composite estimator has zero mean square error, i.e. $\hat{\bar{Y}}_d = \bar{Y}_d$, when the actual errors in the two component estimates are used in estimating the minimum mean square error weght given in equation (2). Under Models 2 and 3 where information from all States is used to estimate the minimum mean square error weight for State d this is, of course, not true. Under Model 1 the restriction of the weight to the interval [0,1] increased the average squared errors and decreased correlation coefficients in all variables. Under Models 2 and 3 this restriction produced negligible changes in average squared errors and correlation coefficients. Under all models the approdmate weight (3) produced averaged squared errors and correlation coefficients similar to those of the restricted minimum mean square error weight. The average squared errors of the composite estimators were as small as or smaller than the corresponding average squared errors of either component estimator. Reductions in average squared error ranged from 0 percent to 45 percent when the composite estimator average squared error was compared to the smaller of the average squared errors of the two component estimators, and from 40 percent to 90 percent when compared to the larger of the two average squared errors. A similar trend is evident in the correlation coefficients.

Model 3 assumes that MSE $\bar{Y}'_d = b'/n_d$ and MSE $\bar{Y}''_d = b''$ so that the approximate MSME weight (3) is determined by

$$R_d = b'/b''n_d = R/n_d$$

For the results presented in table 1 R was estimated as specified in appendix II. The information used to estimate R in this paper will not be available in practice, so that the effect of poor estimates of R needs to be investigated. Table 2 gives an indication of the flat-

ness of the average squared error curve for a range of values near the optimum weight. Even when large errors in estimates of the ratio R occur, the average squared error of the composite estimator is often smaller than that of either component estimator. Also, as would be expected, in no instance is the average squared error of the composite estimator greater than the larger average squared error of the two component estimators. This insensitivity to poor estimates of R is an important characteristic of the composite estimator. Methods for estimating weights for composite estimators are still being developed, and without this characteristic the usefulness of this composite estimator would be limited. These empirical results are consistent with results reported by Royall (1977) which show that in the case of unbiased component estimators the variance curve of the composite estimator is relatively flat in the vicinity of the optimum weight.

TABLE 1

Average Squared Errors and Correlation Coefficients of the Direct, Synthetic and Several Composite Estimators for Five Variables, Forty-Nine States, Health Interview Survey 1969-1971.

Percent of Population	Direct	Synthetic	Model 1 $MSE\bar{Y}'_d = b'_d$, $MSE\bar{Y}'_d = b''_d$			Model 2 $MSE\bar{Y}'_d = b'$, $MSE\bar{Y}'_d = b''$			Model 3 $MSE\bar{Y}'_d = b'/n_d$, $MSE\bar{Y}'_d = b''$		
			MMSE	MMSE $[0,1]$	Approx.	MMSE	MMSE $[0,1]$	Approx	MMSE	MMSE $[0,1]$	Approx
	AVERAGE SQUARED ERROR										
Less than one	.16	.02	.00	.01	.01	.02	.02	.02	.02	.02	.02
Married	1.47	1.08	.00	.24	.34	.60	.60	.60	.64	.64	.64
Separated	.05	.08	.00	.01	.02	.03	.03	.03	.05	.05	.05
Completing High School	12.36	6.72	.00	1.44	2.07	5.20	5.20	5.22	4.16	3.96	3.79
Completing College	1.67	1.15	.00	.43	.53	.85	.85	.85	.87	.86	.80
	CORRELATION COEFFICIENT										
Less than one	.43	.74	1.00	.89	.83	.76	.76	.76	.73	.72	.72
Married	.76	.81	1.00	.96	.94	.89	.89	.89	.89	.89	.89
Separated	.91	.86	1.00	.98	.97	.94	.94	.94	.92	.92	.92
Completing High School	.79	.86	1.00	.97	.96	.89	.89	.89	.91	.92	.92
Completing College	.66	.62	1.00	.86	.84	.71	.71	.71	.75	.75	.75

TABLE 2

Average Squared Errors of the Model 3, Approximate MMSE Composite Estimator for Various Values of the Ratio R and for Five Variables, Forty Nine States, Health Interview Survey, 1969-1971

R	VARIABLE				
	Less Than One	Married	Separated	High School	College
0 ($\bar{Y}'$)	.16	1.47	.05	12.36	1.67
100	.13	1.24	.05	9.97	1.44
500	.08	.86	.04	6.14	1.05
1,000	.06	.72	.04	4.67	.89
2,000	.04	.64	.04	3.80	.81
3,000	.03	.64	.04	3.61	.80
4,000	.03	.64	.05	3.61	.81
5,000	.02	.66	.05	3.69	.82
6,000	.02	.67	.05	3.78	.83
7,000	.02	.69	.05	3.89	.84
10,000	.02	.73	.05	4.21	.88
15,000	.02	.78	.06	4.63	.92
20,000	.02	.82	.06	4.94	.95
∞ ($\bar{Y}''$)	.02	1.08	.08	6.72	1.15

V. SUMMARY

The composite estimator (1), a weighted sum of two component estimators, has a mean square error that is smaller than the larger of the mean square errors of the two component estimators. This statement is not as trivial as it may first seem when it is noted that little information is usually available concerning the magnitude of the mean square errors of the Component estimators. The composite estimator has a mean square error which is smaller than that of either component estimator when an appropriate weighting scheme is used. The estimation of the optimum weight for the composite estimator is a major problem which deserves further attention. However, the composite estimator is surprisingly insensitive to poor estimates of the optimumweight. This insensitivity depends on the relative sizes of the mean square errors of the component estimators. The composite estimator is most insensitive when the mean square errors of the two component estimators do not differ greatly. The percent reduction in mean square error of the composite estimator over those of component estimators also depends on the relationship between the mean square errors of the component estimators.

Data were used to produce composite estimates and to calculate squared errors and correlation coefficients of estimates versus actual values. Only small differences were apparent in average squared errors or in correlation coefficients when an approximation rather than the minimums mean square error weight was used. This was true even when a fairly unrealistic model was used to produce estimates. In all cases the composite estimator produced an average squared error as small as, or smaller than, that of either component estimator. In some cases the percent reductions in average squared errors were large.

Although composite estimators have been used to produce small area estimates, there are two major problems which need additional attention. The first problem is to decide how to estimate the composite estimator weight. Under a simple model the weighting scheme for the James-Stein estimator can be viewed as one method of estimating the composite minimum mean square error weight, but other methods may be better. Under more realistic models the relationship between the James-Stein weighting scheme and the minimum mean square error weight is not so clear. An alternative approach, which has been used to produce weights for composite estimates in the report State Estimates of Disability and Utilization of Medical Services (NCHS, 1978), is to assume specific error functions for the component estimators and for a given sample and set of small

areas to estimate the relative magnitude of the parameters for a selected group of variables. This approach, although not ideal, may be useful since the composite estimator is quite insensitive to bad estimates of minimum mean square error weights. The second problem is to discover how to provide measures of error for a composite estimator for a given small area. This problem is common to all biased small area estimators and is likely to be a difficult one to solve. One way to provide information on the performance of biased small area estimators is to compute average measures of error using variables for which actual errors can be computed. Although this information is useful, it is more useful to have some measure of how well the estimator is likely to perform in a particular small area of interest.

APPENDIX I

SIMPLE DIRECT AND SYNTHETIC ESTIMATORS

Let $Y_{d\alpha i}$ denote the observation of interest for the ith sample unit ($i=1,2,\ldots n_{d\alpha}$) in the α th ($\alpha=1,2,\ldots K$) demographic class in the dth ($d=1,2,\ldots D$) small area. The simple direct estimator for small area d is then

$$\bar{Y}'_d = \sum_{\alpha=1}^{K} \sum_{i=1}^{n_{d\alpha}} Y_{d\alpha i} / n_{d.}$$

The simple direct estimator is more widely used than the synthetic or composite estimators. Its simplicity is appealing and with appropriate sample design it is unbiased and its variance can be estimated. However, when used to estimate for small areas from samples designed for large areas, the conventional sampling theory model yields little information about the properties of this estimator. For this reason alternative estimators have been proposed.

In addition to the above notation let $N_{d\alpha}$ represent the number of units in the population in area d and class α. The sample mean of the αth demographic class for the large area is then

$$\bar{Y}_{.\alpha} = \sum_{d=1}^{D} \sum_{i=1}^{n_{d\alpha}} Y_{d\alpha i} / n_{.\alpha},$$

and the synthetic estimator for small area d is

$$\bar{Y}''_d = \sum_{\alpha=1}^{K} \frac{N_{d\alpha}}{N_{d.}} \bar{Y}_{.\alpha}$$

The α-cells for State synthetic estimates in this paper were defined to be the 64 cells created by cross-classifying the following variables:

1. Color: white; other

2. Sex: male, female

3. Age: under 17 years; 17-44 years; 45-64 years; 65 years and over

4. Family size: fewer than 5 members; 5 metiers or more

5. Industry of head of family: Standard Industrial Classifications: (1) forestry and fisheries, agriculture, contruction, mining and manufacturing; (2) all other industries.

APPENDIX II

WEIGHTING SCHEMES

The expressions used to estimate composite estimator weights are specified below. The models and weighting schemes correspond to those in text table I.

The minimum mean square error (MMSE) weight under Model 1 was estimated by

$$\hat{C}_d^* = \frac{\left(\bar{Y}'_d-\bar{Y}_d\right)^2 - \left(\bar{Y}'_d-\bar{Y}_d\right)\left(\bar{Y}''_d-\bar{Y}_d\right)}{\left(\bar{Y}'_d-Y_d\right)^2 + \left(\bar{Y}''_d-\bar{Y}_d\right)^2 - 2\left(\bar{Y}'_d-\bar{Y}_d\right)\cdot\left(\bar{Y}''_d-\bar{Y}_d\right)}$$

Note: In this case $\hat{\bar{Y}}_d=\bar{Y}_d$

The minimum mean square error (MMSE) weight under Model 2 was estimated by

$$\hat{C}_d^* = \frac{\sum^{49}\left(\bar{Y}''_d-\bar{Y}_d\right)^2/49 - \sum^{49}\left(\bar{Y}'_d-\bar{Y}_d\right)\left(\bar{Y}''_d-\bar{Y}_d\right)/49}{\sum^{49}\left(\bar{Y}'_d-\bar{Y}_d\right)^2/49 + \sum^{49}\left(\bar{Y}''_d-\bar{Y}_d\right)^2/49 - 2\sum^{49}\left(\bar{Y}'_d-\bar{Y}_d\right)\left(\bar{Y}''_d-\bar{Y}_d\right)/49}$$

The minimum mean square error (MMSE) weight under Model 3 was estimated by

$$\hat{C}_d^* = \frac{\sum^{49}\left(\bar{Y}''_d-\bar{Y}_d\right)^2/49 - \sum^{49}\left(\bar{Y}'_d-\bar{Y}_d\right)\left(\bar{Y}''_d-\bar{Y}_d\right)/49}{\hat{b}'/n_d + \sum^{49}\left(\bar{Y}''_d-\bar{Y}_d\right)^2/49 - 2\sum^{49}\left(\bar{Y}'_d-\bar{Y}_d\right)\left(\bar{Y}''_d-\bar{Y}_d\right)/49}$$

where $\hat{b}'$ was estimated by fitting a curve of the form b'/n_d to the individual squared errors of the direct estimates.

The minimum mean square error weights restricted to the interval zero to one (MMSE [0,1]) were estimated for each model as specified above except that they were restricted to the interval [0,1].

The approximate MMSE weights were estimated for each model as specified above except that the crossproduct terms were omitted.

REFERENCES

Efron, Bradley, and Morris, Carl. Stein's estimation rule and its competitors - An empirical Bayes approach. Journal of the American Statistical Association, 68(341):117-130, 1973.

Efron, Bradley, and Morris, Carl, Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association, 70 (350): 311-313, 1975.

Ericksen, Eugene P. Recent developments in estimation for local areas. Proceedings of the American Statistical Association, Social Statistics Section, 1973. pp. 37-41.

Gonzalez, Maria E. Use and evaluation of synthetic estimates. Proceedings of the American Statistical Association, Social Statistics Section, 1973. pp. 33-36.

Gonzalez, Maria E., and Waksberg, Joseph E. Estimation of the error of synthetic estimates. Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria. 1973

Gonzales, Maria E., and Hoza, Christine. Small area estimation of unemployment. Proceedings of the American Statistical Association, Social Statistics Section, 1975. pp.437-443.

James, W., and Stein, C. Estimation with quadratic loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1, Berkeley: University of California Press, 1961. pp.361-379.

Levy, Paul S. The use of mortality data in evaluating synthetic estimates. Proceedings of the American Statistical Association, Social' Statistics Section, 1971. pp. 328-331.

Levy, P.S., and French, D.K. Synthetic estimation of State health characteristics based on the Health Interview Survey. Vital and Health Statistics, Series 2-75(78-1349). Public Health Service, National Center for Health Statistics, 1977.

Namekata, Tsukasa; levy, Paul S.; and O'Rourke, Thomas W. Synthetic estimates of work loss disability for each state and the District of Columbia. Public Health Reports, 90: 532-538, 1975.

National Center for Health Statistics. Synthetic State Estimates of Disability, Public Health Service Pub. No. 1759. 1968.

National Center for Health Statistics. State Estimates of Disability and Utilization of Medical Services: United States, 1974-76. 1978 (in press).

Royall, Richard M. Discussion of two papers on recent developments in estimation of local areas. Proceedings of the American Statistical Association, Social Statistics Section, 1973. pp. 43-44.

Royall, Richard M. Statistical Theory of Small Area Estimates - Use of Prediction Models. Unpublished report prepared under contract from the National Center for Health Statistics. 1977.

Schaible, Wesley L. A Comparison of the Mean Square Errors of the Postratified, Synthetic and Modified Synthetic Estimators. Unpublished report, Office of Statistical Research, National Center for Health Statistics. 1975.

Schaible, Wesley L.; Brock, Dwight B.; and Schnack, George A. An Empirical Comparison of Two Estimators for Small Areas. Presented at the Second Annual Data Use Conference of the National Center for Health Statistics, Dallas, Texas. 1977a.

Schaible, Wesley L.; Brock, Dwight B.; and Schnack, George A. An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics. Proceedings of the American Statistical Association, Social Statistics Section, 1977b. pp. 1017-1021.

ACKNOWLEDGMENT

The author would like to thank Barry Peyton of the Office of Statistical Research, NCHS, for the computation of estimates for this paper.

Discussion

Barbara A. Bailar

If one defines a composite estimator as a weighted average of two or more estimators, one finds they have been used for many years for many different kinds of characteristics because of their desirable properties. In the two applications I know best, the Current Population Survey for labor force estimates, and the Retail Trade Survey for retail sales estimates, their variance reduction property is extremely important. It is interesting to see an application of this technique in the area of small area estimation.

In one of the earliest reports on the use of synthetic estimation by the National Center for Health Statistics, Synthetic State Estimates of Disability (1968) a composite estimate combining two different kinds of synthetic estimates was investigated. However, that composite estimator was not the estimator suggested by Schaible. Interestingly enough. at the 1973 meeting of the American Statistical Association, at which Gonzalez and Ericksen presented papers on estimators and evaluation of estimators for small areas, each of the discussants suggested composite estimators. Royall speculated that a combination of the direct estimator and the synthetic estimator would be better than either alone. Kaitz suggested a combination of the synthetic and the regression estimators to yield an estimator superior to either alone.

Let me now turn to specific comments on the Schaible paper. He introduces the composite estimator as:

$$\hat{\bar{Y}}_d = C_d\bar{Y}_d' + (1-C_d)\bar{Y}_d''$$

and then proceeds to write the mean square error (MSE) of the estimator as:

$$MSE(\hat{\bar{Y}}_d) = C_d MSE\ \bar{Y}_d' + (1-C_d)MSE\ \bar{Y}_d'' - C_d(1-C_d)E(\bar{Y}_d' - \bar{Y}_d'')^2$$

This is a curious way of writing the MSE, though correct, considering that it wasn't used in this way throughout the rest of the paper. All of the results claimed seem much easier to derive if the estimator is written as a difference estimator. I will return to this later.

The conditions that Schaible mentions that might help in estimating the weights for the optimum C_d^* are unrealistic. The first condition mentioned is when each component estimator is unbiased and the two are independent. Since one of the component estimators is the synthetic estimator, which is usually biased, this condition would rarely be met. The second condition that makes the estimation of C_d^* more manageable is when $E(\bar{Y}_d' - \bar{Y}_d)(\bar{Y}_d'' - \bar{Y}_d)$ is small relative to $MSE\bar{Y}_d''$. This, again, would occur rarely. On the other hand, the empirical results show that even if $E(\bar{Y}_d' - \bar{Y}_d)(\bar{Y}_d'' - \bar{Y}_d)$ is not small in relation to MSE $\bar{Y}_d''$ it doesn't seem to matter, at least for the characteristics studied.

It is interesting to observe that the weight is not restricted to the interval (0.1). Most of the applications would seem to confine it to this interval, but the theory holds even when this is not the case.

It was noted in the paper that Royall (1977) had shown that if the component estimators are unbiased then the composite estimator has smaller variance than either component if the weight lies between $2C_d^* - 1$ and $2C_d^*$. If the composite estimator is written in the form of a difference estimator, one can see this is an old familiar problem.

Suppose $\bar{Y}''$ the estimator with smaller variance

$$\hat{\bar{Y}} = \bar{Y}'' + C(\bar{Y}' - \bar{Y}'')$$

$$Var(\hat{\bar{Y}}) = Var(\bar{Y}'') + C^2E(\bar{Y}' - \bar{Y}'')^2 + 2CE\,\bar{Y}''(\bar{Y}' - \bar{Y}'').$$

Then, if

$$Var(\hat{\bar{Y}}) \leq Var(\bar{Y}'')$$

$$C^2E(\bar{Y}' - \bar{Y}'')^2 + 2CE\,\bar{Y}''(\bar{Y}' - \bar{Y}'') \leq 0$$

$$C\left\{\frac{CE(\bar{Y}' - \bar{Y}'')^2}{E(\bar{Y}'' - \bar{Y}')^2} + \frac{2E\,\bar{Y}''(\bar{Y}' - \bar{Y}'')}{E(\bar{Y}'' - \bar{Y}')^2}\right\} \leq 0$$

$$C\{C - 2C^*\} \leq 0$$

where

$$C^* = \frac{E\,\bar{Y}''(\bar{Y}'' - \bar{Y}')}{E(\bar{Y}'' - \bar{Y}')^2}$$

now $C \in (0,1)$ so

$$C - 2C^* \leq 0$$
$$C \leq 2C^* \quad .$$

Now reverse the roles of $\bar{Y}'$ and $\bar{Y}''$, and replace C by (1-C) to get

$$1 - C \leq 2(1 - C^*)$$

or

$$C \geq 2C^* - 1 \quad .$$

In Schaible's paper, as presented at the Workshop, his statement about the percent reduction in the mean square error was proffered without identifying some unstated assumptions. In reviewing this aspect of his paper, we again write Y as a difference estimator,

$$\hat{\bar{Y}} = \bar{Y}'' + C(\bar{Y}' - \bar{Y}'')$$

and letting $\bar{Y}''$ have the smaller MSE (the other argument is analogous and will be omitted),

$$MSE\,\hat{\bar{Y}} = E(\hat{\bar{Y}} - \bar{Y})^2$$

where $\bar{Y}$ is the population value.

$$MSE\,\hat{\bar{Y}} = E[(\bar{Y}'' - \bar{Y}) + C(\bar{Y}' - \bar{Y}'')]^2$$

$$= E(\bar{Y}'' - \bar{Y})^2 + C^2E(\bar{Y}' - \bar{Y}'')^2 + 2CE(\bar{Y}'' - \bar{Y})(\bar{Y}' - \bar{Y}'').$$

Taking the derivative with respect to C, we get

$$\frac{\partial MSE(\hat{\bar{Y}})}{\partial C} = 2CE(\bar{Y}' - \bar{Y}'')^2 + 2E(\bar{Y}'' - \bar{Y})(\bar{Y}' - \bar{Y}'')$$

and

$$C^* = \frac{E(\bar{Y}'' - \bar{Y})(\bar{Y}'' - \bar{Y}')}{E(\bar{Y}' - \bar{Y}'')^2}$$

$$\text{MSE } \hat{\bar{Y}} = \text{MSE } \bar{Y}'' + C^2 E(\bar{Y}' - \bar{Y}'')^2 - 2CE(\bar{Y}'' - \bar{Y})(\bar{Y}'' - \bar{Y}')$$

$$\text{MSE } \hat{\bar{Y}}_{opt} = \text{MSE } \bar{Y}'' + \frac{[E(\bar{Y}'' - \bar{Y})(\bar{Y}'' - \bar{Y}')]^2}{[E(\bar{Y}' - \bar{Y}'')^2]^2} [E(\bar{Y}' - \bar{Y}'')^2]$$

$$- \frac{2[E(\bar{Y}'' - \bar{Y})(\bar{Y}'' - \bar{Y}')]^2}{E(\bar{Y}' - \bar{Y}'')^2}$$

$$= \text{MSE } \bar{Y}'' - \frac{[E(\bar{Y}'' - \bar{Y})(\bar{Y}'' - \bar{Y}')]^2}{E(\bar{Y}' - \bar{Y}'')^2}$$

$$= \text{MSE } \bar{Y}'' \{1 - \rho^2\} \quad ,$$

where

$$\rho = \frac{E(\bar{Y}'' - \bar{Y})(\bar{Y}'' - \bar{Y}')}{(\text{MSE } \bar{Y}'')^{1/2}[E(\bar{Y} - \bar{Y}'')^2]^{1/2}} \quad .$$

The percent reduction in the mean square error of $\hat{\bar{Y}}$ from $\text{MSE}\hat{\bar{Y}}''$

$$R = \frac{\text{MSE } \bar{Y}'' - \text{MSE } \hat{\bar{Y}}}{\text{MSE } \bar{Y}''}$$

$$= \frac{\text{MSE } \bar{Y}'' - \text{MSE } \hat{\bar{Y}}'' (1 - \rho^2)}{\text{MSE } \bar{Y}''}$$

$$= \rho^2$$

$$= \frac{[E(\bar{Y}'' - \bar{Y})(\bar{Y}'' - \bar{Y}')]^2}{(\text{MSE } \bar{Y}'')E(\bar{Y} - \bar{Y}'')^2}$$

$$= C^* \frac{E(\bar{Y}'' - \bar{Y})(\bar{Y}'' - \bar{Y}')}{\text{MSE } \bar{Y}''} \quad .$$

Now if $\bar{Y}''$ and $\bar{Y}'$ are uncorrelated

$$R = C^* \frac{[\text{MSE } \bar{Y}'' - E(\bar{Y}'' - \bar{Y})E(\bar{Y}' - \bar{Y})]}{\text{MSE } \bar{Y}''}$$

and if $\bar{Y}''$ and $\bar{Y}'$ are unbiased,

$$R = C^*.$$

So these two conditions are necessary.

Turning now to the empirical results, there seems to be some confusion. Model 1 clearly is the most realistic model, since the

MSE's of component estimators undoubtedly vary across small areas. Model 2 is the least realistic, and model 3 is an attempt to remedy the deficiencies of model 2.

The characteristics studied do not seem to represent a wide range on which to test these models. Table 1 shows the percentage of the population having the characteristics. It was not clear from Schaible's paper whether he computed the percentages of the total population or the restricted populations, so table 1 shows the percentages calculated both ways. It also shows percentages for urban and rural populations. Only for the category "college graduates" is there a big difference between urban and rural populations.

For the two characteristics, "population less than 1 year" and "separated," the percentages of the population are very small and the nine composite estimators do not vary much. The largest average squared errors occur for the category "high school graduates," where there is a considerable difference between the models. The category "college graduates," though it showed the most difference between urban and rural populations, showed very little differences between models 2 and 3. Was this the result of its being a relatively small proportion of the population? Or because the models are relatively insensitive to the error structure across small areas?

One result that seemed peculiar was the behavior in model 3 for the categories "completing high school" and "completing college." Why wouldn't the minimum C* give a smaller mean square error than the C* restricted to (0,1) or an approximation? Considering this model as the most useful of the three presented, I would worry about its behavior for certain groups of characteristics.

One of the most interesting results was the behavior of the average squared error using the approximation to C*. The average squared errors seemed insensitive to the assumptions. However, I would like to see the results for other characteristics before I would assume this is generally the case. The census item on disability might have been an appropriate item to study, even though it was a sample item.

I certainly concur with Schaible's assessment that the choice of the method of estimating the weight and the method of providing measures of error for small areas need further attention. In addition, I would suggest further exploration of other types of composite estimators. Since the composite estimators do no worse than the poorer of the two components and often do better than either, their continued investigation may yield helpful results.

TABLE 1

Percentages of Population with
Certain Characteristics: 1970 Census

Characteristics Studied	Total population	Urban population	Rural population
Percent of Population			
Persons less than 1 year	1.7	1.7	1.7
Married persons 14+	61.4	59.9	66.0
Separated persons 14+	1.9	2.2	1.2
High school graduates 25+	31.1	31.6	29.6
College graduates 25+	10.7	12.1	6.7
Percent of Total Population			
Less than 1 year	1.7	1.7	1.7
Married	45.2	44.4	47.4
Separated	1.4	1.6	0.9
High school graduates	16.8	17.1	15.8
College graduates	5.8	6.6	3.6

REFERENCES

Ericksen, Eugene P. (1973), "Recent Developments in Estimation for Local Areas," Proceedings of the Social Statistics Section, American Statistical Association.

Gonzalez, Maria E. (1973), 'Use and Evaluation of Synthetic Estimates," Proceedings of the Social Statistics Section, American Statistical Association.

Kaitz, Hyman B. (1973), "Comments on the Papers by Gonzalez and Ericksen," Proceedings of the Social Statistics Section, American Statistical Association.

National Center for Health Statistics (1968), Synthetic State Estimates of Disability, Public Health Service, Publication No. 1759.

Royall, Richard M. (1973), "Discussion of two papers on Recent Developments in Estimation of Local Areas," Proceedings of the Social Statistics Section, American Statistical Association.

Royall, Richard M. (1977), Statistical Theory of Small Area Estimates - Use of Prediction Models, unpublished report prepared under contract to the Natlonal Center for Health Statistics.

Comments

Wesley L. Schaible

Let me reply to some of the comments made by Barbara Bailar. The question was raised whether the total population or a restricted population was used to calculate percentages. The total population was used, so that, for example, the percent of the population under one year of age used here is more analogous to the crude birth rate than to a fertility rate.

I agree that the behavior of the average squared errors of the model 3 education variables is not exactly that expected. But I don't find this as perplexing as Barbara does, especially since the empirical results do not differ greatly from those expected. Our expectations are based on theoretical mean square errors for a given small area, whereas the empirical results in the paper are observed squared errors averaged over many small areas. These are different concepts, as Maria Gonzalez and Joe Waksberg have pointed out in their papers on small area estimation. In addition, the model 3 estimator requires that smooth curves be fitted to individual squared errors. We considered a variety of minimization criteria to fit these curves. The model 3 results presented in table 1 were produced using parameters estimated with an absolute difference minimization criterion. A different criterion would undoubtedly produce a more accurate estimate of the minimum average squared error, which table 2 shows to be somewhat smaller than that estimated in table 1. Nevertheless, the education variables provide the same basic evidence as the other variables. That is, the differences among the average squared errors and correlation coefficients produced by the three model 3 weighting schemes are negligible.

The question was raised as to what assumptions underlie the figure 2 graph giving the percent reduction in mean square error as a function of the optimum weight. The percent reduction given is that which is expected under the same conditions which lead to the approximate weighting scheme. That is, the crossproduct term in the optimum weight is small relative to the mean square error of the second component estimator. It should be noted that the percent reductions in average squared errors indicated in table 1 are consistent with the reductions which would be expected from figure 2.

General Discussion

* I think, if you were to give the problem to Gene Ericksen, he might do something different. He probably would start with the census breaks in these characteristics and then try to update them in a survey, using some symptomatic areas of change. For example, population under 1 year would not be predicted well by a model in this decade because there have been some big fluctuations in births from one year to the next. Birth statistics in a regression model may give a very good prediction. In either case, you might want to look at the composite estimator. Perhaps this may not be the case where you would start with the synthetic estimator.

Another thing: If you have a lot of weights that are one half, you could have used a change time estimator fairly efficiently. Some have been discussed in the literature recently and may be worth considering. It would give you a much better chance of getting a good current weight using some average basis rather than what you are using guessed from the sample.

* If we really wanted to measure percent of the population under 1, all you need to know is how many births there were last year.

* As noted in the presentation of the data shown in the table below, the use of a weight of one half (b' = b") for each component estimator and the use of the approximate minimum mean squared error weighting scheme both outperform the constant variance Stein estimator in these data sets.

* Given this, it would be interesting to see the results from a generalized James-Stein estimator.

* We have investigated generalized James-Stein estimators corresponding to models 1 and 3 and on these data sets they give smaller average squared errors and larger correlation coefficients than the model 2 constant variance James-Stein estimator. However, they did not perform as well as any of the three minimum mean square error weighting schemes, although in a few instances the differences were small. Our investigations are by no means complete, and we are continuing our evaluation

or a variety of composite estimators, including James-Stein type weighting schemes.

(Contributing to the general discussion during this period were: Eugene Ericksen, Robert Fay, Paul Levy, and Wesley Schaible.)

Average Squared Errors and Correlation Coefficients of the Direct, Synthetic and Three Model 2 Composite Estimators for Five Variables, Forty Nine States, Health Interview Survey, 1969-1971

Percent of Population	Direct	Synthetic	Composite - Model 2		
			b'=b''	Stein	Approx. MMSE
	Average Squared Error				
Less than one	.16	.02	.05	.12	.02
Married	1.47	1.08	.62	1.15	.60
Separated	.05	.08	.03	.04	.03
Completing High School	12.36	6.72	5.72	11.95	5.22
Completing College	1.67	1.15	.88	1.57	.85
	Correlation Coefficient				
Less than one	.43	.74	.59	.46	.76
Married	.76	.81	.88	.80	.89
Separated	.91	.86	.94	.92	.94
Completing High School	.79	.86	.87	.79	.89
Completing College	.66	.62	.71	.66	.71

Prediction Models in Small Area Estimation

Richard M. Royall

1. ABSTRACT

Finite population estimation problems are formulated as prediction problems under superpopulation models. For linear regression models, a general theorem on optimal linear estimation is presented. The theorem is applied to simple cross-classification models to generate and analyze various statistics for estimating small area totals. These statistics include the synthetic and composite estimators, as well as some interesting alternatives.

2. INTRODUCTION

Problems of small area estimation vary widely with respect to available auxiliary information and with respect to the relationship of this information to the variables of interest. There is no useful general model which will accommodate all small area estimation problems. Nevertheless, many of the basic relationships can be approximated reasonably well by simple linear regression models. In section 3 we give a general theorem on finite population estimation under linear regression models, and we use this theorem in section 4 to study small area estimation in populations described by simple cross-classification models. In section 4.1 we consider a model in which an efficient unbiased estimator can use only sample units from the small area of interest. In section 4.2 we examine models under which samples from other areas can also be used. Under these latter models synthetic estimators look reasonable on intuitive grounds and are optimal under extreme conditions. In section 4.3 we study populations having a slightly more general structure. Section 5 consists of a brief discussion, and there are two sketchy appendices, one pertaining to synthetic and one to composite estimators.

We concentrate on the problem of estimating the total for a variable y over a specified small area, or domain, d, within a larger finite population. We have a sample, s, from the larger population, and this sample might be far from ideal for our problem. The sample might have been chosen for some other purpose, and it might contain few, if any, units from our domain.

Let y_i y_i represent the value of y associated with unit i. Denote the sample and non-sample units in domain d by s(d) and $\tilde{s}$(d) respectively. Then we want to estimate

$$T_d = \sum_{s(d)} y_i + \sum_{\tilde{s}(d)} y_i \qquad (1)$$

We will use what has been called the rediction a roach to this problem. This approach has been the subject of some lively critical discussions (Royall and Cwnberland 1977; Smith 1976), but recent empirical work has demonstrated its relevance in actual populations (Royall and Cumberland 1977). In the prediction approach the value of y_i is treated as the realized value of a random variable Y_i, and it is the joint distribution of the random variables $Y_1, \ldots, Y_N$ which is used in definitions of bias, variance, and standard error. From this point of view, after the sample has been observed the first sum in (1) is known, and estimating T_d is logically equivalent to predicting the value, $\sum_{\tilde{s}(d)} y_i$, of the unobserved random variable, $\sum_{\tilde{s}(d)} Y_i$. For making this prediction we can use the sample as well as whatever auxiliary information is available about the population units. We represent the auxiliary information as a matrix X of N rows, where N is the number of units in the whole population. The ith row of X is a vector of known values of auxiliary variables associated with unit i. This vector might include indicators showing whether unit i is of a particular type. It might also include such quantities as the size of unit i or previous values of the y-variable. If the y-variable of interest bears a strong relationship to the auxiliary variables, and if we can use our sample to make accurate inferences about this relationship, then we might make a useful estimate (or prediction) of the non-sample y-values in domain d.

For example, if the population can be divided into a few relatively homogeneous classes, so that y_i is strongly related to a variable x_i which shows the class to which unit i belongs, then we might estimate this class mean from our sample and use the estimate as the predicted value of y. This form of reasoning apparently underlies the synthetic estimator, and it is formalized in the prediction models to follow.

3. BEST LINEAR UNBIASED ESTIMATORS – GENERAL THEORY

Having recognized that our problem has the mathematical structure of a prediction problem, we can draw on the extensive body of prediction techniques in developing our theory. The following theorem, obtained from well-known results in linear prediction theory (Whittle 1963, chapter 4) is a slight generalization of Theorem 2.1 (Royall 1976). It gives the best linear unbiased (BLU) estimator for any linear combination of the population y's under a general linear model which relates the y's to the x's. The N values $y_1, y_2, \ldots, y_N$ are arranged as a column vector $\underset{\sim}{y}$. Without loss of generality we list the n sample units first and partition y:

$$\underset{\sim}{y} = \begin{pmatrix} \underset{\sim}{y}_I \\ \underset{\sim}{y}_{II} \end{pmatrix} ,$$

where $\underset{\sim}{y}_I$ is the n-vector of y-values associated with sample units, and $\underset{\sim}{y}_{II}$ is the vector of (N-n) non-sample y-values. We model $\underset{\sim}{y}$ as a realization of a random variable

$$\underset{\sim}{Y} = \begin{pmatrix} \underset{\sim}{Y}_I \\ \underset{\sim}{Y}_{II} \end{pmatrix}$$

having mean vector $\underset{\sim}{X}\underset{\sim}{\beta}$ and covariance matrix $\underset{\sim}{V}$. We partition $\underset{\sim}{X}$ and $\underset{\sim}{V}$ according to sample and non-sample units:

$$\underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{X}_I \\ \underset{\sim}{X}_{II} \end{pmatrix} , \qquad \underset{\sim}{V} = \begin{pmatrix} \underset{\sim}{V}_I & \underset{\sim}{V}_{I,II} \\ \underset{\sim}{V}_{II,I} & \underset{\sim}{V}_{II} \end{pmatrix} .$$

If the vector $\underset{\sim}{\beta}$ has dimension p, then $\underset{\sim}{X}_I$ is n x p, $\underset{\sim}{X}_{II}$ is (N-n) x p, $\underset{\sim}{V}_I$ is the n x n variance-covariance matrix of $\underset{\sim}{Y}_I$, $\underset{\sim}{V}_{I,II}$ is the n x (N-n) matrix of covariances between the n elements of $\underset{\sim}{Y}_I$ and the (N-n) elements of $\underset{\sim}{Y}_{II}$, etc. We consider estimating a general linear function, $\underset{\sim}{\ell}'\underset{\sim}{y}$, and we partition $\underset{\sim}{\ell}$ as we did $\underset{\sim}{y}$ so that

$$\underset{\sim}{\ell}'\underset{\sim}{y} = \underset{\sim}{\ell}'_I\underset{\sim}{y}_I + \underset{\sim}{\ell}'_{II}\underset{\sim}{y}_{II} .$$

Theorem: Among linear estimators $\underset{\sim}{h}'\underset{\sim}{Y}_I$ satisfying $E\left(\underset{\sim}{h}'\underset{\sim}{Y}_I - \underset{\sim}{\ell}'\underset{\sim}{Y}\right) = 0$, the error variance $\mathrm{Var}\left(\underset{\sim}{h}'\underset{\sim}{Y}_I - \underset{\sim}{\ell}'\underset{\sim}{Y}\right)$ is minimized by

$$\underset{\sim}{h}^{*\prime}\underset{\sim}{Y}_I = \underset{\sim}{\ell}_I'\underset{\sim}{Y}_I + \underset{\sim}{\ell}_{II}'\left[\underset{\sim}{X}_{II}\hat{\underset{\sim}{\beta}} + \underset{\sim}{V}_{II,I}\underset{\sim}{V}_I^{-1}\left(\underset{\sim}{Y}_I - \underset{\sim}{X}_I\hat{\underset{\sim}{\beta}}\right)\right]$$

where

$$\hat{\underset{\sim}{\beta}} = \left(\underset{\sim}{X}_I'\underset{\sim}{V}_I^{-1}\underset{\sim}{X}_I\right)^{-1}\underset{\sim}{X}_I'\underset{\sim}{V}_I^{-1}\underset{\sim}{Y}_I .$$

The error variance of this estimator is

$$\mathrm{Var}\left(\underset{\sim}{h}^{*\prime}\underset{\sim}{Y}_I - \underset{\sim}{\ell}'\underset{\sim}{Y}\right) = \underset{\sim}{\ell}_{II}'\left(\underset{\sim}{V}_{II} - \underset{\sim}{V}_{II,I}\underset{\sim}{V}_I^{-1}\underset{\sim}{V}_{I,II}\right)\underset{\sim}{\ell}_{II}$$

$$+\underset{\sim}{\ell}_{II}'\left(\underset{\sim}{X}_{II} - \underset{\sim}{V}_{II,I}\underset{\sim}{V}_I^{-1}\underset{\sim}{X}_I\right)\left(\underset{\sim}{X}_I'\underset{\sim}{V}_I^{-1}\underset{\sim}{X}_I\right)^{-1}\left(\underset{\sim}{X}_{II} - \underset{\sim}{V}_{II,I}\underset{\sim}{V}_I^{-1}\underset{\sim}{X}_I\right)'\underset{\sim}{\ell}_{II} . \quad \|$$

The optimal estimator consists of the sum of the known part of $\underset{\sim}{\ell}'\underset{\sim}{y}$, namely $\underset{\sim}{\ell}_I'\underset{\sim}{y}_I$, and the BLU predictor of $\underset{\sim}{\ell}_{II}'\underset{\sim}{y}_{II}$,

$$\underset{\sim}{\ell}_{II}'\left[\underset{\sim}{X}_{II}\hat{\underset{\sim}{\beta}} + \underset{\sim}{V}_{II,I}\underset{\sim}{V}_I^{-1}\left(\underset{\sim}{Y}_I - \underset{\sim}{X}_I\hat{\underset{\sim}{\beta}}\right)\right].$$

If the sample and non-sample units are uncorrelated ($\underset{\sim}{V}_{II,I}$ is the zero matrix), the predictor of $\underset{\sim}{\ell}_{II}'\underset{\sim}{Y}_{II}$ is simply $\underset{\sim}{\ell}_{II}'\underset{\sim}{X}_{II}\hat{\underset{\sim}{\beta}}$, the BLU estimator of $E\left(\underset{\sim}{\ell}_{II}'\underset{\sim}{Y}_{II}\right)$. For the present problem of estimating a domain total the $\underset{\sim}{\ell}$ vector consists of ones in the positions corresponding to the domain-d units in $\underset{\sim}{Y}$, and zeros in all other positions.

4. ESTIMATION IN CROSS-CLASSIFIED POPULATIONS

Although the preceding theorem provides estimators for problems of rather general structure, we will study only some relatively simple cases where the population units are cross-classified: each unit falls in one of D domains and also belongs to one of C classes. Thus the population is partitioned into CD class-by-domain cells. If unit i falls into class c and domain d then we say i belongs to cell (c, d). Let s(c, d) denote the sample from cell (c, d) and let $\tilde{s}$(c, d) denote the set of non-sample units in this cell. The domain total T_d to be estimated can now be written

$$T_d = \sum_c \sum_{s(c,d)} y_i + \sum_c \sum_{\tilde{s}(c,d)} y_i . \qquad (2)$$

We will denote by N_{cd} the number of units in cell (c, d) and by n_{cd} the number of units in the sample from this cell. Of course the sample s(c, d) can be the empty set, in which case $n_{cd} = 0$ and $\tilde{s}(c, d)$ is the set of all N_{cd} units in cell (c, d).

All of the models we will study here treat the Y 's within a given class-by-domain cell as being exchangeable. For our purposes this means that if different units i, j, k, and ℓ belong to the same cell, then Y_i, Y_j, Y_k, and Y_ℓ all have the same probability distribution, and the pair (Y_i, Y_j) have the same joint distribution as (Y_k, Y_ℓ). Exchangeability implies that within a given cell all units have a common mean and variance, and all pairs of units have a common covariance. This implies that if $\bar{Y}_{s(c,d)}$ is the average for sample units in cell (c, d), there are constants μ_{cd}, ρ_{cd}, and σ^2_{cd} such that

$$E\left(\bar{Y}_{s(c,d)}\right) = \mu_{cd}$$

$$\mathrm{Var}\left(\bar{Y}_{s(c,d)}\right) = \rho_{cd}\sigma^2_{cd} + \left(1 - \rho_{cd}\right)\sigma^2_{cd}/n_{cd},$$

$$\mathrm{Cov}\left(\bar{Y}_{s(c,d)}, \bar{Y}_{\tilde{s}(c,d)}\right) = \rho_{cd}\sigma^2_{cd},$$

$$\mathrm{Cov}\left(Y_i, Y_j\right) = \rho_{cd}\sigma^2_{cd} \text{ for every pair } i \neq j \text{ in cell (c, d).}$$

4.1 Cell Means Unrelated

With no further assumptions we can give an unbiased estimator of the domain total, T_d, provided that all cells in domain d are sampled. This is the "post-stratified" estimator

$$\hat{T}^{(A)}_d = \sum_c \hat{T}^{(A)}_{cd}, \qquad (3)$$

where $\hat{T}^{(A)}_{cd}$ is the expansion estimator $N_{cd}\bar{y}_{s(c,d)}$.

In fact Theorem 1 can be applied to show that if the Y's are exchangeable within cells and if Y_i and Y_j are uncorrelated whenever units i and j belong to different cells, then $\hat{T}_d^{(A)}$ is the optimal (BLU) estimator. That is, $\hat{T}_d^{(A)}$ is optimal under

Model A: For every class c and domain d,

$$E\,Y_i = \mu_{cd} \qquad i \text{ in cell } (c, d)$$

$$\mathrm{Cov}(Y_i, Y_j) = \begin{cases} \sigma_{cd}^2 & i = j, \text{ in cell } (c, d) \\ \rho_{cd}\,\sigma_{cd}^2 & i \neq j, \text{ i and j in cell } (c, d) \\ 0 & \text{i, j in different cells.} \end{cases}$$

Under Model A the error variance of $\hat{T}_d^{(A)}$ is

$$\mathrm{Var}\left(\hat{T}_d^{(A)} - T_d\right) = \sum_c \frac{N_{cd}^2}{n_{cd}} \left(1 - \frac{n_{cd}}{N_{cd}}\right)\left(1 - \rho_{cd}\right)\sigma_{cd}^2 . \qquad (4)$$

An unbiased estimate of the error-variance is obtained when, for $\left(1 - \rho_{cd}\right)\sigma_{cd}^2$ is replaced in (4) by its unbiased estimate $\sum_{i \in s(c,d)} \left(y_i - \bar{y}_{s(c,d)}\right)^2 \Big/ \left(n_{cd} - 1\right)$.

The post-stratified estimator $\hat{T}_d^{(A)}$ is unbiased under the minimal assumption of exchangeability within cells, and is optimal when additional assumptions are made concerning the variance-covariance matrix. There are two main reasons why we do not stop here. The first reason is simply that in many applications $\hat{T}_d^{(A)}$ is not available because not all cells in domain d are

sampled. (In fact we might find that domain d is not represented at all in the sample.) Then we must look for alternatives to the post-stratified estimator. The second reason for considering other estimators is that if we can use a more restrictive model than Model A, then sample units from other domains might be used to construct an estimator which is significantly more efficient than $\hat{T}_d^{(A)}$.

If we rewrite (3) as

$$\hat{T}_d^{(A)} = \sum_c n_{cd} \bar{y}_{s(c,d)} + \sum_c \left(N_{cd} - n_{cd}\right) \bar{y}_{s(c,d)}$$

and compare this with expression (2) for T_d, we see that the estimation error is

$$\left(\hat{T}_d^{(A)} - T_d\right) = \sum_c \left(N_{cd} - n_{cd}\right)\left(\bar{y}_{s(c,d)} - \bar{y}_{\tilde{s}(c,d)}\right) .$$

Clearly, the total for non-sample units in cell (c, d), $\sum_{\tilde{s}(c,d)} y_i = \left(N_{cd} - n_{cd}\right) \bar{y}_{\tilde{s}(c,d)}$, is being estimated (or predicted) by the quantity $\left(N_{cd} - n_{cd}\right)\bar{y}_{s(c,d)}$. That is, the average value over non-sample units, $\bar{y}_{\tilde{s}(c,d)}$, is estimated by the average over sample units from the same cell, $\bar{y}_{s(c,d)}$. The post-stratified estimator is unbiased under Model A because $E\left(\bar{Y}_{s(c,d)} - \bar{Y}_{\tilde{s}(c,d)}\right) = \mu_{cd} - \mu_{cd} = 0$. No assumptions relating one cell mean μ_{cd} to any other are required.

If we have no sample units from cell (c, d) then we cannot estimate μ_{cd} unless this parameter is related somehow to the parameters in cells which are sampled. This is the unfortunate and unavoidable fact which makes small-area estimation difficult. We must either draw an adequate sample from cell (c, d) or we must rely on whatever assumptions are required for estimating T_{cd} from observations on other cells. To the extent that each cell is unique, we will be frustrated in all efforts to provide

estimates for small groups of cells where only small samples are available. To the extent that there are similarities and regularities among the cells, we might use observations from some cells to make inferencesabout others, and thus produce useful small area estimates. These "similarities and regularities" are just the relationships which we express through models like those which follow.

4.2 Cell Means Determined By Class But Uncorrelated

A simple model under which unbiased estimation of T_d is possible even when some classes are not represented in the sample from domain d is the following. It treats each class as a distinct population in which the class-by-domaincells represent clusters. The model is Model B: For every class c and domain,

$$E\,Y_i = \mu_c \qquad i \text{ in class } c$$

$$\mathrm{Cov}\left(Y_i, Y_j\right) = \begin{cases} \sigma^2_{cd} & i = j \text{ in cell } (c, d) \\ \rho_{cd}\,\sigma^2_{cd} & i \neq j \text{ in cell } (c, d) \\ 0 & \text{otherwise .} \end{cases}$$

Model B would apply if the population vector y were generated by a two-stage process in which the class-c cell means μ_{cd} are themselves realized values of uncorrelated random variables having mean μ_c and variance τ^2_c and if, given μ_{cd}, the Y_i in cell (c, d) are exchangeable with mean μ_{cd}, variance σ'^2_{cd}, and covariance $\rho'_{cd}\,\sigma'^2_{cd}$. The $\sigma^2_{cd} = \tau^2_c + \sigma'^2_{cd}$ and $\rho_{cd}\,\sigma^2_{cd} = \tau^2_c + \rho'_{cd}\,\sigma'^2_{cd}$.

Model B says, in effect, that there is a common expected value for all units in a given class, regardless of their domain. It recognizes, however, through ρ_{cd}, by-domain cell (c, d) are more alike than class-c units which do not belong to the same domain. It is under this sort of model that the synthetic estimator looks reasonable:

$$\hat{T}^{(sy)}_d = \sum_c N_{cd}\,\hat{\mu}_c \qquad (5)$$

where $\hat{\mu}_c$ is some weighted average, $\sum_j \ell_{cj} \bar{y}_{s(c,j)}$, of sample means

from all class-c cells sampled. Since each of these sample means has expected value μ_c, and the ℓ_{cj} sum to one, the synthetic estimator is unbiased under Model B. Schaible (1977) has pointed out that when (5) is rewritten as $\hat{T}_d^{(sy)} = \sum_c n_{cd} \hat{\mu}_c + \sum_c \left(N_{cd} - n_{cd}\right) \hat{\mu}_c$

it becomes clear that the known sample sum, $\sum_c n_{cd} \bar{y}_{s(c,d)}$ is being

estimated, in effect, by $\sum_c n_{cd} \hat{\mu}_c$. Replacing this estimate by the known true value would appear to be an obvious way of improving the synthetic estimator. The resulting "modified synthetic estimator," $\sum_c \left[n_{cd} \bar{y}_{s(c,d)} + \left(N_{cd} - n_{cd}\right) \hat{\mu}_c \right]$ is also unbiased

under Model B. Some comparisons of this estimator's variance with the synthetic estimator's variance are shown in Appendix I. Of course, in many potential applications the effect of the modification will be slight.

Clearly the post-stratified estimator remains unbiased under Model B. We will look at the variances of synthetic and post-stratified estimators under this model after finding the BLU estimator and its variance.

We assume that every class c = 1, ... C is represented in the sample, although the sample from class c might not contain any observations from domain d. That is, although n_{cd} may be zero, $n_c = \sum_j n_{cj} > 0$ for all c. We denote the variance of a sample mean from cell (c, j) by

$$v_{cj} = \mathrm{Var}\left(\bar{Y}_{s(c,j)}\right) = \rho_{cj} \sigma_{cj}^2 + \left(1 - \rho_{cj}\right) \sigma_{cj}^2 / n_{cj} .$$

Then under Model B the BLU estimator for the cell (c, d) total is (Royall 1976)

$$\hat{T}_{cd}^{(B)} = n_{cd} \bar{y}_{s(c,d)} + \left(N_{cd} - n_{cd}\right) \left[\omega_{cd} \bar{y}_{s(c,d)} + \left(1 - \omega_{cd}\right) \hat{\mu}_c \right] \quad (6)$$

where $\omega_{cd} = n_{cd}\rho_{cd} \Big/ \left(1 - \rho_{cd} + n_{cd}\rho_{cd}\right)$, and $\hat{\mu}_c = \Sigma_j u_{cj} \bar{y}_{s(c,j)}$

with u_{cj} defined for all sampled cells (c, j) by

$$u_{cj} = v_{cj}^{-1} \Big/ \Sigma_\ell v_{c\ell}^{-1} .$$

The sum of the estimators for cell totals in domain d gives the BLU estimator of the domain total:

$$\hat{T}_d^{(B)} = \Sigma_c \hat{T}_{cd}^{(B)} . \tag{7}$$

Optimality of this estimator under Model B can be verified using the Theorem in section 3.

Before examining the error-variance of $\hat{T}_d^{(B)}$ we consider a variation on the problem: Suppose Model B applies and we have, in addition to the sample, a supplementary estimate $\dot{\mu}_c$ of the class mean μ_c, for c = 1, ... C. Now consider linear estimators of the form

$$\hat{T}_d = \Sigma_c \alpha_{cd} \bar{y}_{s(c,d)} + \Sigma_c \beta_{cd} \dot{\mu}_c .$$

If $\hat{T}_d$ unbiased under Model B we must have

$$E\left(\hat{T}_d - T_d\right) = \Sigma_c \left(\alpha_{cd} + \beta_{cd} - N_{cd}\right)\mu_c = 0$$

which implies $\beta_{cd} = N_{cd} - \alpha_{cd}$, so that we can write

$$\hat{T}_d = \Sigma_c \alpha_{cd}\left(\bar{y}_{s(c,d)} - \dot{\mu}_c\right) + \Sigma_c N_{cd} \dot{\mu}_c .$$

Reparameterizing, we let $\theta_{cd} = \left(\alpha_{cd} - n_{cd}\right) \Big/ \left(N_{cd} - n_{cd}\right)$ and see that unbiasedness implies that the estimate must have the form

$$\hat{T}_d = \sum_c n_{cd} \bar{y}_{s(c,d)} + \sum_c \left(N_{cd} - n_{cd}\right)\left[\theta_{cd} \bar{y}_{s(c,d)} + \left(1 - \theta_{cd}\right) \dot{\mu}_c\right]$$

for some constants θ_{cd}, c = 1, ... C. If $\dot{\mu}_c$ is uncorrelated with y-values in classes other than c, then optimal θ's are, for c = 1, 2 , ..., c,

$$\theta^*_{cd} = \mathrm{Cov}\left(\bar{y}_{s(c,d)} - \dot{\mu}_c , \bar{y}_{\tilde{s}(c,d)} - \dot{\mu}_c\right) \Big/ \mathrm{Var}\left(\bar{y}_{s(c,d)} - \dot{\mu}_c\right) ;$$

and with these weights $\mathrm{Var}\left(\hat{T}_d - T_d\right)$ equals

$$\sum_c \left(N_{cd} - n_{cd}\right)^2 \mathrm{Var}\left(\bar{y}_{\tilde{s}(c,d)} - \dot{\mu}_c\right)\left[1 - \rho^2\left(\bar{y}_{s(c,d)} - \dot{\mu}_c , \bar{y}_{\tilde{s}(c,d)} - \dot{\mu}_c\right)\right] \tag{8}$$

where p(a, b) denotes the correlation coefficient of a and b. In case $\mathrm{Var}\left(\dot{\mu}_c\right)$ zero (μ_c is known) the optimal weights, θ^*_{cd}, are the same weights, ω_{cd}, in (6) which are optimal when μ_c is estimated from the sample by $\hat{\mu}_c$. In this case the error-variance (8) becomes (after some reorganization)

$$\sum_c \frac{N_{cd}^2}{n_{cd}}\left(1 - \frac{n_{cd}}{N_{cd}}\right)\left(1 - \rho_{cd}\right)\sigma_{cd}^2\left[1 - \left(1 - \frac{n_{cd}}{N_{cd}}\right)\left(1 - \omega_{cd}\right)\right] . \tag{9}$$

We have written (9) as though $n_{cd} > 0$ for all c. If in fact $n_{cd} = 0$ then we take $\omega_{cd}/n_{cd} = \dfrac{\rho_{cd}}{1 - \rho_{cd}}$ and the summand in (9) is $N_{cd}^2\left[\rho_{cd}\sigma_{cd}^2 + \left(1 - \rho_{cd}\right)\sigma_{cd}^2/N_{cd}\right]$.

Now in the absence of supplementary estimates of the μ_c, $\hat{T}_d^{(B)}$ given in (7) is the BLU estimator under Model B, and its error variance, $\mathrm{Var}\left(\hat{T}_d^{(B)} - T_d\right)$ can be written

$$\underbrace{\sum_c \frac{N_{cd}^2}{n_{cd}}\left(1 - \frac{n_{cd}}{N_{cd}}\right)\left(1 - \rho_{cd}\right)\sigma_{cd}^2}_{a}\underbrace{\left\{1 - \left(1 - \frac{n_{cd}}{N_{cd}}\right)\left(1 - \omega_{cd}\right)\left(1 - u_{cd}\right)\right\}}_{b}.$$

The part labelled "a" is the variance of the post-stratified estimator, and that labelled "b" is the variance (9) attainable if the μ_c were known. For estimating the cell (c, d) total the relative efficiency of the post-stratified estimator to the BLU estimator $\hat{T}_{cd}^{(B)}$ is $1 - \left(1 - \frac{n_{cd}}{N_{cd}}\right)\left(1 - \omega_{cd}\right)\left(1 - u_{cd}\right)$,

which is at least as large as the maximum of the three quantities $\frac{n_{cd}}{N_{cd}}$, ω_{cd}, and u_{cd}. If the σ^2's are constant and the ρ's all equal ρ this relative efficiency lies between 1 (the efficiency when $\rho = 1$) and

$$\frac{n_{cd}}{N_{cd}} + \frac{n_{cd}}{n_c} - \frac{n_{cd}^2}{N_{cd}\, n_c} \quad \text{(the efficiency when } \rho = 0\text{)}.$$

The optimal estimator $\hat{T}_d^{(B)}$ depends on the ρ's and the σ^2's, which are generally unknown. However, even when incorrect values of these parameters are used, the estimator is unbiased under Model B. This suggests that estimators of this form (7) obtained using simple variance structures might prove useful under a fairly wide range of conditions. For example, if all ρ's are zero and the σ^2's are constant, $\hat{T}_d^{(B)}$ is simply $\hat{T}_d^{(S)} = \Sigma_c \hat{T}_{cd}^{(S)}$, where

$$\hat{T}_{cd}^{(S)} = n_{cd}\,\bar{y}_{s(c,d)} + \left(N_{cd} - n_{cd}\right)\Sigma_j n_{cj}\,\bar{y}_{s(c,j)} \Big/ n_c .$$

The estimator $\hat{T}_d^{(S)}$ is the modified synthetic estimator studied by Schaible (1977). Its error variance under Model B is

$$\mathrm{Var}\left(\hat{T}_d^{(S)} - T_d\right) = \sum_c \frac{N_{cd}^2}{n_{cd}}\left(1 - \frac{n_{cd}}{N_{cd}}\right)\left(1 - \rho_{cd}\right)\sigma_{cd}^2 \left\{1 - \left(1 - \frac{n_{cd}}{N_{cd}}\right)\right.$$

$$\left.\left(1 - \frac{n_{cd}}{1-\rho_{cd}}\left[\rho_{cd}\left(1 - 2\,\frac{n_{cd}}{n_c}\right) + \frac{1}{\sigma_{cd}^2}\sum_j \left(\frac{n_{cj}}{n_c}\right)^2 v_{cj}\right]\right)\right\}.$$

More generally, if the σ^2's are constant the estimator $\hat{T}_d^{(B)}$ does not depend on the value of that constant. This is clear from (6) since the σ^2's enter that expression only through the weights u_{cj}, and when these variances are constant,

$$u_{cj} = \left[n_{cj}/1 + \rho_{cj}\left(n_{cj} - 1\right)\right] \Big/ \left[\Sigma_\ell\, n_{c\ell}/1 + \rho_{cj}\left(n_{cj} - 1\right)\right].$$

If the ρ's are also set equal to a constant ρ, a family of

estimators is generated. The estimator $\hat{T}_d^{(S)}$ is obtained when $\rho = 0$, $\hat{T}_d^{(A)}$ is obtained when $\rho = 1$, and other members of this family, with the value of ρ estimated from historical or sample data, are potentially useful. We denote the estimator obtained for a given value of ρ by $\hat{T}_d^{(\rho)}$. The estimator $\hat{T}_d^{(\rho)}$ with $0 < \rho < 1$ represents a compromise between the modified synthetic estimator $\hat{T}_d^{(S)}$ and the post-stratified estimator $\hat{T}_d^{(A)}$.

Another way of striking a compromise between these two is to take a weighted average for each cell (c, d) (cf. expression (6) for $\hat{T}_{cd}^{(B)}$):

$$\hat{T}_{cd}^{(W)} = w_{cd}\hat{T}_{cd}^{(A)} + \left(1 - w_{cd}\right)\hat{T}_{cd}^{(S)}$$

$$= n_{cd}\bar{y}_{s(c,d)} + \left(N_{cd} - n_{cd}\right)\left[w_{cd}\bar{y}_{s(c,d)} + \left(1 - w_{cd}\right)\Sigma_j n_{cj}\bar{y}_{s(c,j)}/n_c\right]$$

and to estimate T_d by the sum $\Sigma_c \hat{T}_{cd}^{(W)}$. Weighted averages of this sort are often referred to as composite estimators. (See, for example, Schaible, Brock and Schnack 1973).

In Appendix II we give some simple conditions under which a composite estimator has smaller error-variance than either of its two components, and we show that these conditions are satisfied for a relatively wide range of weights. Under Model B with constant σ^2's and ρ's, say $\rho_{cj} = \rho$, the optimal weights are given by

$$w^*_{cd} = 1\Big/\left(1 + r_{cd}\right)$$

where $n_{cd}r_{cd} = \left(1-\rho\right)\left(1 - \dfrac{n_{cd}}{n_c}\right)\Big/\rho\left[\Sigma_{j\neq d}\left(\dfrac{n_{cj}}{n_c}\right)^2 + \left(1 - \dfrac{n_{cd}}{n_c}\right)^2\right]$.

For a given value of ρ, the composite estimator $\hat{T}_d^{(W)}$ which uses the weights in (10) is closely related to $\hat{T}_d^{(\rho)}$. In both cases, increasing either n_{cd} or ρ gives relatively more weight to the cell sample mean $\bar{y}_{s(c,d)}$ in estimating the total T_{cd}.

When $\rho = 0$, $\hat{T}_d^{(W)} = \hat{T}_d^{(\rho)} = \hat{T}_d^{(S)}$, while when $\rho = 1$, $\hat{T}_d^{(W)} = \hat{T}_d^{(\rho)} = \hat{T}_d^{(A)}$.

For intermediate value of ρ the main difference between $\hat{T}_d^{(W)}$ and $\hat{T}_d^{(\rho)}$ is in their respective estimates of μ_c,

$$\frac{\sum_j n_{cj} \bar{y}_{s(c,j)}}{n_c} \quad \text{and} \quad \frac{\sum_j \bar{y}_{s(c,j)} \Big/ \left[1 + (1-\rho)/n_{cj}\,\rho\right]}{\sum_j 1 \Big/ \left[1 + (1-\rho)/n_{cj}\,\rho\right]}.$$

The estimate of μ_c used in $\hat{T}_d^{(W)}$ gives sample mean $\bar{y}_{s(c,j)}$ weight proportional to n_{cj},, while the estimate used in $\hat{T}_d^{(\rho)}$ gives the sample means more nearly equal weights. For this reason $\hat{T}_d^{(\rho)}$ appears to provide better protection from domination by cells with unusually large sample sizes.

4.3 Cell Means Determined By Class And Correlated Within Domains

Model B allows, through the parameter ρ_{cd}, for possibly important differences between domains within each class c. That is, for i and j both belonging to class c, the expected value of $\left(Y_i - Y_j\right)^2$ might be much smaller when i and j belong to the same domain than when they belong to different ones. A weakness of this model is that it does not express the possibility that the differences between domains might be fairly consistent from class to class. For example, when T_{cd} exceeds its expected value $N_{cd}\,\mu_c$ the other cell totals in the same domain, $T_{c'd}$, might tend to exceed their

expected values. There are various ways in which we can allow for this possibility. One is simply to modify Model A, setting $\mu_{cd} = \mu_c + \mu_d$, so that there is an additive "domain effect." Another is to treat the μ_{cd} as realized values of random variables (so that Model A is a conditional model, given the μ_{cd}'s); the joint distribution of the μ_{cd}'s is such that μ_{cd} and $\mu_{c'd'}$ are positively correlated if either $c = c'$ or $d = d'$. This leads to a model in which all the Y_i's have the same expected value (the a priori expected value of the μ_{cd}'s) but in which Y_i and Y_j are positively correlated if i and j belong either to the same class or to the same domain. A third possible alternative generalizes Model B, treating μ_{cd} as a random variable with expected value μ_c. However it allows μ_{cd} and $\mu_{c'd'}$ to be positively correlated whenever $d = d'$. This model specifies fixed effects for classes, but allows class means to be correlated within a domain. All of these models should be investigated, but for now we consider only the third: Model C: For every class c and domain d

$$E(Y_i) = \mu_c \qquad \text{i belongs to class c}$$

$$\mathrm{Cov}(Y_i, Y_j) = \begin{cases} \sigma^2_{cd} & \text{i = j, i in cell (c, d)} \\ \rho_{cd}\,\sigma^2_{cd} & \text{i} \neq \text{j, i, j in cell (c, d)} \\ \tau_d & \text{i in cell (c, d), j in cell (c', d), c} \neq \text{c'} \\ 0 & \text{i, j in different domains.} \end{cases}$$

Under this model the cell averages satisfy:

$$E\left(\bar{Y}_{s(c,d)}\right) = E\left(\bar{Y}_{\tilde{s}(c,d)}\right) = \mu_c ,$$

$$\mathrm{Cov}\left(\bar{Y}_{s(c,d)}, \bar{Y}_{s(c',d')}\right) = \begin{cases} \rho_{cd}\sigma^2_{cd} + (1-\rho_{cd})\sigma^2_{cd}/n_{cd} & c = c' \text{ and } d = d' \\ \tau_d & c \neq c',\ d = d' \\ 0 & d \neq d' \end{cases}$$

$$\mathrm{Cov}\left(\bar{Y}_{s(c,d)}, \bar{Y}_{\tilde{s}(c',d')}\right) = \begin{cases} \rho_{cd}\sigma^2_{cd} & c = c',\ d = d' \\ \tau_d & c \neq c',\ d = d' \\ 0 & d \neq d'. \end{cases}$$

Note that $\mathrm{Var}\left(\bar{Y}_{s(c,d)}\right)$ which we denote by v_{cd}, is the same as under Models A and B.

A thorough analysis of Model C cannot be undertaken here. We will content ourselves with examining (i) the effects of the correlations introduced in Model C on the estimators already considered and (ii) the optimal estimator $\hat{T}^{(C)}_d$ obtained for a computationally simple special case of Model C.

Note that Models B and C differ only in their covariance structure. For this reason linear estimators such as $\hat{T}^{(A)}_d$, $\hat{T}^{(B)}_d$, and $\hat{T}^{(W)}_d$ which are unbiased under B remain unbiased under C. We now consider the effect of the covariances, τ_d, on the variances of these estimators. All three estimators have the general form

$\sum_c n_{cd}\bar{y}_{s(c,d)} + \left(N_{cd} - n_{cd}\right)\sum_j \ell_{cj}\bar{y}_{s(c,j)}$ for some constants

$0 \le \ell_{cj} \le 1$ which sum to one, and for which $\ell_{cj} = 0$ if $n_{cj} = 0$. For any estimator of this form the error-variance under Model C

$$\mathrm{Var}\left(\hat{T}_d - T_d\right) = \mathrm{Var}\sum_c \left(N_{cd} - n_{cd}\right)\left(\sum_j \ell_{cj}\bar{y}_{s(c,j)} - \bar{y}_{s(c,d)}\right)$$

$$= \mathrm{Var}\sum_c \left(N_{cd} - n_{cd}\right)\left(\bar{y}_{s(c,d)} - \bar{y}_{s(c,d)}\right)$$

$$+ \sum_c \left(N_{cd} - n_{cd}\right)^2 \left[\sum_j \ell_{cj}^2 v_{cj} - v_{cd} + 2\left(1 - \ell_{cd}\right)\rho_{cd}\sigma_{cd}^2\right]$$

$$+ \sum\sum_{c \neq c'} \left(N_{cd} - n_{cd}\right)\left(N_{c'd} - n_{c'd}\right)\left[\sum_j \ell_{cj}\ell_{c'j}\tau_j - \ell_{cd}\tau_d - \ell_{c'd}\tau_d + \tau_d\right].$$

Now the summand in the third term is

$$\left(N_{cd} - n_{cd}\right)\left(N_{c'd} - n_{c'd}\right)\left[\sum_{j \neq d} \ell_{cj}\ell_{c'j}\tau_j + \tau_d\left(1 - \ell_{cd}\right)\left(1 - \ell_{c'd}\right)\right],$$

which is non-negative if the τ_j's are non-negative and the ℓ's do not exceed one. Thus the positive covariance terms $\{\tau_j\}$ increase the variance of the estimators. An exception is the post-stratified estimator, which is obtained when, for every c, $\ell_{cd} = 1$ and $\ell_{cj} = 0$ for all $j \neq d$. For the post-stratified estimator the third term in (11) vanishes.

The BLU estimator $\hat{T}_d^{(C)}$ under Model C depends on the ρ's, σ^2's, and τ's, but as before, use of incorrect values in $\hat{T}_d^{(C)}$ does not introduce a bias under the model. If the values used are approximately correct, the estimator will be approximately optimal. By setting these parameters equal to constants, ρ, σ^2, and τ we generate a family of estimators. This proves to be a two-parameter family in which the estimator depends only on ρ and the ratio, τ/σ^2. Using historical or sample data to estimate these two quantities, we can choose a member of this family which

might compare favorably with the estimator $\hat{T}_d^{(\rho)}$ obtained using the same value of ρ but taking $\tau = 0$.

Because of the exchangeability within cells, we can find $\hat{T}_d^{(C)}$ by applying the Theorem in section 3 to the condensed problem in which $\underset{\sim}{Y}_I$ is the vector of all cell sample means and $\underset{\sim}{Y}_{II}$ is the vector of means of non-sample units in domain d cells. Even with simplification, and restricting the ρ's, σ^2's, and τ's to be constants, the formula for $\hat{T}_d^{(C)}$ needs more than a casual inspection for its appreciation. We will not undertake the necessary analysis here but will look at the very special case in which all of the cell sample sizes n_{cj} equal the same constant, m. This can only suggest the direction in which use of Model C will carry us away from the estimators appropriate under Model B.

We denote by $\bar{y}_{c\cdot}$ the average of sample means from cells in class c, by $\bar{y}_{\cdot d}$ the average of sample means from domain d.

$$\bar{y}_{\cdot d} = \frac{1}{C} \sum_{c=1}^{C} \bar{y}_{s(c,d)} ,$$

andby $\bar{y}_{\cdot\cdot}$ the average of all the sample means

$$\bar{y}_{\cdot\cdot} = \frac{1}{C} \sum_{c=1}^{C} \bar{y}_{c\cdot} .$$

Then the BLU estimator under Model C, for constant n's, ρ's, σ^2's and τ's is

$$\hat{T}_d^{(C)} = \sum_{c=1}^{C} m \bar{y}_{s(c,d)} + \sum_{c=1}^{C} \left(N_{cd} - m\right)\left[\dot{\omega} \bar{y}_{s(c,d)} + (1-\dot{\omega}) \bar{y}_{c\cdot}\right]$$

$$+ \left[\left(\sum_c N_{cd}\right) - Cm\right] \alpha (1-\dot{\omega})\left(\bar{y}_{\cdot d} - \bar{y}_{\cdot\cdot}\right)$$

where $\dot{\omega} = \left(\rho\sigma^2 - \tau \right) \Big/ \left[\rho\sigma^2 - \tau + (1-\rho)/m \right]$

$\alpha = C\tau \Big/ \left[C\tau + \rho\sigma^2 - \tau + (1-\rho)/m \right]$.

The final term in this estimator can be interpreted as a correction for the "domain effect" estimated by $\bar{y}_{\cdot d} - \bar{y}_{\cdot\cdot}$. This effect is a result of the correlation among the class-cell means within each domain, and the correction term vanishes as this correlation vanishes ($\tau \to 0$). The estimator can also be written

$$\hat{T}_d = T_d^{(ps)} - (1 - \dot{\omega}) \sum_{c=1}^{C} \left(N_{cd} - m \right) \left[\bar{y}_{s(c,d)} - \bar{y}_{c\cdot} - \alpha \left(\bar{y}_{\cdot d} - \bar{y}_{\cdot\cdot} \right) \right].$$

As m becomes large the weight $1 - \dot{\omega}$ approaches zero and the estimator is approximately the post-stratified estimator.

5. DISCUSSION

We have focussed on simple cross-classification models as tools for studying the synthetic estimator and some alternatives. We have assumed that the numbers of units falling into the different classes within our domain of interest are known. Often much more is known, and as Gonzalez and Waksberg (1973) have suggested, this additional local area information might be used to improve on synthetic estimates. Here again, prediction models can be used to express the relationships among all the variables, and to suggest and compare alternative estimators.

A very important use of prediction models, which we have not been able to treat here, is in suggesting and analyzing variance estimators (Royall and Cumberland 1977; Royall and Eberhardt 1975). The variance estimation theory based on prediction models, in contrast to the theory based on random choice of sample units, pertains to the actual sample used in estimation, not to the estimator's average performance over some other samples which might have been selected, or on average properties over different domains. The calculations are all made conditionally, given the sample s which was actually observed.

A workshop of this sort, focused on a specific technique, can spur development, but it can also be dangerous. The danger is that, from hearing many people speak many words about synthetic estimation we become comfortable with the technique. The idea and the jargon become familiar, and it is easy to accept that "Since all these people are studying synthetic estimation, it must be okay." We must remain skeptical and not allow

familiarity to dull our healthy skepticism. There is reason for some optimism, but it must be guarded optimism. One of the benefits of the prediction approach is that by holding s fixed, it forces us to examine carefully those relationships between variables which in fact enable us to use observations on some units to make inferences about others. When these relationships are weak and uncertain, then so are our inferences. There is no "repeated sampling" distribution to use to gloss over this fact. If most of our sample data from North Carolina come from one region, and if we do not know much about the relationships among the variables, then we cannot make reliable estimates for the state. This is true regardless of whether or not a repetition of our sampling plan might provide a larger sample, or a better-distributed one, from this state. Using data from South Carolina and Virginia in estimating the North Carolina total entails assumptions that certain relationships among variables are the same in North Carolina as in these other places; using the prediction approach forces us to make these assumptions explicit and in doing so to realize just how essentially difficult small area estimation problems are.

APPENDIX I: VARIANCES OF SYNTHETIC AND MODIFIED SYNTHETIC ESTIMATORS

For a given synthetic estimator (5) the corresponding modified synthetic estimator will have smaller variance under Model B if the differences $\hat{\mu}_c - \bar{y}_{s(c,d)}$ and $\hat{\mu}_c - \bar{y}_{\tilde{s}(c,d)}$ are positively correlated for all classes c. This is because

$$\mathrm{Var}\left(\hat{T}_d^{(sy)} - T_d\right) = \mathrm{Var}\sum_c \left(N_{cd} - n_{cd}\right)\left(\hat{\mu}_c - \bar{y}_{\tilde{s}(c,d)}\right)$$

$$+ 2\sum_c n_{cd}\left(N_{cd} - n_{cd}\right)\mathrm{Cov}\left(\hat{\mu}_c - \bar{y}_{\tilde{s}(c,d)}, \hat{\mu}_c - \bar{y}_{s(c,d)}\right)$$

$$+ \sum_c n_{cd}^2\,\mathrm{Var}\left(\hat{\mu}_c - \bar{y}_{s(c,d)}\right) \quad ,$$

and the first term on the right-hand side is the error-variance of the modified synthetic estimator. For the particular case in which $\hat{\mu}_c = \sum_j n_{cj}\bar{y}_{s(c,j)}/n_c$, the modified estimator $\hat{T}_d^{(S)}$ has smaller error-variance under a wide range of conditions. For example, if within class c the σ^2's and the ρ's are constants, say σ_c^2 and ρ_c, then

$$\mathrm{Cov}\left(\hat{\mu}_c - \bar{y}_{\tilde{s}(c,d)}, \hat{\mu}_c - \bar{y}_{s(c,d)}\right) = \rho_c\,\sigma_c^2\left[\sum_j\left(n_{cj}/n_c\right)^2 - 2\frac{n_{cd}}{n_c} + 1\right] > 0$$

and the modified estimator's error-variance is smaller.

APPENDIX II: COMPOSITE ESTIMATORS

Given two unbiased predictors, X and Y, of a random variable Z, we consider composite estimators (predictors) of the form α X + (1 - α) Y where $0 \leq \alpha \leq 1$ and ask

(i) What value of α is optimal?

(ii) For what range of values of α is the composite estimator better than either X or Y?

We assume only that both X and Y are unbiased predictors of Z: E (X - Z) + E (Y - Z) = 0.

Let Var X = σ_X^2, Cov (X , Z) + σ_{XZ}^2, etc. Then the error-variance of the composite estimator, Var (α X + (1 - α) Y - Z), is easily shown to be minimized when α is

$$\alpha^* = \frac{\text{Cov}(X - Y,\ Z - Y)}{\text{Var}\ (X - Y)} = \frac{\sigma_Y^2 + \sigma_{XZ} - \sigma_{YZ} - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\,\sigma_{XY}}\ .$$

In case X, Y, and Z are all uncorrelated, this is just the usual receipe - weights for X and Y should be inversely proportional to their variances, σ_X^2 and σ_Y^2.

To answer the second question we ask what values of α satisfy the inequality

$$\text{Var}\left(\alpha\ X + (1 - \alpha)\ Y - Z\right) < \text{Var}\ (Y - Z),$$

and easily find the answer to be

$$\alpha < 2\ \alpha^*\ .$$

By symmetry,

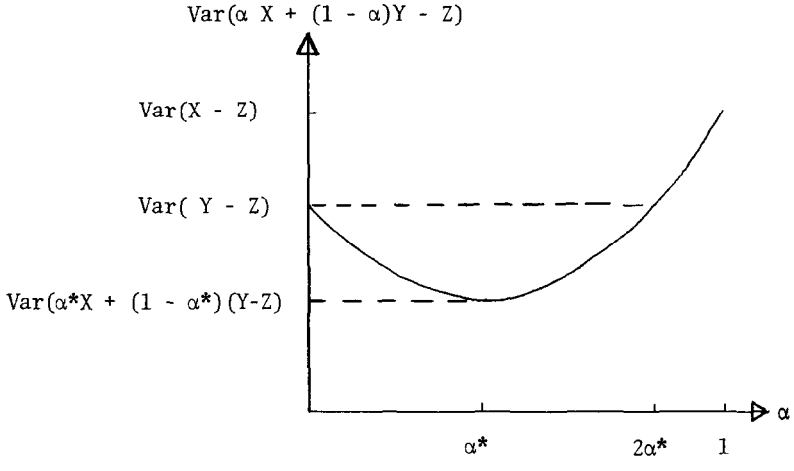
$$\text{Var}\left(X + (1 - \alpha)\ Y - Z\right) < \text{Var}\ (X - Z)$$

if and only if $(1 - \alpha) < 2\ (1 - \alpha^*)$, which is equivalent to $\alpha > 2\ \alpha^* - 1$. Thus if the optimal weight α^* is less than 1/2, the composite estimator is better than either X or Y alone if the weight assigned to X is less than twice the optimal weight α^*.

If $\alpha^* > \frac{1}{2}$, then the composite estimator is better if the weight assigned to Y, $(1 - \alpha)$, is less than twice the optimal weight, $1 - \alpha^*$. The composite estimator is better so long as

$$2\alpha^* - 1 < \alpha < 2\alpha^*.$$

The following graph illustrates the situation when Y is a better predictor than X:



From this sketch it is clear not only that the composite estimator is better for all $\alpha < 2\alpha^*$, but also that the variance curve is relatively flat in the vicinity of the optimum α^*. When $\alpha^* < 12/2$, as in the sketch, the composite estimator achieves at least 75% of the variance reduction possible,

$$\left[\mathrm{Var}(Y-Z) - \mathrm{Var}\left(\alpha X + (1-\alpha)Y - Z\right)\right] \geq .75 \left[\mathrm{Var}(Y-Z) - \mathrm{Var}\left(\alpha^* X + (1-\alpha^*)Y - Z\right)\right],$$

$$\text{if } \frac{\alpha^*}{2} < \alpha < \frac{3\alpha^*}{2}.$$

REFERENCES

Gonzalez, M.E., and Waksberg, J. Estimation of the error of synthetic estimates. Presented at first meeting of the International Association of Survey Statisticians in Vienna, Austria on August 18-25, 1973.

Royall, R.M. The linear least-squares prediction approach to two-stage sampling. J Am Stat Assoc, 71:657-664, 1976.

Royall, R.M., and Cumberland, W.G. An empirical study of prediction theory in finite population sampling: Simple random sampling and the ratio estimator. Presented at Symposium on Survey Sampling and Measurement at Chapel Hill, N.C. in April 1977. In: Namboodiri, N.K., ed. vol. to be published by Academic Press.

Royall, R.M., and Cumberland, W.G. Variance estimation in finite population sampling. J Am Stat Assoc, 72:(to appear), 1978.

Royall, R.M., and Eberhardt, K. Variance estimates for the ratio estimator. Sankhya, 37(Series C):43-52, 1975.

Royall, R.M., and Herson, J. Robust estimation in finite populations 1. J Am Stat Assoc, 68:880-889, 1973.

Schaible, W.L. A comparison of the mean square errors of the post-stratified synthetic, and modified synthetic estimators. NCHS, unpublished report draft, June 12, 1975.

Schaible, W.L., Brock, D.B., and Schnack, G.A. An empirical comparison of the simple inflation, synthetic, and composite estimators for small area statistics. Proceedings of the Social Statistics Section of the American Statistical Association, 1977.

Smith, T.M.F. The foundations of survey sampling: A review (with discussion). J R Stat Soc, 139(Series A):183-204, 1976.

Whittle, P. Prediction and Regulation by Linear Least-Square Methods. London: The English Universities Press, Ltd., 1963.

Discussion

Harold Nisselson

As Dick Royall indicated, his paper is a rather dense one, and my comments about it will be rather general.

First of all, at the risk of opening old wounds -- arguments that have been fought in many places -- I would like to distinguish between model-based design and model-based inference. I think from the point of view of helping our understanding of what makes an estimator good -- what are useful factors -- what are circumstances under which something is likely to work or not -- what Dick has done here is, I think, very interesting and useful. I would like to see a lot of work done with it, both theoretical and empirical.

I have some problems with the model of the prediction approach in this case which seems to me to be more relevant to a response error model. In fact, it would be interesting to have the concepts of this model applied to a situation in which we were concerned with response variation.

The correlation coefficients actually have a very strong role, because if you start out with a simple model (the first that Dick has in his paper), then it turns out that the estimator and the estimator of variance that Dick gave are exactly the ones that one would use from a finite population sampling model. However, if you take the second model where he assumes that the mean in each stratum is the same across all domains, then if you take the optimum estimator and assume that all the correlations are zero and the variance is constant, I don't think you would get what would be the intuitive estimator that one would use. Somehow, if you are trying to estimate a domain, let us say a statistic for a particular area, and you are using a post-stratified estimator, it doesn't seem intuitively valid that the observations that fall in a particular stratum in a particular domain should get extra weight when there doesn't seem to be any sort of reason why they should.

I think that one thing encouraging about the methods is that they seem to be fairly robust. But I don't think that the real criteria for evaluation can come from the model assumptions themselves -- they have to come from some kind of empirical work. I think, in general, this touches on the point that has been raised repeatedly at this Workshop -- the idea that we have to start devoting more attention to what are some measures of quality; what our assurance is that in producing a lot of small area estimates by analytic methods (if I may call them that rather than synthetic methods) that we're not doing as much harm as good. It may well be that the mean square error is not a very satisfying criterion, particularly from the user's point of view.

We have been finding in our own experience that most of our evaluations are, so to speak, statistician-or survey design-oriented, rather than user oriented. From this point of view, the kinds of evaluations that are used in the Gonzalez-Waksberg paper -- where they look at what is the probability that you'll do more good than harm -- or the kind of evaluation that Ericksen makes where he says: what have I done to the percent of extreme error (let's say 10 percent or more) -- I believe that kind of evaluation is goingto be more and more important.

Small area estimation is getting to be more and more important because Federal program funds are being allocated on the basis of small area estimates. There has been reference to the number of places for which Revenue Sharing estimates are made -- that estimates for 39,000 geographic areas are being made (some of which end up being combined). Of those 39,000 areas, some 29,000 have populations under 2,500; 22,000 have populations under 1,000; and 15,000 have populations under 500.

When we apply these analytic techniques to so many places of 500 or less, it may well be that besides looking at different kinds of evaluation from the user's point of view, we might want to impose different kinds of constraints on the estimates we make. One kind of constraint might be that we won't make an adjustment of more than a certain amount because we're not sure of what we are doing. Another kind of constraint would be to make sure that our estimates agree with some kind of controls, established on a more satisfying basis, at a higher level. You have seen the evidence that over and over again these methods work better for larger areas than they do for very small areas. I think that we could probably give a lot more attention to that.

Having said that, anyway, I will repeat again, I think it is an interesting paper and one that bears a lot of looking at and a lot of playing around with.

REFERENCES

Gonzalez, Maria Elena, and Waksberg, Joseph (1973), "Estimations of the Error of Synthetic Estimates." Paper presented at the meeting of the International Association of Survey Statisticians, Vienna, Austria. U.S. Bureau of the Census, Processed.

Ericksen, Eugene P. (1974), "A Regression Method for Estimating Population Changes of Local Areas," Journal of the American Statistical Association, 69, 867-875.

General Discussion

* We might think of using different techniques for different categories of areas because you have to have information for them. For example, we can establish analytic estimates at, say, the State and county level. We might have a certain amount of confidence at the county level, more confidence at the State level, and still more confidence at the national level. We can do this for categories of places, large places and small places, say, and for the balance of the counties. eSometimes there are some very large places for which we can make very good estimates. I haven't heard any discussion and would be interested in useful ways to impose constraints. This could get particularly messy if you used methods which imposed constraints, e.g., do not make an adjustment of more than so many standard errors or so many percent or something like that. Has anybody here been working with this problem? Geographic areas are the units of estimation and you then want to make a prediction. This is a problem that occurs in many applications. For example, in time series analysis, the seasonally adjusted numbers for total housing starts is not the sum of the estimates for single family starts and multiple family starts. In fact, the single family start estimate can sometimes be bigger than the total housing start estimate, for example.

* I hope you don't let that happen!

* Stein has looked at something like this problem for equal variances. An unequal variance situation is difficult. Stein did consider that you might have different sorts of shrinkage for different levels.

* Suppose we consider another aspect. What effect is there of putting a constraint like State level data on small areas that then are to add up to the State totals.

* When you start dealing with small areas each of which is a rather small part of a State (e.g. town, township), the constraint you put on the State level will have a rather trivial effect on each small area. It seems very comparable to the fact that ratio estimates don't really do much good when you're dealing with statistics that are small relative to the controls that you're using for the ratio estimate. It seems

that the effect would be about the same. However, I have no empirical information.

* Well, suppose you had a set of estimates and a large place in the balance and suppose you had a lot of confidence in the estimate for the large place and you make an adjustment to reach the county total, independently arrived at, Then you might have made a big change in the balance.

* It should--but I'm not sure whether it would be good or bad.

* NCHS has made synthetic estimates for States and also for regions. The HIS probability estimate is then used for ratio estimates. The ratio adjustment procedure did improve the average mean square error. However, not much work has as yet been done for counties. From what little has been done, it seems that it's much more difficult to do a good job of making a synthetic estimate for a county than it is for a State.

* This is not unlike a problem encountered by the Census Bureau when sample data were ratio estimated for fairly small-sized areas. When these estimates are aggregated they didn't give quite as good an estimate at the higher level of aggregation as if they were estimated directly. Sometimes there were inconsistencies. Many of these matters were studied prior to the 1960 census and after the 1960 census. There didn't seem to be any practical solutions; except, however for one thing, and that was that the people who were looking at the small area data were interested only in the small area data, and that's where the payoff is. The fact is that you're going to do some harm for the higher levels of aggregation, but in the setting where people are interested in small area data this does not carry as much weight simply because of the demands for the small area data.

* Consider now the question that was raised: how do you decide when you're doing more good than harm? A measure that has been proposed by Waksberg and Gonzalez is the "average mean square error," and it is one way of measuring how good an estimate is. There seems to be two problems with the measure. One is, how to interpret the measure. The other is, it seems to imply that you have an independent estimate for each of the small areas. Perhaps there are other ways of evaluating the synthetic estimates from the point of view of a statistical agency. How can the agency decide whether or not the synthetic estimates are good enough to publish, so that other people would accept them as usable and use them?

* This is the heart of the issue. First of all, consider the measure defined as "the average mean square error." The computation of the measure does not require any outside information. It can be done directly from the survey that was used to create the synthetic estimate. The unbiased estimates that you take for the areas are the sample estimates for the areas that you would use if you were not creating synthetic estimates. The problem here in estimation of the average mean square error for small areas is similar to the situation in which you make a variance estimate from a sample when you don't have a sufficiently large enough sample to make a good variance estimate. In that situation you probably don't have enough information to make a good synthetic

estimate either. There should be no more trouble explaining the average mean square error in a probability sense than the measures reported as average standard errors. One could start by identifying that the value of the average mean square error was calculated under some fairly general conditions. Paralleling the presentation of data on average standard errors, there would be a table or tables of average mean square errors. Then the interpretation would be that, given estimates created for individual counties, the synthetic estimate for a given county has approximately two chances in three of being within-one root mean square error of the results of a current census. The real problem is the problem of the outliers, the ones that will not be within the normal range but will be in the tails of the distribution. One criterion is that if on the average you are going to do well, this would be a reasonable way to operate. If you are going to be concerned about the few outliers, where you may be way off on your estimate, and this is a very serious concern, then you've got to hold back and say, "I'm not sure how to operate." One of the big advantages of the composite estimator is that when you have outliers at least you use some information to reduce the size of the error. Maybe you can even use it to identify the areas where the estimates may be outliers and decide not to use the synthetic estimate for these areas.

* Suppose that there is a national sample and one wanted State estimates. But suppose that there is no sample in ten of the States. Can the average mean square error be used and estimated even though there are no direct estimates for the ten States?

* Yes, just as in estimating variances, you don't have to have a sample in all States in order to compute a between-State variance. When you have data for a sample of counties you can make synthetic estimates for counties. The fact that you don't have any observations in, say, Tennessee, may stop you from making a composite estimate for Tennessee but it won't stop you from making a synthetic estimate.

* While it won't stop you from making a synthetic estimate for Tennessee, will it stop you from making a good estimate of your average mean square error?

* It doesn't appear to.

* I'm struck by a relationship between this discussion and a discussion which took place a number of years ago: When do we have a good estimate of variance? In order to study this, one of the things that people have done is to use replication methods. If one were to use a sample of counties and consider a large number of independent subsets of samples and find whether there is stability in the average mean square error, that may be a step beyond where you are now. In fact, many times when variances are calculated one does not have a specific variance for each of the statistics that are published. What you do is use average relationships and you use regression functions of variances that vary quite a lot. If we can observe that the average variation among the average mean square errors is not any worse than the average variation among variances that are used as a basis for the regression function, then you might begin to have a little more confidence in the average

mean square errors. It's something that might be worth examining as a further step towards a criterion for whether or not the average mean square error measurement is acceptable.

* For people who have a responsibility to produce synthetic estimates for program purposes it would be extremely useful to have some guidelines as to when they should and when they should not disseminate synthetic estimates or use them as a basis for their program and policy decisions.

* There is one suggestion that could be considered. If you decide to disseminate or use synthetic estimates, you can say, "I don't really know whether they are true, but I can tell you this: If you were thinking of acting on them, see if symptomatic indicators show that the action would not be unreasonable. For example, if you were thinking of building a hospital, let's say to treat cardiac arrest patients, I would put that hospital in a county that had a lot of patients who had high pressure jobs or a lot of people who were greatly overweight ."

* I don't think that you can go much further than that. You can talk about the error as much as you want, but you're still going to have outliers. There is nothing you can do. The only thing I can say is if you are planning for facilities or programs be certain there is plenty of population which generally has the problem or that in the long run probably will.

* This issue is a very difficult one--the problems of estimating errors in local estimates. One thing you can do is to take all your estimates and correlate them with the sample estimates that you have. It seems obvious that the estimates most highly correlated with the sample estimates would be the most accurate. In the area of population growth, the places that are growing fastest are more likely to have a positive bias, and the places that are growing slowest are more likely to have a negative bias, and likely the errors tend to be bigger in the areas that are growing fastest. If we assume that you're going to have to put out some kind of estimate anyway (e.g., if you have to put out an estimate as you do for revenue sharing) then one way to evaluate alternatives would simply be to look at the rank order correlations.

* One could go just a bit further and compute the regression coefficient with the sample data as the dependent variable and the final synthetic estimates as the independent variable. If, ignoring all the covariances, you get a standard error for the coefficient and if the coefficient turns out to be significantly different from one, the synthetic estimate is likely to be useful. It would be a test of the synthetic estimate provided you had enough sample, and presumably you would.

* I really get very worried about the notion of publishing and using a synthetic estimate when an agency decides to give out significant amounts of funds. For example, if one is going to give out CETA funds to those places that have high estimates of unemployment, the biggest errors in synthetic estimates are likely to be where you have high

unemployment. Perhaps the estimates are also likely to be biased consistently in one direction. That kind of bias for program use in giving out money, in my book, makes it almost inadmissible.

* What would be the choice, if it is inadmissible? How else would you advise the policymaker to give out the money if the law requires that the money be distributed?

* This raises an important point. Congress makes laws that provide formulas for distributing money. But I'm not so sure that Congress is getting the best input from statisticians that it should get to advise them on what it is reasonable to do. A committee at the National Academy of Sciences is interested in this problem and is considering the possibility (although I don't know how they'll go about doing it) of suggesting to Congress that it would be glad to advise them on pending legislation that involves the application of statistics. When you get right down to it, it's not going to solve our problems today, but as statisticians we have a strong responsibility to try to do something about that problem.

* The General Accounting Office, acting as an arm of the legislative branch, does have oversight in the research area in helping Congress. They have recently turned to an outside social science group in order to get advice. Thus, there is a model and perhaps the legislative branch through the General Accounting Office can be sensitized to this general area.

* Consider some answers to the question, 'What are the alternatives?" First, there's one simple alternative--get enough money to conduct a survey which gives local area statistics. (I'm only being half facetious on this.) There is certainly a role for synthetic estimates in dealing with the kinds of problems that are being discussed. But, there are conditions where it turns out that the distributions are such that synthetic estimates are not good. We may, unless we are careful, be getting into a Gresham's law situation: Bad statistics will drive out good statistics from the marketplace. For some purposes there may be no solution but to say: "If you really want to distribute billions of dollars a year, then you have to appropriate money to do surveys in order to get the needed state and local area data; synthetic estimates are just not good enough for this purpose." For example, it required virtually no effort to get the funds for the Survey of Income and Education in order to distribute funds for the Title I Education Act. As soon as it was pointed out that the money could not be distributed without an appropriate statistical base, the funds were provided.

* I want to add a technical observation to what you are saying. It seems to me that from some points of view, particularly if you're talking about some of these outliers, we ought to be looking at a different regression. When there is population change, our estimates tend to lag whether there is a decrease or an increase. This suggests that there is a lag in the indicator variables and the way this works itself out in providing estimates on a community-by-cormnunity basis. It might be that we ought to be projecting either using some kind of

lagged relationships in a regression, or projecting indicator variables, or something else that might be a help in some of these outlier cases.

* I'd like to clarify a little bit a point made this morning which partially answers your question: "What are the alternatives?" It is sometimes not necessary to make small-area-specific estimates even if it is necessary only in the sense of legislation requiring it. Maybe it should be proposed that the legislation requiring it should be modified and turn to estimates for classes of small areas. By this I mean that even with the direct estimator techniques we can get pretty good estimates at relatively tolerable costs for, say, the collection of cities that in the last census had between 100,000 and 500,000 population and that are in the North Central reeion. We'd make a svnthetic average estimate for any city that falls in that category. Thus, the grant would be determined by the estimates that we could make by direct means for the class in which the city belongs. This principle can be extended quite a bit. Thus, one could make estimates for, say, fifty categories in a direct way. It would seem out of the question to make direct estimates for 39,000 places through any realized set of resources.

* It may be worth noting that? in relation to CETA, there is some consideration about modifying existing legislation.

* Since this workshop is on synthetic estimates, it would be well if somewhere along the line there is some attempt to summarize the criteria that people have suggested or may suggest that could be used in deciding whether synthetic estimates met quality assurance criteria, whatever they would be, so that the estimates could be used in grant formulas, or published.

While this workshop is on synthetic estimates, there are other methods that deserve research. First, statisticians have to get involved in the subject matter areas and understand the mechanisms producing the data much better than they appear to, so that they will come up with models that are specific to certain types of interaction. Statisticians need to get involved with the data so that, perhaps, they could have a better understanding of what are the likely predictor factors and could think of models that are not necessarily linear models but that have appropriate parameters in them. The problem would be to estimate which specific parameters are relevant for the particular kind of data that you are trying to estimate. Secondly, there is a lot that can be done in survey methodology. The possibility of computerized telephone surveys is going to substantially reduce the cost of doing surveys, and it will become feasible to substantially increase sample size and distribute it over more areas with less clustering, so that we may possibly be able to produce statistics for areas for which we are now incapable of producing them. That's another avenue that I think should be investigated as an alternative to depending on only one method like synthetic estimates.

* Well, I don't want to throw cold water on your telephone procedure, since you've mentioned it a couple of times. I think it has a lot of potential, particularly with the rising costs of personal interview surveys. But, at the same time I think that a note of caution is in-

dicated, especially when the behaviors being investigated are of such a nature that persons are unlikely to admit them to:

(a) anybody;

(b) someone they can't see face to face, or

(c) somebody that they must tell--out loud, not in writing--that they have done "X."

So, despite considerable potential for telephone-collected data, their potential is and should be limited, especially in regard to covert behaviors of any kind, intrapsychic behavior, unreported crimes, and other behaviors of a highly private nature. I think there is greater opportunity for risk to human subjects in these kinds of interviews as well. Researchers should be responsible in establishing some con sensual limits.

* Consider the following about our current discussion: You might be able to use the telephone survey to gather more information at the local level on the independent or systematic variables and use your national survey, with interviewers, on the outcome variable. In synthetic estimates you are interested in improving the extent of information that is available at the local level, having something more than what is currently available, as well as in the outcome relationships.

It may also be worth keeping in mind, relative to the needs for data for, say, 39,000 units, of the possibilities of certain kinds of design strategy. For example, you could use a national sample and produce synthetic estimates for all of the local units. Then you could draw a followup sample for specific local units and see how well the synthetic estimates perform versus a specifically constructed direct estimate for each of the local areas. From these data you could try to see if you can identify the variables that might explain the residuals, then devise a modified estimator, and examine the residuals again. Thus, through the use of a sequential survey design strategy, you would get some insight. We should consider that design strategies are as important as the estimation strategy.

* It may be worth noting that for some survey designs, analysts have in the past identified the need to oversample a few illustrative types of areas of various kinds. Thus, instead of having a national self-weighted sample, the survey had a disproportionate allocation. This overlay of the additional sample in a subsample of the areas provided the analysts with a set of illustrative results specific to various types of situations. This would appear to be analogous to the current proposal for testing synthetic estimates.

(Contributing to the general discussion during this period were: Ira Cisin, Eugene Ericksen, Robert Fay, Maria Gonzalez, Gary Koch, Paul Levy, Harold Nisselson, Louise Richards, Joan Rittenhouse, Wesley Schaible, Walt Simmons, Monroe Sirken, Joseph Steinberg and Joseph Waksberg.)

A Modified Approach to Small Area Estimation

Steven B. Cohen

ABSTRACT

The ever-growing need for good estimates of the health, social, political, and economic parameters of local areas has served as the motivating force for new developments in methodology. Due to the constraints of sample size, design, and cost, accessible data from large areas for criterion variables of interest is often used jointly with local data on symptomatic variables. Furthermore, several procedures have derived local area estimators by combining symptomatic information and sample data into a multiple regression format. In those situations where assumptions are too strict or unrealistic, as when a nonlinear model is more appropriate, the merits of a more flexible approach are obvious.

Our research focuses upon a further investigation of an alternative strategy for which the most limiting assumption is the availability of good symptomatic information. A more formal representation of the model is developed within the framework of a poststratification scheme. The methodology involves ratio estimation of the respective stratum means via indicator variables which serve the purpose of classification.

To determine the accuracy of the proposed small area estimator and allow for comparisons of precision with respect to other strategies, we express the relationship between criterion and symptomatic variables by relevant continuous multivariate distributions. Specifically, comparisons are made with the results obtained using a regression estimator which is applicable to the same general setting. The theoretical framework considers multivariate stratification, where boundary determination is achieved by application of practical methods which use minimum variance stratification as a criterion.

1. INTRODUCTION

The ever-growing need for good estimates of the social, political, economic, and health parameters of local areas has been rapidly gaining recognition. The allocation of Federal aid to both States and municipalities is often dependent upon information pertaining to population, unemployment, and housing. Candidates vying for political office are particularly concerned with obtaining reliable estimates of voter pre-

ference and participation at the subnational level. Similarly, rather precise small area estimates of retail trade are essential indicators for the commercial sector.

Some useful information has been obtained from sources which include the decennial census and vital registration systems. Generally, Federal agencies have relied upon sample surveys to provide estimates of the data they require, though such estimates pertain to the entire United States or each of its four broad geographical regions. Direct estimates of data for small areas are unavailable, primarily due to sample size requirements, which are prohibitive with respect to cost, and strata designs which often cross State and county limits. Consequently, several procedures have been developed which utilize available data from large areas, local data on population, and accessible local data on ancillary (symptomatic) variables, in order to produce synthetically the desired estimates. Synthetic estimation is perhaps the most well known, defined by the United States Bureau of Census as "the method of reference to a standard national distribution." Gonzalez (1974) has offered a more comprehensive explanation--"An unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for subareas on the assumption that the small areas have the same characteristics as the larger area, we identify these estimates as synthetic estimates." Developed at the National Center for Health Statistics, the method was initially used to provide synthetic State estimates of disability from the results of the National Health Interview Survey (H.I.S.).

Procedurally, a number of demographic variables are selected (i.e., race, income, sex, age) , and when possible, national sample surveys are used to determine estimates of a characteristic (criterion variable) of interest for each of the G mutually exclusive and exhaustive domains defined by the respective demographic cross classifications. To produce the synthetic estimate of a criterion variable (Y) for local area ℓ, the NCHS model takes the form of a weighted average.

$$Y^*_\ell = \sum_{j=1}^{G} P_{\ell j} \, Y_{.j} \qquad (1.1)$$

where $P_{\ell j}$ is the proportion of local area ℓ's population represented by domain j so that $\sum_{j}^{G} P_{\ell j} = 1$, and $Y_{.j}$ is the probability estimate of the criterion variable for domain j obtained from a national sample. The more detailed estimating equation includes a regional adjustment.

Due to the nature of their derivation, the synthetic estimates will generally cluster near the mean for a specific geographic region. Consequently, the method is not particularly sensitive to many of the internal forces operating at the local level. By assuming the small areas share the same characteristics as a standard national distribution, they can only be distinguished by their respective demographic configurations. Recognizing this inherent limitation, Levy (1971) proposed a method which utilized available information at the local level on predictor (symptomatic) variables in conjunction with the NCHS estima-

tor. The following model was considered:

$$Y_{\ell}^{**} = \alpha + \beta X_{\ell} + \varepsilon_{\ell} \qquad (1.2)$$

where X_{ℓ} is the value of the symptomatic variable X for the ℓ^{th} subarea,

$$Y_{\ell}^{**} = (Y_{\ell} - Y_{\ell}^{*})/Y_{\ell}^{*} \times 100$$

where ε_{ℓ} = a term representing random error, and α and β, regression coefficients to be estimated. Here, the percentage difference between the synthetic estimate and the true value is treated as a linear function of some related predictor variable X_{ℓ}. Were the estimates $\hat{\alpha}$ and $\hat{\beta}$ available and ε_{ℓ} omitted, an estimator $\hat{Y}_{\ell}$ of Y_{ℓ} could be derived from (1.2), taking the form:

$$\hat{Y}_{\ell} = Y_{\ell}^{*}[(\hat{\alpha} + \hat{\beta} X_{\ell})/100 + 1] \qquad (1.3)$$

It is assumed that X_{ℓ} is available for every local area, but since Y_{ℓ}^{**} is a function of the true value Y_{ℓ} (which is unknown), a different strategy is used to estimate the linear coefficients. Briefly, α and β are estimated by least squares after combining local areas to form strata. The method can be extended to consider $\underset{\sim}{X}_{\ell}$ as a vector of symptomatic data, whereby $\hat{Y}_{\ell}$ is treated as a multiple regression estimator.

Ericksen (1973b) developed another technique for computing local area estimates which, unlike the NCHS estimator, solely combines symptomatic information and sample data into a multiple regression format (assuming an underlying linear model). Referred to as the regression-sample data method of local area estimation, the procedure can be outlined as follows:

> Initially, a sample of n local areas, referred to as primary sampling units (PSU's), is selected from the N local areas in the population. Estimates of the criterion variable are then computed for the respective PSU's in the sample.
>
> Symptomatic information is collected for both sample and nonsample PSU's. Typical predictor variables are the number of births, deaths, and school enrollment.
>
> The linear least squares regression estimate is computed using data for the sample PSU's only. Estimates for all subareas are then determined by substituting values of the symptomatic indicators, whether included in the respective sample or not.

Although the method is applicable for estimating any parameter for which the sample and symptomatic data is available, attention has been directed to postcensal estimates of population growth. To reduce the variability and skewness of the distribution, it is suggested that variables be written in ratio form. The procedure resembles the ratio correlation

technique, first introduced by Snow (1911) and developed by Crosetti and Schmitt (1956), which estimates the multivariate relationship among population growth and predictor variables. Postcensal estimates derived using the ratio correlation method require the fitting of a linear model to selected variables represented in terms of a ratio of measurements taken at the endpoints of the immediately preceeding intercensal period. The availability and inclusion of information pertaining to each subarea of the total population is essential. In addition, satisfactory results can only be expected when the functional form of the actual and predicted models vary only slightly. Assuming the stability of relationships between the intercensal and postcensal periods, desired small area estimates are obtained by entering the respective postcensal changes in the values of the symptomatic variables into the resulting equation. Ericksen's procedure uses data which is exclusively postcensal and obtained from sample surveys. Consequently, fewer restrictions are specified for the method to yield reliable results.

The model assumes the availability of criterion variable estimates for each of n sample PSU's and the values of p symptomatic indicators for the universe of N local areas. It takes the matrix representation:

$$\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{B} + \underset{\sim}{u} \qquad (1.4)$$

where Y, an nxl vector, is the criterion variable consisting of a set

of actual unobserved values; X, an n x (p+1) matrix denoting the set of predictor variables;

$\underset{\sim}{B}$, the (p+1) x 1 vector of regression coefficients; and $\underset{\sim}{u}$, an nxl vector, a stochastic error term.

2. LOCAL AREA ESTIMATION USING THE KALSBEEK MODEL

2.1. Methodology

The method advanced by Ericksen is most feasible when the linearity assumption is satisfied and the observed multiple correlation is high. But what decision is reached when the multiple correlation level is moderate (.5-.8) and a nonlinear model is more suitable? The inclusion of all possible symptomatic variables into the regression would increase the R^2, but most probably at the expense of an "overfit" model which increases the mean square error of the final estimate. More generally, in those situations where assumptions are too strict or unrealistic, the need for a more flexible approach is most obvious. Kalsbeek (1973) has developed one such procedure in which the most limiting assumption is the availability of good symptomatic information.

It has usually been common practice to treat the local area units as the smallest level for which the estimates are made. Contrarily, Kalsbeek suggests breaking up the local unit into constituent geographical sectors called "base units," such as townships, enumeration districts, or other geographical submits of a county. The local area for which

a variable of interest is to be estimated is referred to as the "target area" and further subdivided into constituent units called "target area base units." Unlike other methods which use symptomatic information directly for the purposes of estimation, this procedure uses the information to group base units (sample base units) from the total population. The symptomatic information is also used to classify "target area base units" into the appropriate group.

Initially, a random sample of n base units is selected from the total population of N base units. The sample base units (possibly including some "target area base units") are required to possess both symptomatic and criterion information. These units are divided into K groups (strata) using either or both types of information available. The object is to form groups which are most homogeneous within while dissimilar between themselves. Grouping can be handled by any one of several iterative procedures in cluster analysis (i.e., Automative Interaction Detection (A.I.D.), Multivariate Interactive K-Means Cluster Analysis (MIKCA).

All "target area base units" belonging to the local area in question are then assigned (classified) to one of the K groups with respect to symptomatic information. Consequently, each "target area base unit" is associated with a group of base units both similar to itself and internally homogeneous. An estimate for each of the "target area base units" with respect to the criterion variable is obtained from the sample base units in the group to which it has been assigned. These estimates are then pooled to arrive at a final estimate for the respective target area.

Our research focuses upon a further investigation of the strategy proposed by Kalsbeek. Here, a more formal representation of the model is developed within the framework of a poststratification scheme. The methodology involves ratio estimation of the respective stratum means via indicator variables which serve the purpose of classification.

2.2. Notation

Consider a population consisting of L local areas, indexed by $\ell = 1, 2, \ldots . L$, which have further been subdivided into constituent geographical sectors called "base units." There are $N_{\ell\cdot}$ base units in the ℓ^{th} local area, and

$$\sum_{\ell=1}^{L} N_{\ell\cdot} = N_{\cdot\cdot} \qquad (2.2.1)$$

in the population, individually indexed by $i = 1, 2, \ldots . N_{\ell\cdot}$, to denote the i^{th} base unit from the ℓ^{th} local area. When the local area reference is dropped, each base unit is indexed by $i = 1, 2, \ldots ., N_{\cdot\cdot}$.

Furthermore, each base unit i consists of a cluster of M_i smaller units referred to as elements. Hence, there are $M_{\ell\cdot} = \sum_{i=1}^{N_{\ell\cdot}} M_i$ elements in

the ℓ^{th} local area and $M_{..} = \sum_{\ell=1}^{L} M_{\ell.} = \sum_{i=1}^{N_{..}} M_i$ elements in the population. Let y_{ij} represent the observed value of the criterion variable for the j^{th} element within the i^{th} base unit, where

$$Y_i = \sum_{j=1}^{M_i} y_{ij} \tag{2.2.2}$$

is the i^{th} unit total.

In practice, a multistage sampling design is most appropriate. To facilitate the presentation, we assume a two stage sampling design whereby a simple random sample of n base units (first stage units) is initially drawn from the $N_{..}$ base units in the population. A subsample of m_i out of the M_i elements is then selected with equal probabilities of selection from each of the chosen sample base units. The subunits are chosen independently in different base units. The units are then divided into K strata (groups) which are rectilinearly defined. nonoverlapping, and exhaustive. Here, stratum boundary determination is achieved by application of clustering algorithms or other practical methods which consider minimum variance stratification as a criterion. Consequently, estimates of the stratum means are obtained by a method which closely resembles poststratification. To determine the criterion variable estimator for the ℓ^{th} local area, each "target base unit" is assigned to the stratum most similar with respect to symptomatic information. Thus, we have a two way classification of all base units in the population by respective strata and local areas, where $N_{\ell g}$ is the total number of base units in the g^{th} stratum from the ℓ^{th} local area.

2.3. Representation of the Model

The local area estimator of the criterion variable may be expressed in terms of an average, a proportion, or a total. Initially, we direct attention to the mean per element representation.

Assuming a two stage sampling design with subunits of unequal sizes, we define

$$\bar{y}_i = \sum_{j=1}^{m_i} \frac{y_{ij}}{m_i} \tag{2.3.1}$$

as the sample mean per element in the i^{th} base unit and

$$\bar{Y}_i = \sum_{j=1}^{M_i} y_{ij}/M_i \tag{2.3.2}$$

as the overall mean per element in the i^{th} base unit. To obtain an estimate of the g^{th} stratum mean per element, we also define the indicator variables I_{gi} (once more dropping the local area reference), such that

I_{gi} = 1 if the (first stage) base unit falls in the g^{th} stratum;
= 0 otherwise

for g = 1, 2, ... K and i = 1, 2, . . $N_{\cdot\cdot}$. Here, $\sum_{i=1}^{n} I_{gi} = n_g$, the number of sample base units belonging to the gth stratum, and $\sum_{i=1}^{N_{\cdot\cdot}} I_{gi} = N_{\cdot g}$. Consequently, let

$$\hat{\bar{\bar{y}}}_g = \frac{\sum_{i=1}^{n} I_{gi} M_i \bar{y}_i}{\sum_{i=1}^{n} I_{gi} M_i} = \frac{\sum^{n_g} M_i \bar{y}_i}{\sum^{n_g} M_i} \qquad (2.3.3)$$

(summed only over the n_g sample base units from the g^{th} stratum) be our (post-stratified) estimator of the g^{th} stratum mean per element. Since $\hat{\bar{\bar{y}}}_g$ is a ratio estimator of

$$\bar{\bar{Y}}_g = \frac{\sum_{i=1}^{N_{\cdot\cdot}} I_{gi} M_i \bar{Y}_i}{\sum_{i=1}^{N_{\cdot\cdot}} I_{gi} M_i} = \frac{\sum^{N_{\cdot g}} M_i \bar{Y}_i}{M_{\cdot g}} \qquad (2.3.4)$$

(where the sum is over the $N_{\cdot g}$ base units assigned to the g^{th} stratum), it is biased to the order of l/n. Yet, when n is large (i.e., n $\geq$ 100), the bias is negligible and the expectation of $\hat{\bar{\bar{y}}}_g$ is approximately equivalent to $\bar{\bar{Y}}_g$,

$$E(\hat{\bar{\bar{y}}}_g) \doteq \bar{\bar{Y}}_g \qquad g = 1, 2, \ldots, K \; . \qquad (2.3.5)$$

Returning to the ℓ^{th} local area, we focus attention on the "target base unit" alignment in order to weight appropriately the stratum estimators

$(\hat{\bar{\bar{y}}}_g)$ by the proportion of elements in the base units so classified. Therefore, the estimator of the criterion variable for the ℓ^{th} local area takes the following form:

$$\hat{\bar{\bar{y}}}_{\ell.} = \sum_{g=1}^{K} \frac{M_{\ell g}}{M_{\ell.}} \hat{\bar{\bar{y}}}_g \qquad (2.3.6)$$

such that

$$E(\hat{\bar{\bar{y}}}_{\ell.}) = \sum_{g=1}^{K} \frac{M_{\ell g}}{M_{\ell.}} E(\hat{\bar{\bar{y}}}_g) \doteq \sum_{g=1}^{K} \frac{M_{\ell g}}{M_{\ell.}} \hat{\bar{\bar{Y}}}_g \qquad (2.3.7)$$

when n is large. Often the sizes of $M_{\ell g}$ and $M_{\ell.}$ are only known approximately. When this occurs, the respective estimators of the stratum means are weighted by the ratio of available estimates $M'_{\ell g}$ and $M'_{\ell.}$ or by the cruder ratio $N_{\ell g}/N_{\ell.}$.

Due to the nature of its derivation, the local area estimator $\hat{\bar{\bar{y}}}_{\ell.}$ of $\bar{\bar{Y}}_{\ell}$ is biased. The observed value of the criterion variable mean per element is

$$\bar{\bar{Y}}_{\ell.} = \frac{\sum_{}^{N_{\ell.}} M_i \bar{Y}_i}{\sum_{}^{N_{\ell.}} M_i} = \frac{\sum_{}^{N_{\ell.}} M_i \bar{Y}_i}{M_{\ell.}} \qquad (2.3.8)$$

summed across only those base units in the ℓ^{th} local area. The bias,

$$B = [E(\hat{\bar{\bar{y}}}_{\ell.}) - \bar{\bar{Y}}_{\ell.}] \qquad (2.3.9)$$

can be approximated by

$$B' = \left[\sum_{g=1}^{K} \frac{M_{\ell g}}{M_{\ell.}} \bar{\bar{Y}}_g - \frac{\sum_{}^{N_{\ell.}} M_i \bar{Y}_i}{M_{\ell.}} \right] . \qquad (2.3.10)$$

Similarly, to express the local area estimator in terms of a proportion, y_{ij} is redefined, so that

y_{ij} = 1 when the j^{th} element in the i^{th} base unit has the characteristic of interest;

= 0 otherwise,

so that

$$\sum_{j=1}^{M_i} y_{ij} = Y_i \qquad (2.3.11)$$

is the total number of elements in the i^{th} base unit with the characteristic of interest.

2.4. An Expression for the Mean Squared Error of the Local Area Estimator

It has already been observed that the local area estimator $\hat{\bar{\bar{y}}}_{\ell.}$ is biased. Consequently, the mean squared error term takes the form:

$$E[(\hat{\bar{\bar{y}}}_{\ell.} - \bar{\bar{Y}}_{\ell.})^2] = E(\hat{\bar{\bar{y}}}_{\ell.} - E(\hat{\bar{\bar{y}}}_{\ell.}))^2 + (E(\hat{\bar{\bar{y}}}_{\ell.}) - \bar{\bar{Y}}_{\ell.})^2$$

$$= \text{Variance}(\hat{\bar{\bar{y}}}_{\ell.}) + (\text{Bias})^2 \qquad (2.4.1)$$

Since

$$E(\hat{\bar{\bar{y}}}_{\ell.}) \doteq \sum_{g=1}^{K} \frac{M_{\ell g}}{M_{\ell.}} \bar{\bar{Y}}_g \quad ,$$

where $\hat{\bar{\bar{y}}}_{\ell.}$ is a linear combination of the ratio estimators $\hat{\bar{\bar{y}}}_g$ g = 1, 2, ..., K (with negligible bias), the variance of $\hat{\bar{\bar{y}}}_{\ell.}$ can be approximated

$$\text{Var}(\hat{\bar{\bar{y}}}_{\ell.}) \doteq \sum_{g=1}^{K} \left(\frac{M_{\ell g}}{M_{\ell.}}\right)^2 \text{Var}(\hat{\bar{\bar{y}}}_g)$$

$$+ \sum_{g \neq g'} \sum \left(\frac{M_{\ell g}}{M_{\ell.}}\right)\left(\frac{M_{\ell g'}}{M_{\ell.}}\right) \text{Cov}(\hat{\bar{\bar{y}}}_g, \hat{\bar{\bar{y}}}_{g'}) \quad . \qquad (2.4.2)$$

If we also assume

$$\frac{\sum_{i=1}^{n} I_{gi} M_i \bar{y}_i}{\sum_{i=1}^{n} I_{gi} M_i} - \bar{\bar{Y}}_g \doteq \frac{\sum_{i=1}^{n} I_{gi} M_i (\bar{y}_i - \bar{\bar{Y}}_g)}{n\left(\frac{M_{\cdot g}}{N_{\cdot\cdot}}\right)} \quad , \qquad (2.4.3)$$

then

$$\mathrm{Var}(\hat{\bar{y}}_g) \doteq \frac{(N_{..} - n)}{n N_{..}} \frac{N_{..}}{(M_{.g})^2} \frac{\sum_{i=1}^{N_{..}} I_{gi}^2 M_i^2 (\bar{Y}_i - \bar{\bar{Y}}_g)^2}{(N_{..} - 1)} + \frac{N_{..}}{n M_{.g}^2} \sum_{i=1}^{N_{..}} I_{gi}^2 M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{\sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2}{(M_i - 1)} \quad .$$

This is the standard form of the approximate variance of a ratio estimator for a two-stage sampline design where the base units have eaual probabilities of selection. Here, the first term represents the between base unit component of the variance, whereas the second denotes the within base unit contribution.

Since our two stage sampling design requires the independent selection of subsamples from different sample base units, and the respective strata estimators are defined in terms of the indicator variables I_{gi}, it can be shown that $\mathrm{Cov}(\hat{\bar{y}}_g, \hat{\bar{y}}_{g'}) \doteq 0$. Hence, the mean squared error of our small area estimator can be expressed as:

$$\mathrm{MSE}(\hat{\bar{y}}_{\ell.}) \doteq \sum_{g=1}^{K} \left(\frac{M_{\ell g}}{M_{.g}}\right)^2 \mathrm{Var}(\hat{\bar{y}}_g) + (\mathrm{Bias})^2 \quad . \qquad (2.4.5)$$

3. A REFORMULATION OF THE KALSBEEK MODEL; SOME ANALYTIC AND EMPIRICAL INVESTIGATIONS

3.1. Introduction

An analytical expression for the mean squared error of our local area estimator has been derived in the previous chapter. Yet, the inherent bias of the model does not allow for tests of its precision unless another unbiased estimate or the true value of the criterion variable is obtained at the local level. In practice, this is usually unavailable and is the reason that alternative strategies must be considered.

In order to determine the accuracy of the small area estimator and allow for comparisons of precision with respect to other strategies, we attempt to express the relationship between criterion and symptomatic variables by means of a probabilistic model. The model enables one to determine the true value of the criterion variable for target areas of interest and to approximate the bias and mean squared error of the respective local area estimators, and provides a framework for comparisons.

3.2 Determination of Stratum Boundaries

As noted, our small area estimator of the criterion variable for the ℓ^{th} local area using the Kalsbeek model takes the form:

$$\hat{\bar{\bar{y}}}_{\ell\cdot} = \sum_{g=1}^{K} \frac{M_{\ell g}}{M_{\ell\cdot}} \hat{\bar{\bar{y}}}_g \quad . \tag{3.2.1}$$

To avoid unnecessary complications which would occur with the multistage sampling design, we consider the single stage cluster sample design, adding the restriction that all target base units consist of the same number of elements. As described in the first chapter, strata (groups) are to be formed which are optimally homogeneous within, while simultaneously dissimilar between themselves. When the underlying relationship between the criterion and symptomatic variables is unknown, the strategy that has been entertained consists of forming groups by minimizing their within sum of squares while maximizing their between sum of squares using only the sample data. However, when a certain probabilistic model is entertained, one could determine those boundaries of the predictor variables which minimize the mean square error of $\bar{\bar{y}}_{\ell\cdot}$. Since each local area estimator usually consists of a different weighted linear combination of the respective stratum estimators, the boundaries which are optimal for small area 1 would not necessarily be so for small area ℓ'. Consequently, another reasonable strategy would be to determine the optimal strata boundaries on the symptomatic variables which minimize the mean squared error of the criterion variable estimator for the over-all population. This estimator is actually the weighted average of all small area estimators, weighted by the respective proportion of elements belonging to the particular small area. As before,

$$\hat{\bar{\bar{y}}}_g = \frac{\sum_{i=1}^{n} I_{gi} M_i \bar{y}_i}{\sum_{i=1}^{n} I_{gi} M_i} \tag{3.2.2}$$

where $M_i = M$ for i = 1, 2, ... $N_{\cdot\cdot}$

and because we are now considering a single stage cluster design,

$$\bar{y}_i = \frac{\sum_{j=1}^{M} y_{ij}}{M} = \bar{Y}_i \quad ,$$

and therefore,

$$\hat{\bar{\bar{y}}}_g = \frac{M \sum_{i=1}^{n} I_{gi}\, \bar{y}_i}{M \sum_{i=1}^{n} I_{gi}} = \frac{\sum^{n_g} \bar{y}_i}{n_g} \quad . \tag{3.2.3}$$

Consequently,

$$\begin{aligned} \hat{\bar{\bar{y}}}_{..} &= \sum_{\ell=1}^{L} \frac{M_{\ell\cdot}}{M_{..}} \hat{\bar{\bar{y}}}_{\ell\cdot} = \sum_{\ell=1}^{L} \frac{M_{\ell\cdot}}{M_{..}} \sum_{g=1}^{K} \frac{M_{\ell g}}{M_{\ell\cdot}} \hat{\bar{\bar{y}}}_g \\ &= \sum_{\ell=1}^{L} \sum_{g=1}^{K} \frac{M_{\ell g}}{M_{..}} \hat{\bar{\bar{y}}}_g = \sum_{g=1}^{K} \frac{M_{\cdot g}}{M_{..}} \hat{\bar{\bar{y}}}_g \\ &= \sum_{g=1}^{K} \frac{M N_{\cdot g}}{M N_{..}} \hat{\bar{\bar{y}}}_g = \sum_{g=1}^{K} \frac{N_{\cdot g}}{N_{..}} \hat{\bar{\bar{y}}}_g \quad , \end{aligned} \tag{3.2.4}$$

since $M_{..} = N_{..} M$ and $M_{\cdot g} = N_{\cdot g} M$.

We also note that this linear combination of local area estimators is an approximately unbiased estimator of the criterion variable for the overall population.

Since the estimator is approximately unbiased, our mean squared error term is actually the variance of the overall population estimator. We must determine the boundaries on the symptomatic variables which will minimize $\mathrm{Var}(\hat{\bar{\bar{y}}}_{..})$. Here, we are faced with the additional problem of working with a linear combination of poststratified estimators. For any fixed sample size n out of $N_{..}$ base units, the n_g g = 1, 2, ... K (K fixed) are random, subject only to the restriction $\sum_{g=1}^{K} n_g = n$. Because the variance of a poststratified estimator is most similar to that of a stratified estimator with proportional allocation, it would be reasonable to use those boundaries on the symptomatic variables which are optimal here. The strategy is most appropriate when $\sum_{g=1}^{K} n_g / K$ is reasonably large, since the poststratified estimator's variance approaches that of the stratified estimator's variance (considering proportional allocation) when this occurs.

Dalenius (1957) and Singh and Sukhatme (1972) have considered the case of minimum variance stratification when a single auxiliary variable was used as the stratification variable. They showed that for a particular allocation (i.e., Neyman, proportional) the boundaries on the auxiliary variable must satisfy a set of minimal equations. Since these equations are ill adapted to practical computation, a quick approximate method has been developed by Dalenius and Hodges (1959)) known as the CUM$\sqrt{f}$ rule, and has been shown to be quite efficient. Thomsen (1975) has found that by taking equal intervals using the CUM$^3\sqrt{f}$ rule, approximately optimum stratum boundaries are determined which compare favorably with those derived by the CUM$\sqrt{f}$ rule.

Often, the stratification scheme will depend on more than one variable. Here as well, several methods have been developed which consider the problem of determining those stratum boundaries which are optimal in the sense of minimum variance stratification.

Anderson (1976) suggests a method which uses the CUM$\sqrt{f}$ rule (or CUM$^3\sqrt{f}$ rule) along each marginal stratifier such that the product of the number of strata for each variable equals:

$$K\left(\prod_{i=1}^{P} K_i = K\right).$$

The method is not optimal, but is practical. It has been shown to yield estimators that are more precise than when only one strong stratifier is used. Another practical method, suggested by Kalsbeek (1973)) allows for the determination of boundaries at successive stages of stratification. Approximately optimum boundaries are obtained for the most significant stratifier, then for the second, conditioned on the stratum means of the first, and so forth until all the stratification variables have been included. In the research that follows. both the methods advanced by Anderson and Kalsbeek are considered.

3.3. A Reformulation of the Kalsbeek Model

We wish to consider the case of sampling from populations with specified continuous multivariate distributions. To use such an approach requires rather strong underlying assumptions regarding the nature of relationships between the criterion and symptomatic variables. To be consistent in getting the finite population results to conform to the new scheme, we disregard the finite population correction factors. Since we have initially considered a single stage cluster sampling design with the restriction that all target base units consist of the same number of elements, our small area estimator is expressed as

$$\hat{\bar{y}}_{\ell\cdot} = \sum_{g=1}^{K} \frac{M_g}{M_{\ell\cdot}} \hat{\bar{y}}_g \tag{3.3.1}$$

$$= \sum_{g=1}^{K} \frac{N_{\ell g}}{N_{\ell\cdot}} \frac{M}{M} \hat{\bar{y}}_g = \sum_{g=1}^{K} \frac{N_{\ell g}}{N_{\ell\cdot}} \hat{\bar{y}}_g \tag{3.3.2}$$

where $\frac{N_{\ell g}}{N_{\ell .}}$ = $\frac{\text{\# of target base units falling in } g^{th} \text{ strata for } \ell^{th} \text{ local area}}{\text{Total \# of target base units in the } \ell\text{th local area}}$

and

$$\hat{\bar{\bar{y}}}_g = \frac{M \sum_{i=1}^{n} I_{gi} \bar{y}_i}{M \sum_{i=1}^{n} I_{gi}} = \frac{\sum^{n_g} \bar{y}_i}{n_g} , \tag{3.3.3}$$

where n_g (the number of sample base units falling in the g^{th} stratum) is random. Consequently,

$$E(\hat{\bar{\bar{y}}}_{\ell .}) = \sum_{g=1}^{K} \frac{N_{\ell g}}{N_{\ell .}} E(\hat{\bar{\bar{y}}}_g) \tag{3.3.4}$$

where

$$E(\hat{\bar{\bar{y}}}_g) = E(\sum^{n_g} \frac{\bar{y}_i}{n_g})$$

and, if we assume $n_g \neq 0$ for g = 1, 2, ... K,

$$E(\hat{\bar{\bar{y}}}_g) = E\Big|_{\substack{g \text{ strata} \\ n_g \text{ fixed}}}^{th} (\bar{y}_i) \quad . \tag{3.3.5}$$

Similarly, we have shown

$$\text{Var}(\hat{\bar{\bar{y}}}_{\ell .}) \doteq \sum_{g=1}^{K} \left(\frac{N_{\ell g}}{N_{\ell .}}\right)^2 \text{Var}(\hat{\bar{\bar{y}}}_g) \tag{3.3.6}$$

where

$$(\bar{\bar{y}}_g) \doteq \left[\frac{1}{n W_g} + \frac{1 - W_g}{n^2 W_g^2}\right] \text{Var}\Big|_{\substack{g^{th} \text{ stratum} \\ n_g \text{ fixed}}} (\bar{y}_i) \tag{3.3.7}$$

with W_g, the respective stratum weights, and again assuming $n_g \neq 0$ for g = 1, 2, ..K.

Therefore,

$$\mathrm{Var}(\hat{\bar{y}}_{\ell\cdot}) \doteq \sum_{g=1}^{K} \left(\frac{N_{\ell g}}{N_{\ell\cdot}}\right)^2 \left[\frac{1}{n\,W_g} + \frac{1 - W_g}{n^2\,W_g^2}\right] \mathrm{Var}\Big|_{\substack{g^{th}\ \text{stratum} \\ n_g\ \text{fixed}}} (\bar{y}_i) \quad . \tag{3.3.8}$$

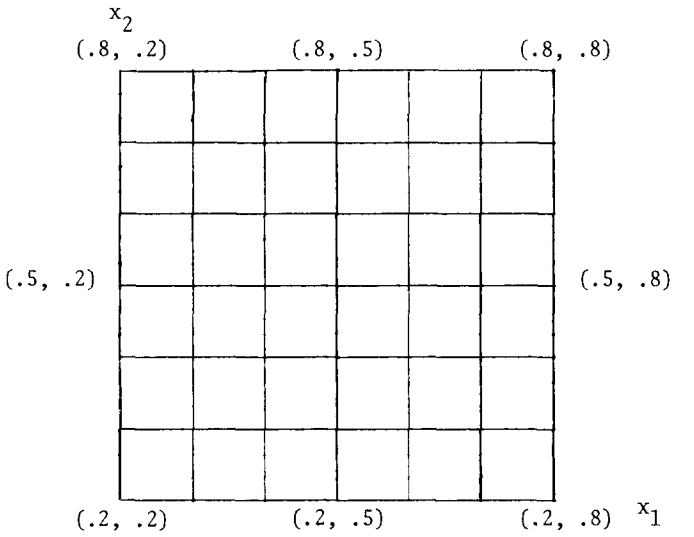
3.4. The Theoretical Framework

Assume a simple random sample of size n is drawn from an infinite p+1 dimensional multivariate population (with continuous distribution) whose observations take the form of the ((p+l)xl) random vector $(y, x_1, x_2, \ldots x_p)'$. Here, the y element conforms to the $\bar{y}_i$ cluster mean, while the $(x_1, x_2, \ldots, x_p)'$ are symptomatic indicators which conform to those for each target base unit. The joint density of the multivariate super population is $f(y, x_1, x_2, \ldots, x_p)$ with marginal probability density functions $f_1(y), f_2(x_1), \ldots, f_{p+1}(x_p)$.

$E_{((p+1)x1)} \begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} \mu_y \\ \underset{\sim}{\mu}_x \end{pmatrix}$ are the respective means of the criterion and symptomatic variables while $\mathrm{Var} \begin{pmatrix} y \\ \underset{\sim}{x} \end{pmatrix} = \begin{bmatrix} \sigma_y^2 & \underset{\sim}{\sigma}'_{yx} \\ \underset{\sim}{\sigma}_{yx} & \underset{\sim}{\sigma}_{x_i x_j} \end{bmatrix} = \underset{(p+1)x(p+1)}{\Sigma}$ is the respective variance covariance matrix assumed to be positive definite.

Once the underlying multivariate distribution has been specified, we are able to construct target areas of interest for fixed values of $N_{\ell\cdot}$. Here, the respective target base units are represented by $N_{\ell\cdot}$ (lxp) vectors of symptomatic information taking the form $(x_{i1}, x_{i2}, \ldots, x_{ip})$. These are determined by taking the values of equally spaced percentiles on the respective marginal distributions of the symptomatic variables over different ranges of interest such that their product is equal to $N_{\ell\cdot}$. To be more explicit, consider the bivariate case with $N_{\ell\cdot} = 49$ and the 20th

to 80th percentile as the range of interest on each marginal stratifier. The values of the equally spaced, cross classified percentiles observed in the following diagram determine the target area's symptomatic information configuration.



With the number of strata (k) fixed, the multivariate stratum boundaries are of form

$$(a_1 < x_1 < b_1,\ a_2 < x_2 < b_2,\ \ldots,\ a_p < x_p < b_p)$$

which are rectilinear, nonoverlapping and exhaustive. Consequently, the expected value of $\bar{y}_i$ for the g^{th} strata (n_g fixed),

is equivalent to

$$E(y \mid a_{1g} < x_1 < b_{1g},\ a_{2g} < x_2 < b_{2g},\ \ldots,\ a_{pg} < x_p < b_{pg}) \quad ,$$

assuming the underlying multivariate distribution. Anderson (1976) has shown that

$$E(y \mid a_{1g} < x_1 < b_{1g},\ \ldots,\ a_{pg} < x_p < b_{pg}) =$$

$$\int_{a_{1g}}^{b_{1g}} \cdots \int_{a_{pg}}^{b_{pg}} \frac{(E(y \mid \underset{\sim}{x})\ g(\underset{\sim}{x})\ d\ \underset{\sim}{x}}{W_g} \qquad (3.4.1)$$

where $E(y|\underset{\sim}{x})$ is the conditional expectation of y given $\underset{\sim}{x}$;

$$g(\underset{\sim}{x}) = \int_{-\infty}^{\infty} f(y, x_1, x_2, \ldots, x_p)\, d\, y$$

is the respective joint density function of the symptomatic variables; and

$$W_g = \int_{a_{1g}}^{b_{1g}} \cdots \int_{a_{pg}}^{b_{pg}} g(\underset{\sim}{x})\, d\, \underset{\sim}{x} = Pr\{a_{1g} < x_1 < b_{1g}, \ldots, a_{pg} < x_p < b_{pg}\} \tag{3.4.3}$$

is the probability of being in the g^{th} strata. Therefore,

$$E(\hat{\bar{y}}_{\ell.}) \doteq \sum_{g=1}^{K} \frac{N_{\ell g}}{N_{\ell.}} E(y|a_{1g} < x_1 < b_{1g}, \ldots, a_{pg} < x_p < b_{pg}) \quad . \tag{3.4.4}$$

Similarly, the variance of $\bar{y}_i$ for the g^{th} strata (n_g fixed),

$$Var(y|a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g}, \ldots, a_{pg} < x_p < b_{pg}) \quad , \tag{3.4.5}$$

for which Anderson has derived the expression

$$\int_{a_{1g}}^{b_{1g}} \cdots \int_{a_{pg}}^{b_{pg}} \frac{Var(y|\underset{\sim}{x})\, g(\underset{\sim}{x}) d(\underset{\sim}{x})}{W_g} + \int_{a_{1g}}^{b_{1g}} \cdots \int_{a_{pg}}^{b_{pg}} \frac{[E(y|\underset{\sim}{x}) - E(y|a_{1g} < x_1 < b_{1g}, \ldots, a_{pg} < x_p < b_{pg})]^2\, g(\underset{\sim}{x})\, d\, \underset{\sim}{x}}{W_g} \tag{3.4.6}$$

where $Var(y|\underset{\sim}{x})$ is the conditional variance of y given $\underset{\sim}{x}$. Consequently, $Var(\hat{\bar{y}}_{\ell.}) \doteq$

$$\sum_{g=1}^{K} \left(\frac{N_{\ell g}}{N_{\ell.}}\right)^2 \left[\frac{1}{n W_g} + \frac{1 - W_g}{n^2 W_g^2}\right] Var(y|a_{1g} < x_1 < b_{1g}, \ldots, a_{pg} < x_p < b_{pg}) \,. \tag{3.4.7}$$

3.4.1. Determination of the Bias

We defined the true value of a criterion variable of interest for local area ℓ

$$\bar{\bar{Y}}_{\ell\cdot} = \sum^{N_{\ell\cdot}} M_i \bar{Y}_i / M_{\ell\cdot} \qquad (3.4.8)$$

for the two stage sampling design. Similarly,

$$\bar{\bar{Y}}_{\ell\cdot} = \frac{\sum^{N_{\ell\cdot}} M \bar{y}_i}{M N_{\ell\cdot}} = \frac{\sum^{N_{\ell\cdot}} \bar{y}_i}{N_{\ell\cdot}} \qquad (3.4.9)$$

for the single stage cluster design with target base units having the same number of elements. In the theoretical framework considered, $\bar{\bar{Y}}_{\ell\cdot}$ has been defined as a function of the vector of symptomatic information, $(x_1, x_2, \ldots, x_p)$, for different target areas of interest. Here,

$$\bar{\bar{Y}}_{\ell\cdot} = \frac{\sum^{N_{\ell\cdot}} E(y|\underset{\sim}{x})}{N_{\ell\cdot}} \qquad (3.4.10)$$

for $x = (x_1, x_2, \ldots, x_p)$ fixed. Consequently, the bias of our poststratified local area estimator $(\hat{\bar{\bar{y}}}_{\ell\cdot})$ can be approximated by:

$$\text{Bias } (\hat{\bar{\bar{y}}}_{\ell\cdot}) \doteq$$

$$\sum_{g=1}^{K} \frac{N_{\ell g}}{N_{\ell\cdot}} E(y|a_{1g} < x_1 < b_{1g}, \ldots, a_{pg} < x_p < b_{pg}) - \frac{\sum^{N_{\ell\cdot}} E(y|\underset{\sim}{x})}{N_{\ell\cdot}} \quad . \qquad (3.4.11)$$

Also, the mean squared error of $\hat{\bar{\bar{y}}}_{\ell\cdot}$ can be approximated by

$$\text{M.S.E.}(\hat{\bar{\bar{y}}}_{\ell\cdot}) \doteq$$

$$\sum_{g=1}^{K} \left(\frac{N_{\ell g}}{N_{\ell\cdot}}\right)^2 \left[\frac{1}{n W_g} + \frac{1-W_g}{n^2 W_g^2}\right] \text{Var}(y|a_{1g} < x_1 < b_{1g}, \ldots, a_{pg} < x_p < b_{pg})$$

$$+ (\text{Bias } (\hat{\bar{\bar{y}}}_{\ell\cdot}))^2 \quad . \qquad (3.4.12)$$

3.5. Estimation Using the Ericksen Model

To allow for a comparison of the method's accuracy, we reconsider the Ericksen model which is applicable in the same general setting. Here, the least squares regression estimator is determined using data obtained from the sample base units. Estimates of the criterion variable for the respective target area base units are then derived by substituting their vectors of symptomatic information into the resulting equation. The model of Ericksen is represented by:

$$\hat{\bar{y}}_{\ell\cdot(E)} = \frac{\sum^{N_{\ell\cdot}} \underset{\sim}{X}'_{(E)} \underset{\sim}{\hat{B}}_{(E)}}{N_{\ell\cdot}} = \hat{B}_0 + \bar{X}_{\ell 1}\hat{B}_1 + \bar{X}_{\ell 2}\hat{B}_2 + \ldots + \bar{X}_{\ell p}\hat{B}_p \qquad (3.5.1)$$

where $\underset{\sim}{X}'_{(E)} = (1, x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip})$ is a (1x(p+1)) vector of symptomatic information from the i^{th} base unit in the ℓ^{th} target area;

$\underset{\sim}{\hat{B}}_{(E)} = \begin{bmatrix} \hat{B}_0 \\ \hat{B}_1 \\ \cdot \\ \cdot \\ \cdot \\ \hat{B}_p \end{bmatrix}$ is the ((p+1)x1) vector of the least squares regression coefficients determined by the criterion and symptomatic variable information for the n sample base units;

$N_{\ell\cdot}$ is the number of base units in the ℓ^{th} target area;

and $\bar{X}_{\ell s}$ s = 1, 2, ... p is the s^{th} symptomatic variable's mean for the ℓ^{th} target area.

3.6 Distribution Specific Results

To give our findings a degree of validity beyond the scope of the theoretical framework, the relationship between criterion and symptomatic variables must be characterized by those distributions most relevant to the practical setting. Since the vector $(y, x_1, x_2, \ldots, x_p)$ of criterion and symptomatic variables has been defined to represent a vector of cluster means, their distributions approach the normal when the underlying distributions are not markedly skewed. Consequently, the first distribution we have chosen to consider is the multivariate normal. To facilitate the presentation, we examine the trivariate case where the random vector $\underset{\sim}{v}' = (y, x_1, x_2)$ has a three dimensional multivariate normal distribution with joint density function

$$f(\underset{\sim}{v}) = \frac{\exp[-\frac{1}{2}(\underset{\sim}{v} - \underset{\sim}{\mu_v})' \underset{\sim}{\Sigma_v}^{-1}(\underset{\sim}{v} - \underset{\sim}{\mu_v})]}{(2\pi)^{3/2} |\underset{\sim}{\Sigma_v}|^{1/2}} \qquad (3.6.1)$$

$$-\infty < y, x_1, x_2 < \infty$$

with

$$E(\underset{\sim}{v}) = \underset{\sim}{\mu_v} = \begin{bmatrix} \mu_y \\ \mu_{x_1} \\ \mu_{x_2} \end{bmatrix}$$

and

$$\underset{\sim}{\Sigma_v} = \begin{bmatrix} \sigma_y^2 & \sigma_{yx_1} & \sigma_{yx_2} \\ \sigma_{yx_1} & \sigma_{x_1}^2 & \sigma_{x_1x_2} \\ \sigma_{yx_2} & \sigma_{x_1x_2} & \sigma_{x_2}^2 \end{bmatrix}$$

(assumed to be positive definite).

Another continuous distribution of major interest to our research is the multivariate logistic distribution. The logistic curve has long been a valuable tool to demographers as a model for estimating population growth in designated geographical areas. Also, the marginal distributions of the multivariate logistic are quite similar to the normal. More importantly, since its curve of regression is nonlinear in x, we have a setting for which the Ericksen estimator is biased. As before, we shall consider the trivariate case where the random vector $\underset{\sim}{v}' = (v_1, v_2, v_3) = (y_1, x_1, x_2)$ has the density function described by Gumbel (1961),

$$f(\underset{\sim}{v}) = \frac{3![1 + \sum_{i=1}^{3} \exp\{-(v_i - \mu_{v_i})/\zeta_{v_i}\}]^{-4} \exp[-\sum_{i=1}^{3} (v_i - \mu_{v_i})/\zeta_{v_i}\}}{\prod_{i=1}^{3} \zeta_{v_i}}$$

$$-\infty < \underset{\sim}{v} < \infty \qquad (3.6.2)$$

and $\zeta_{V_i} = \sigma_{V_i} / (\pi/\sqrt{3})$ such that the cumulative distribution function of V_i is

$$F_{V_i}(v_i) = [1 + \exp\{\frac{-(v_i - \mu_{V_i})}{\zeta_{V_i}}\}]^{-1} \quad . \tag{3.6.3}$$

To determine the accuracy of our poststratified target area estimator and compare its precision with respect to the Ericksen model, the following settings are specified:

(1) Underlying Trivariate Normal Distribution with a high association level ($R \doteq .95$)

$$E\begin{bmatrix} y \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 60 \\ 50 \\ 50 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} 100 & 42.5 & 42.5 \\ 42.5 & 25 & 15 \\ 42.5 & 15 & 25 \end{bmatrix}$$

(2) Underlying Trivariate Normal Distribution with a low association level ($R \doteq .58$)

$$E\begin{bmatrix} y \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 60 \\ 50 \\ 50 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} 100 & 25 & 25 \\ 25 & 25 & 12.5 \\ 25 & 12.5 & 25 \end{bmatrix}$$

(3) Underlying Trivariate Logistic Distribution with level of association corresponding to ($R \doteq .58$)

$$E\begin{bmatrix} y \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 60 \\ 50 \\ 50 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} 100 & 25 & 25 \\ 25 & 25 & 12.5 \\ 25 & 12.5 & 25 \end{bmatrix} \quad .$$

The target areas we consider consist of $N_\ell = 49$ target base units whose representation is given in Section 3.4 with (.2, .8), (.05, .95) and (.35, .95) as the ranges of interest. These two values of n, the number of base units in the design, are given: n=120, n=480. For these settings, large sample approximations are used when necessary to derive the expectation, variance, and bias for the Ericksen (linear) estimator.

The number of strata we consider varies as $K = b^2$, where b = 2, 3, 4 is the number of boundaries on each marginal stratifier. When the underlying

distribution is trivariate normal, two alternative strategies are used in the determination of the stratum boundaries. The first method is attributed to Anderson (1976) whereby marginally optimum stratum boundaries for proportional allocation are selected. These are given by Sethi for the standardized normal variate:

b	Optimum Boundaries
2	0.0
3	-.61, .61
4	-.99, 0.0, .99

The other, attributed to Kalsbeek, is a hierarchical scheme described in Section 3.2. The respective stratum boundaries are shown in figure (3.6.1) for the standardized normal variates when $\rho_{x_1x_2} = .6$. The same boundaries are used for $\rho_{x_1x_2} = .5$ to improve the target area estimator's precision.

When the underlying distribution is trivariate logistic, we use Anderson's approach with the CUM$\sqrt{f}$ rule. To implement this procedure on each marginal stratifier, we consider the theoretically infinite population to be finite and of size 10,000. Selecting the 0.5^{th} and 99.5^{th} percentiles as endpoints, we construct 100 equally spaced intervals on the range of the distribution, determine their respective frequencies, and apply the CUM$\sqrt{f}$ rule. Here,

b	Stratum Boundaries Using CUM$\sqrt{f}$ Rule
2	0.0
3	-0.99 0.99
4	-1.57 0.0 1.57

We knew a priori that the Ericksen model was most appropriate to the linear setting, being an unbiased target area estimator when the underlying continuous distribution is multivariate normal. This is confirmed in the high association model under study ($R \doteq .95$). When the level of association is seriously reduced ($R \doteq .58$). the supoerioritv of the linear estimator is nowhere as clear. At the same time, we note gains in precision for the poststratified estimator when the hierarchical scheme of stratum boundary determination is employed. This is reflected in both the variance and mean squared error terms. Similarly, we note gains in precision for both estimators with an increase in sample size. Consequently, when the sample size is large and a hierarchical scheme is employed, the poststratified estimator does reasonably well for the linear setting.

When attention is directed to the nonlinear setting of the trivariate logistic distribution, the merits of the proposed approach become more obvious. As before, we also note gains in precision for both estimators as reflected in the variance and mean equared error terms with increased sample size. Here, the inherent bias in the linear estimator generally dominates that of the poststratified estimator. This is primarily a function of the lack of fit of the Ericksen model. Had we considered a trivariate setting with an even more striking nonlinear curve of regression, the relative bias of the linear estimator would be greater. For each target area under consideration, there is at least one stratification

FIGURE (3.6.1)

Stratum Boundaries Using Hierarchical Scheme

b=2

$X_1 \leq .0$ | $X_2 \leq -.479$ | $X_2 > -.479$

(Stratum Mean = -.798)

$X_1 > 0.0$ | $X_2 \leq .479$ | $X_2 > .479$

(S.M. = .798)

b=3

$X_1 \leq -.61$ | $X_2 \leq -1.222$ | $-1.22 < X_2 \leq -.246$ | $X_2 > -.246$

(S.M. = -1.223)

$-.61 < X_1 \leq .61$ | $X_2 \leq -.488$ | $-.488 < X_2 \leq .488$ | $X_2 > .488$

(S.M. = 0.0)

$X_1 > .61$ | $X_2 \leq .246$ | $.246 < X_2 \leq 1.222$ | $X_2 > 1.222$

(S.M. = 1.223)

b=4

$X_1 \leq -.99$ | $X_2 \leq -1.702$ | $-1.702 < X_2 \leq -.910$ | $-.910 < X_2 \leq -.118$ | $X_2 > -.118$

(S.M. = -1517)

$-.99 < X_1 \leq 0.0$ | $X_2 \leq -1.066$ | $-1.066 < X_2 \leq -.274$ | $-.274 < X_2 \leq .518$ | $X_2 > .518$

(S.M. = -.456)

$.0 < X_1 \leq .99$ | $X_2 \leq -.518$ | $-.518 < X_2 \leq .274$ | $.274 < X_2 \leq 1.066$ | $X_2 > 1.066$

(S.M. = .456)

$X_1 > .99$ | $X_2 \leq .118$ | $.118 < X_2 \leq .910$ | $.910 < X_2 \leq 1.702$ | $X_2 > 1.702$

(S.M. = 1.517)

scheme for n=120, and at least two for n=480, which demonstrate the post-stratified estimators' superiority using the mean squared error as the measure of precision. Had a more optimal scheme for the determination of strata boundaries been available, further increases in the precision of our poststratified estimator could have been observed.

Generally, when stratification is the strategy used to yield an estimator of a criteriaon variable for a particular target population, an increase in the number of strata, K, is followed by an increase in the estimator's precision (as measured by a decrease in the variance) for relatively small values of K. Subsequent increases in K coincide with diminishing returns with respect to further proportional reductions in the estimator's variance. Since each target area estimator under consideration consists of a different weighted linear combination of stratum estimators, and the sampled population does not completely coincide with the target population, we do not expect to find strong evidence of a consistent relationship between the proposed method's precision and the number of strata to be specified (see tables 3.1 - 3.9).

4. SUMMARY

To summarize, reliable estimates of parameters at the local level are difficult, if not impossible, to obtain directly from sample surveys, primarily due to the constraints of sample size and design. Yet, the very nature of the problem has served as the motivating force in the development of several alternative procedures. When underlying assumptions are too strict or unrealistic, the need for a more flexible approach is obvious. The method considered in our research is particularly attractive in that no functional model between criterion and symptomatic variables must be specified. Here, the most limiting assumption is the availability of symptomatic information. Estimates for the respective "base units" of "target areas" are available as a byproduct of the technique. Finally, the method performs reasonably well even for the linear setting, though here it would be better to choose Ericksen's approach.

TABLE 3.1

TARGET AREA ESTIMATION FOR TRIVARIATE NORMAL DISTRIBUTION WITH R ≐ .95

Stratification Scheme	Model	Strata (n=120)	Range (.2, .8) Approximate Values for Criterion Parameters				True Value of Criterion Variable
			Expectation	Variance	Bias	M.S.E.	
Optimal Boundaries on Marginals	Ericksen Modified Kalsbeek Model	4 9 16	60.000 58.625 60.000 59.288	0.081 0.292 0.228 0.263	0.000 -1.375 0.000 -0.712	0.081 2.181 0.228 0.770	60.000 60.000 60.000 60.000
Hierarchical	Ericksen Modified Kalsbeek Model	4 9 16	60.000 59.316 60.000 59.773	0.081 0.235 0.211 0.255	0.000 -0.684 0.000 -0.227	0.081 0.753 0.211 0.306	60.000 60.000 60.000 60.000
		(n=480)					
Optimal Boundaries on Marginals	Ericksen Modified Kalsbeek Model	4 9 16	60.000 58.625 60.000 59.288	0.020 0.071 0.055 0.063	0.000 -1.375 0.000 -0.712	0.020 1.961 0.055 0.570	60.000 60.000 60.000 60.000
Hierarchical	Ericksen Modified Kalsbeek Model	4 9 16	60.000 59.316 60.000 59.773	0.020 0.067 0.050 0.059	0.000 -0.684 0.000 -0.227	0.020 0.538 0.050 0.111	60.000 60.000 60.000 60.000

TABLE 3.2

TARGET AREA ESTIMATION FOR TRIVARIATE NORMAL DISTRIBUTION WITH R= .58

Stratification Scheme	Model	Strata (n=120)	Range (.2, .8) Approximate Values for Criterion Parameters Expectation	Variance	Bias	M.S.E.	True Value of Criterion Variable
Optimal Boundaries	Ericksen		60.000	0.556	0.000	0.556	60.000
	Modified	4	59.145	0.731	-0.855	1.462	60.000
	Kalsbeek	9	60.000	0.925	0.000	0.925	60.000
	Model	16	59.554	1.292	-0.446	1.490	60.000
Hierarchical	Ericksen		60.000	0.556	0.000	0.556	60.000
	Modified	4	59.580	0.720	-0.420	0.907	60.000
	Kalsbeek	9	60.000	0.813	0.000	0.813	60.000
	Model	16	59.862	1.162	-0.138	1.181	60.000
		(n=480)					
Optimal Boundaries on Marginals	Ericksen		60.000	0.139	0.000	0.139	60.000
	Modified	4	59.145	0.178	-0.855	0.909	60.000
	Kalsbeek	9	60.000	0.224	0.000	0.224	60.000
	Model	16	59.554	0.309	-0.446	0.507	60.000
Hierarchical	Ericksen		60.000	0.139	0.000	0.139	60.000
	Modified	4	59.580	0.180	-0.420	0.356	60.000
	Kalsbeek	9	60.000	0.196	0.000	0.196	60.000
	Model	16	59.862	0.274	-0.138	0.293	60.000

TABLE 3.3

TARGET AREA ESTIMATION FOR TRIVARIATE LOGISTIC DISTRIBUTION (CORRESPONDING TO R ≐ .58)

Stratification Scheme	Model	Strata (n=120)	Range (.2, .8) Approximate Values for Criterion Parameters Expectation	Variance	Bias	M.S.E.	True Value of Criterion Variable
Approximate Optimal Boundaries on Marginals Using Cum$\sqrt{f}$ Rule	Ericksen Modified Kalsbeek Model	4 9 16	60.000 59.334 60.956 61.108	0.517 0.763 0.869 1.276	-1.296 -1.962 -0.340 -0.188	2.195 4.612 0.984 1.311	61.296 61.296 61.296 61.296
		(n=480)					
	Ericksen Modified Kalsbeek Model	4 9 16	60.000 59.334 60.956 61.108	0.129 0.186 0.210 0.304	-1.296 -1.962 -0.340 -0.188	1.808 4.036 0.325 0.340	61.296 61.296 61.296 61.296

TABLE 3.4

TARGET AREA ESTIMATION FOR TRIVARIATE NORMAL DISTRIBUTION WITH $R \doteq .95$

Stratification Scheme	Model	Strata (n=120)	Range (.05, .95) Approximate Values for Criterion Parameters: Expectation	Variance	Bias	M.S.E.	True Value of Criterion Variable
Optimal Boundaries on Marginals	Ericksen		60.000	0.081	0.000	0.081	60.000
	Modified	4	58.625	0.292	-1.375	2.181	60.000
	Kalsbeek	9	60.000	0.311	0.000	0.311	60.000
	Model	16	59.278	0.476	-0.722	0.997	60.000
Hierarchical	Ericksen		60.000	0.081	0.000	0.081	60.000
	Modified	4	59.316	0.235	-0.684	0.753	60.000
	Kalsbeek	9	60.000	0.215	0.000	0.215	60.000
	Model	16	59.588	0.218	-0.412	0.388	60.000
		(n=480)					
Optimal Boundaries on Marginals	Ericksen		60.000	0.020	0.000	0.020	60.000
	Modified	4	58.625	0.071	-1.375	1.960	60.000
	Kalsbeek	9	60.000	0.065	0.000	0.065	60.000
	Model	16	59.278	0.069	-0.722	0.590	60.000
Hierarchical	Ericksen		60.000	0.020	0.000	0.020	60.000
	Modified	4	59.316	0.070	-0.684	0.538	60.000
	Kalsbeek	9	60.000	0.051	0.000	0.051	60.000
	Model	16	59.588	0.048	-0.412	0.218	60.000

TABLE 3.5

TARGET AREA ESTIMATION FOR TRIVARIATE NORMAL DISTRIBUTION WITH $R \doteq .58$

Stratification Scheme	Model	Strata (n=120)	Range (.05, .95) Approximate Values for Criterion Parameters Expectation	Variance	Bias	M.S.E.	True Value of Criterion Variable
Optimal Boundaries on Marginals	Ericksen		60.000	0.556	0.000	0.556	60.000
	Modified	4	59.145	0.731	-0.855	1.461	60.000
	Kalsbeek	9	60.000	0.026	0.000	0.926	60.000
	Model	16	59.548	1.118	-0.452	1.322	60.000
Hierarchical	Ericksen		60.000	0.556	0.000	0.556	60.000
	Modified	4	59.580	0.731	-0.420	0.907	60.000
	Kalsbeek	9	60.000	0.722	0.000	0.722	60.000
	Model	16	59.745	0.818	-0.255	0.883	60.000
		(n=480)					
Optimal Boundaries on Marginals	Ericksen		60.000	0.139	0.000	0.139	60.000
	Modified	4	59.145	0.178	-0.855	0.909	60.000
	Kalsbeek	9	60.000	0.205	0.000	0.205	60.000
	Model	16	59.548	0.215	-0.452	0.419	60.000
Hierarchical	Ericksen		60.000	0.139	0.000	0.139	60.000
	Modified	4	59.580	0.179	-0.420	0.356	60.000
	Kalsbeek	9	60.000	0.172	0.000	0.172	60.000
	Model	16	59.745	0.187	-0.255	0.252	60.000

TABLE 3.6

TARGET AREA ESTIMATION FOR TRIVARIATE LOGISTIC DISTRIBUTION (CORRESPONDING TO R ± .58)

Range (.05, .95)

Stratification Scheme	Model	Strata (n=120)	Approximate Values for Criterion Parameters				True Value of Criterion Variable
			Expectation	Variance	Bias	M.S.E.	
Approximate Optimal Boundaries on Marginals Using Cum√f Rule	Ericksen		60.000	0.517	+0.835	1.214	59.165
	Modified	4	59.334	0.763	+0.169	0.791	59.165
	Kalsbeek	9	59.925	0.903	+0.760	1.481	59.165
	Model	16	59.796	0.923	+0.631	1.322	59.165
		(n=480)					
	Ericksen		60.000	0.129	+0.835	0.827	59.165
	Modified	4	59.334	0.186	+0.169	0.215	59.165
	Kalsbeek	9	59.925	0.201	+0.760	0.779	59.165
	Model	16	59.796	0.191	+0.631	0.590	59.165

TABLE 3.7

TARGET AREA ESTIMATION FOR TRIVARIATE NORMAL DISTRIBUTION WITH $R \doteq .95$

Range (.35, .95)

Stratification Scheme	Model	Strata (n=120)	Approximate Values for Criterion Parameters				True Value of Criterion Variable
			Expectation	Variance	Bias	M.S.E.	
Optimal Boundaries on Marginals	Ericksen		65.094	0.104	0.000	0.104	65.094
	Modified	4	64.124	0.366	-0.970	1.306	65.094
	Kalsbeek	9	65.751	0.307	0.657	0.739	65.094
	Model	16	65.369	0.274	0.276	0.350	65.094
Hierarchical	Ericksen		65.094	0.104	0.000	0.104	65.094
	Modified	4	63.799	0.340	-1.294	2.015	65.094
	Kalsbeek	9	65.102	0.286	0.008	0.286	65.094
	Model	16	64.962	0.246	-0.132	0.264	65.094
		(n=480)					
Optimal Boundaries on Marginals	Ericksen		65.094	0.026	0.000	0.026	65.094
	Modified	4	64.124	0.090	-0.970	1.030	65.094
	Kalsbeek	9	65.751	0.074	0.657	0.506	65.094
	Model	16	65.369	0.060	0.276	0.136	65.094
Hierarchical	Ericksen		65.094	0.026	0.000	0.026	65.094
	Modified	4	63.799	0.083	-1.294	1.759	65.094
	Kalsbeek	9	65.102	0.068	0.008	0.068	65.094
	Model	16	64.962	0.056	-0.132	0.073	65.094

TABLE 3.8

TARGET AREA ESTIMATION FOR TRIVARIATE NORMAL DISTRIBUTION WITH R ≐ .58

Range (.35, .95)

Stratification Scheme	Model	Strata (n=120)	Approximate Values for Criterion Parameters				True Value of Criterion Variable
			Expectation	Variance	Bias	M.S.E.	
Optimal Boundaries on Marginals	Ericksen		63.196	0.726	0.000	0.726	63.196
	Modified	4	62.565	0.843	-0.631	1.242	63.196
	Kalsbeek	9	63.622	1.103	0.426	1.284	63.196
	Model	16	63.385	1.069	0.189	1.104	63.196
Hierarchical	Ericksen		63.196	0.726	0.000	0.726	63.196
	Modified	4	62.350	0.850	-0.846	1.566	63.196
	Kalsbeek	9	63.202	1.072	0.006	1.072	63.196
	Model	16	63.130	1.071	-0.066	1.075	63.196
		(n=480)					
Optimal Boundaries on Marginals	Ericksen		63.196	0.182	0.000	0.182	63.196
	Modified	4	62.565	0.207	-0.631	0.606	63.196
	Kalsbeek	6	63.622	0.265	0.426	0.446	63.196
	Model	16	63.385	0.241	0.189	0.277	63.196
Hierarchical	Ericksen		63.196	0.182	0.000	0.182	63.196
	Modified	4	62.350	0.209	-0.846	0.925	63.196
	Kalsbeek	9	63.202	0.257	0.006	0.257	63.196
	Model	16	63.130	0.247	-0.066	0.251	63.196

TABLE 3.9

TARGET AREA ESTIMATION FOR TRIVARIATE
LOGISTIC DISTRIBUTION (CORRESPONDING TO R ≐ .58)

Stratification Scheme	Model	Strata (n=120)	Range (.35, .95) Approximate Values for Criterion Parameters: Expectation	Variance	Bias	....	True Value of Criterion Variable
Approximate Optimal Boundaries on Marginals Using Cum$\sqrt{f}$ Rule	Ericksen Modified Kalsbeek Model	4 9 16	63.035 62.530 63.865 63.340	0.659 0.742 0.924 0.856	-0.680 -1.184 0.151 -0.314	1.122 2.144 0.947 0.955	63.714 63.714 63.714 63.714
		(n=480)					
	Ericksen Modified Kalsbeek Model	4 9 16	63.034 62.530 63.865 63.340	0.165 0.182 0.223 0.198	-0.680 -1.184 0.151 -0.314	0.627 1.584 0.246 0.296	63.714 63.714 63.714 63.714

BIBLIOGRAPHY

Anderson, D.W. (1976). Gains from multivariate stratification. Unpublished doctoral dissertation. University of Michigan, Ann Arbor.

Anderson, D.W., Kish, L., and Cornell, R.G. (1976). Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model. Journal of the American Statistical Association 71, 887-892.

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

Burnham, W.D. and Sprague, J. (1970). Additive and multiplicative models of the voting universe: the case of Pennsylvania, 1960-1968. American Political Science Review 64, 471-490.

Causey, B.C. (1972). Sensitivity of raked contingency table totals to changes in problem conditions. Annals of Mathematical Statistics 43, 656-658.

Cochran, W.G. (1963). Sampling Techniques. New York: John Wiley and Sons.

Cohen, A.C. (1957). On the solution of estimating equations for truncated and censored samples from normal populations. Biometrika 44, 225-236.

Cramer, H. (1963). Mathematical Methods of Statistics. Princeton: Princeton University press.

Crosetti, A.H. and Schmitt, R.C. (1956). A method of estimating the intercensal populations of counties. Journal of the American Statistical Association 51, 587-590.

Curnow, R.N. and Dunnett, G.W. (1962). The numerical evaluation of certain multivariate normal integrals. Annals of Mathematical Statistics 33, 571-579.

Dalenius, T. (1957). Sampling in Sweden. Contributions to the Methods and Theories of Sample Survey Practice. Stockholm: Almquist och Wiksell.

Dalenius, T. and Hodges, J.L. (1959). Minimum variance stratification. Journal of the American Statistical Association 54, 88-101.

Duncan, O.D. (1959). Residential Segregation and Social Differentiation. Vienna, Austria: International Population Conference.

Ericksen, E.P. (1973a). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. Demography 10, 137-160.

Ericksen, E.P. (1973b). Recent developments in estimation for local areas. American Statistics Association, Proceedings of the Social Statics Section, 37-41.

Ericksen, E.P. (1974). A regression method for estimating population changes of local areas. Journal of the American Statistical Association 69, 867-875.

Gonzalez, M.E. and Waksberg, J. (1973). Estimation of the error of synthetic estimates. Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

Gonzalez, M.E. (1973). Use and evaluation of synthetic estimates. American Statistical Association, Proceedings of the Social Statistics Section, 33-36.

Gonzalez, M.E. and Hoza, C. (1975). Small area estimation of unemployment. American Statistical Association, Proceedings of the Social Statistics Section.

Gumbel, E.J. (1961). Bivariate logistic distributions. Journal of the American Statistical Association 56, 335-349.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Methods and Theory, Vols. I and II. New York: John Wiley an Sons.

Hinckley, B. (1970). Incumbency and the presidential vote in senate elections: defining parameters of subpresidential voting. American Political Science Review 64, 836-842.

Johnson, N.L. and Kotz, S. (1972). Continuous Multivariate Distributions. New York: John Wiley and Sons.

Kaitz, H. (1973). Comments on the papers by Gonzalez and Ericksen. American Statistical Association, Proceeding of the Social Statistics Section, 44-45.

Kalsbeek, W.D. (1973). A method for obtaining local postcensal estimates for several types of variables. Unpublished doctoral dissertation. University of Michigan, Ann Arbor.

Keyfitz, N. (1957). Estimates of sampling variances where two units are selected from each stratum. Journal of the American Statistical Association 52, 503-510.

Kim, J., Petrocik, J.R., and Enokson, S.N. (1975). Voter turnout among American States: systemic and individual components. American Political Science Review 69, 107-123.

Kish, L. (1967). Survey Sampling. New York: John Wiley and Sons.

Koch, G.G. (1973). An alternative approach to multivariate response error models for sample survey data with applications to estimators involving subclass means. Journal of the American Statistical Association 68, 906-913.

Koch, G.G., Freeman, D.H., and Tolley, H.D. (1975). The asymptotic covariance structure of estimated parameters from contingency table loglinear models. University of North Carolina Institute of Statistics Mimeo Series No. 1046.

Koch, G.G. and Freeman, D.H. (1976). The asymptotic covariance structure of estimated parameters from marginal adjustment of contingency tables. Paper presented to the Washington Statistical Society Frederick F. Stephan Memorial Methodology program.

Levy, P.S. (1971). The use of mortality data in evaluating synthetic estimates. American Statistical Association, Proceedings of the Social Statistics Section, 328-331.

Murthy, M.N. (1967). Sampling Theory and Methods. Calcutta: Statistical Publishing Society.

Namekater, T., Levy, P.S. and O'Rourke, T.W. (1975). Synthetic estimates of work loss disability for each state and the District of Columbia. Public Health Reports 90, 532-538.

Pool, I., Akelson, R.P. and Popkin, S.L. (1975). Candidates, Issues and Strategies: A Computer Simulation of the 1960 and 1964 Presidential Elections. Cambridge, Massachusetts: MIT Press.

Rosenbaum, S. (1961). Moments of a truncated bivariate normal distribution. Journal of the Royal Statistical Society--Series B 23, 405-408.

Royall, R. (1973). Discussion of papers by Gonzalez and Ericksen. American Statistical Association, Proceedings of the Social Statistics Section, 42-43.

Scheuren, F.J. and Oh, H.L. (1975). A data analysis approach to fitting square tables. Communications in Statistics 4, 595-615.

Schneider, A.L. (1973). Estimating aggregate opinion in small political units: the disaggregation of national survey data. Applied Policy Research Series, Vol. 1, Oregon Research Institution.

Seidman, D. (1975). Simulation of public opinion: a caveat. Public Opinion Quarterly 39, 43-54.

Sethi, V.K. (1963). A note on optimal stratification of populations for estimating the population means. Australian Journal of Statistics 5, 20-33.

Shryock, H.S. and Siegel, J.S. (1973). The Methods and Material of Demography. Washington D.C.: U.S. Government Printing Office.

Singh, R. and Sukhatme, B.V. (1972). A note on optimum stratification. Journal of the Indian Society of Agricultural Statistics 24, 91-98.

Snow, E.C. (1911). The application of the method of multiple correlation to the estimation of postcensal population. Journal of the Royal Statistical Society 74, 575-620.

Soares, G. and Hamblin, R. (1967). Socioeconomic variables and voting for the radical left: Chile 1952. American Political Science Review 61, 1053-1065.

Stephen, F.F. (1945). The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian variate. Annals of Mathematical Statistics 16, 50-61.

Thomsen, I. (1976). A comparison of approximately optimal stratification given proportional allocation with other methods of stratification and allocation. Metrika 23, 15-25.

U.S. Bureau of the Census. (1963). The Current Population Survey--A Report on Methodology. Technical Paper No. 7.

U.S. Bureau of the Census. (1973). Concepts and methods used in manpower statistics from the Current Population Survey. Current Population Reports, p. 23, No. 22.

U.S. Bureau of the Census. (1973). Count and City Data Book, 1972 (A Statistical Abstract Supplement). Washington D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census. (1973). Federal state cooperative program for local population estimates: test results--April 1, 1970. Current Population Reports, P-26, No. 21.

U.S. Bureau of the Census. (1975). Coverage of the population in the 1970 census and some implications for public programs. Current Population Reports, P-23, No. 56.

U.S. National Center for Health Statistics. (1968). Synthetic Estimates of Disability. PHS publication, No. 1759.

Weber, R.E. and Shaffer, W.R. (1972). Public opinion and American state policy making. Midwest Journal of Political Science 16, 683-699.

Weber, R.E., Hopkins, A.H., Mezey, M.L., and Munger, F. J. (1973). Computer simulation of state electorates. Public Opinion Quarterly 36, 549-565.

Woodruff, R.S. (1966). Use of a regression technique to produce area breakdowns of the monthly estimates of retail trade. Journal of the American Statistical Association 61, 496-504.

Discussion

Joseph Waksberg

1. I'm not sure that I see where the Kalsbeek-Cohen model is really different from the synthetic estimator model that Simmons, Levy and others at NCHS have described, or that Maria Gonzalez and I discussed in our 1973 paper. The synthetic estimate is defined as $\Sigma \hat{p}_{1i} \bar{x}_i$ where the i is an index for the classification of the population considered most useful for the statistic to be estimated. Most, or possibly all, of the examples discussed in the earlier papers have considered the commonly-used classification variables such as sex, age, race, etc. However, there is no theoretical reason why some type of small area geographic classification cannot be used to define the classes, either solely or in combination with the more usual demographic items. If this is done, then the Kalsbeek-Cohen model merges with the earlier one.

Some of the earlier papers do include geography as a component of the classification scheme, but use fairly large areas; for example, SMSA's versus non-SMSA's, or county size. These are areas that generally correspond to primary sampling units for most of the large-scale national surveys whose results have been used for synthetic estimates. They are easily manipulable since the data can be automatically coded. More important, they comprise classifications for which reasonably accurate data are likely to exist on the population proportions that act as weights for the local area estimates. This is, of course, essential for the theory to have any practical application. Cohen's paper departs from the large area geographic units and shows that smaller areas can also be used.

I have tried to develop criteria for the kinds of areas that could be efficiently utilized in real-life applications of Cohen's approach. It seems to me that there are three conditions that have to be satisfied in defining the areas:

a. The areas must be such that each sample element can be coded in its proper base unit, so that it is clear to which of the G classes it belongs;

b. Current population counts of the number of elements in each of the G classes are necessary so that the appropriate weights can be used in the estimators;

c. The areas must be fairly small so that the population within each area is relatively homogeneous. This is necessary for the poststratification in the estimator to be effective in reducing the mean square error.

Concentrating first on the third condition, I suspect it is necessary to get down to the tract or enumeration district (ED) level to achieve sufficient homogeneity. Many private national surveys using area sampling techniques do use ED as a stage of sample selection permitting the base unit coding. However, the Census Bureau currently does not have this capability easily available for about 15 percent of its sample, the part used to represent new construction. Extra efforts would be required to carry out Cohen's procedure.

The real problem, however, stems from the second criterion. Once one departs from the census dates, the estimates of the population of the G strata in each local area become very uncertain. For example, I suppose that proportion of population in various minority classes would be a fairly obvious stratification variable. There have been dramatic and significant changes in the population of such areas in many cities of the United States, and also in minority proportions in these areas. I doubt that most cities have accurate information on the changes that have occurred. We recently contacted a number of local officials and agencies in Maryland in an attempt to update 1970 census data on the percent of black population per tract, and the information was simply not available. The application of Cohen-Kalsbeek method thus seems to me mostly restricted to a period of possibly two or three years after the census. Of course, with the start of mid-decade censuses in the 1980's, this will not be as much of a restriction as it is at present.

There is one study area where the same time restrictions may not apply: studies of political behavior. Election precincts have some of the characteristics of ED's. The geographic sizes and average populations are not too different. However. unlike ED's. election precinct information is brought up-to-date every two years; and in some areas more often. It would be possible to apply the method described in Cohen's paper to studies which use election precincts as stages in sampling.

2. Let me move to another issue, tests of the accuracy of the various procedures that are being developed. Cohen developed several potential population distributions and studied the bias for each distribution. Many of the other papers have proceeded empirically, using information available for local areas from censuses or other sources, and simulating synthetic estimators. Both of these procedures are valuable in giving insight on the conditions under which one method or another is preferable. However, neither procedure is sufficient for most real-life studies that would call for practical applications of synthetic estimates. It is necessary for a technique to have some means of estimating accuracy from survey results without making assumptions about the nature of the underlying distributions. Ultimately, the accuracy depends on the size of the between local area variance. I didn't see any discussion of between-area estimation methods in Cohen's paper. Possibly, it's sufficient to assume that usual methods of estimating components of variance exist.

3. I'd like to add one general remark about potential uses of synthetic estimates. In Maria Gonzalez and Christine Hoza's article, "Small Area Estimation with Applications to Unemployment and Housing Estimates" in the March 1978 issue of the Journal of the American Statistical Association, average mean square errors are shown for estimates of unemployment in 1970 crossclassified by unemployment rate in the area. The errors of the estimates are sort of u-shaped, low for areas with average unemployment rates and much higher for areas at both ends of the distribution, in particular for those at the higher end. This is not too surprising. Synthetic estimates tend to squeeze estimates toward the mean. One of the main purposes of using symptomatic data in a regression model is to compensate for this tendency. Ericksen's work on eliminating outliers is another attempt to reduce the same effect.

I think it is unlikely that these devices will be completely successful. This raises a real dilemma when one attempts to make local area estimates for purposes of administrative action at the local level. For example, if we wish to allocate funds for drug abuse treatment or education on the basis of the size of the problem, then it is precisely the areas that need the funds most whose estimates will be most seriously understated. I am not very optimistic about the possibility of finding the right symptomatic variables to significantly reduce this effect.

There are several courses of action that can be taken. One, of course, is simply to live with the problem. A second is to view synthetic estimates as screening devices, designed to identify the areas where it is reasonably safe to assume that only a small problem exists, and do more intensive work to get a better handle on the problem in areas where the synthetic estimator is above a specific cut-off. The third is to use synthetic estimates not to produce statistics for individual areas, but to produce distributions of the areas, for example, number of areas with drug abuse rates at various levels. If the latter is done, some moderate size should be used to establish the upper end of the class intervals. When it is important to have good estimates for areas at the upper end of the distribution, synthetic estimates are likely to be inadequate unless very effective symptomatic variables exist.

4. There has been occasional reference during the meeting to the elimination of outliers in order to get better fit to models. I am somewhat uneasy about mechanical rules to eliminate or reduce the effect of outliers. My inclination is to view outliers from a quality control point of view, that is, to reexamine them to make sure there are no errors in the data, or for that matter as a clue to the use of other, nonlinear models, rather than to follow mechanical rules of rejection.

Some time ago I saw a dramatic illustration of the dangers of automatic rules on outliers. In the 1966 election, one of the TV networks was making early evening projections of state votes on the basis of data from a sample of precincts. As part of quality control, the percentage Democratic vote in each precinct was compared to past performance in that precinct. Wild fluctuations were removed as being either

data errors or some sort of unrepresentative freaks. In that year, there was an unusual election for governor of Maryland. The Democrats nominated an extremely right-wing, prosegregation candidate. The Republicans nominated someone who was largely unknown, and kept quiet on most controversial issues. As a result, precincts in predominantly black and liberal areas, that had been solidly Democratic in previous elections. suddenlv voted solidly Republican. The analysts in New York, apparently completely unfamiliar with the Maryland situations, proceeded to throw out the results of the sample precincts in such areas. These, of course, were the precincts that most clearly illuminated what was going on in Maryland. The network probably made the worst projection in history on that election. I might say that I was not involved in these projections. The experience, however, is indicative of the dangers in too much "fooling around" with the data.

General Discussion

* There is one point which has just been made by Joe Waksberg that may be worth emphasizing. The point is slightly different from one made earlier. That is, perhaps synthetic estimators could be used for distinguishing outliers which should be given special treatment the next time around in a sample survey, so that one could supplement the sample in those areas in particular. Thus, instead of spreading effort over say. 39.000 units. if you could find some small subset of areas in which a rather different cultural, social, or economic phenomenon exists, then this would be useful for designing the second effort. Thus, there may be a number of uses of synthetic estimators as screeners. The one which has just been suggested should be kept in mind.

(Contributing to the general discussion during this period were: Reuben Cohen, Joseph Steinberg and Joseph Waksberg.)

Part III

Case Studies on the Use and Accuracy of Synthetic Estimates: Unemployment and Housing Applications

Maria Elena Gonzalez

ABSTRACT

A description is given of unemployment synthetic estimates for counties, based on the 1970 Census of Population. The distribution of the method error of these estimates is given, as well as the relative accuracy of these estimates. Implications for intercensal estimates based on regression models are considered.

Vacancy rates from the 1970 Census of Housing are discussed. Estimates of 1970 estimates of dilapidated housing units with all plumbing facilities and their accuracy are analyzed.

INTRODUCTION

Small area estimates are required for the planning and evaluation of programs for individual areas, as well as for the distribution of Federal funds to State and local areas. This great demand has created a need to analyze the different methodologies available to obtain small area data and evaluate the accuracy of the data produced.

One such methodology, called synthetic estimation,[1] has been used to obtain estimates for small areas and as a method of imputation for missing data. In the simplest case a synthetic estimate would use a valid estimate for a large area (e.g., a State), and apply it to all the subareas (e.g., counties) within the State: for the subareas (counties in this case) this estimate would in general be biased. The bias for the subareas is due to the difference which usually exists between the estimate for the large area and the various estimates for the subareas. In most of the examples to be discussed in this paper, synthetic estimates are derived by partitioning the universe into a series of mutually exclusive and exhaustive cells and deriving the estimate as a sum of products. In the case of unemployment,the~ightscorrespondto the distribution in the small area of the labor force by age, for

example, and the estimated unemployment by age corresponds to the estimate for the larger area.

A formula expressing the synthetic estimate, is:

$$u_i^* = \sum_{j=1}^{G} p_{ij}\, u_{.j} \qquad (1)$$

where p_{ij} is the labor force for county (or subarea) i and characteristic j, $\sum_{j=1}^{G} p_{ij} = 1$, and $u_{.j}$ is the unemployment rate for characteristic j in the State (or larger area).

In this paper, we review synthetic estimates of unemloyment derived for counties at the time of the 1970 Census of Population; these estimates are compared with the Census 20-percent sample estimates of unemployment to obtain and analyze the distribution of the method error of the synthetic estimates. In addition, some regression estimates of unemployment which might be used intercensally (the years between decennial censuses) are presented..

In the area of housing, we present data on vacancy rates. In the 1970 Census of Population and Housing, it was found that about 11% percent of the housing units initially reported by enumerators as vacant were occupied. An adjustment for these misclassified vacant units was included in the processing, and some effects will be described (see Gonzalez and Waksberg 1973). The pretests for the 1980 census shed some further light on these results. In addition, the possibility of estimating vacancy rates intercensally is explored.

In the 1970 Census of Housing, estimates of housing units dilapidated with all plumbing facilities (DWAPF) were obtained by synthetic methods. The relative accuracy of these estimates is discussed.

UNEMPLOYMENT STATISTICS

The 1970 Census of Population data on unemployment, collected from a 20-percent sample, were used to calculate various alternative synthetic estimates of unemployment for counties in the United States. This allows us to compare the Census and synthetic estimates. The unemployment estimates for geographic divisions were used as the basis for the $u_{.j}$, for a number of different characteristics, j. The characteristics used to compute synthetic estimates included sex, race (black vs. all other races), and alternative classifications of the population by: occupation, age-marital status, industry and occupation-income (see Gonzalez and Hoza 1978). The definition of the cells (mutually exclusive and exhaustive) used to compute the alternative synthetic estimates was determined empirically, trying to minimize the number of cells for which many counties

had zero persons in the labor force.[2] It is possible that by means of a more systematic approach, such as the use of cluster analysis for defining the cells, improved results could be obtained.

The synthetic estimates based on race - sex - occupation classification provided the highest weighted correlation, 0.682, with the county estimates for the 1970 Census. Within each of the nine geographic divisions, the number of cells used to compute the synthetic estimate based on race - sex - occupation was 31: 12 cells for nonblack males, 9 cells for nonblack females, and 5 cells each for black males and black females.

The synthetic estimate based on race - sex - age - marital status resulted in a weighted correlation for all counties of 0.569. This synthetic estimate used 50 cells within each of the nine geographic divisions. The increase in number of cells did not, in this case, result in a higher correlation with the Census estimate. Computing the county synthetic estimates based on the unemployment rates for the geographic divisions where they are located might not lead to the most efficient results. It is possible that a more homogeneous grouping of counties would give better results. In this analysis, however, other groupings of counties were not tried.

Table 1 shows the number of counties classified by the 1970 Census estimate of unemployment, as well as the root mean square error and the relative root mean square error for the synthetic estimates based on race - sex - occupation and those based on race - sex - age - marital status classifications. The root mean square error was estimated as:

$$(MSE_{u_i^*})^{\frac{1}{2}} = \left(\frac{1}{M} \sum_{i=1}^{M} (u_i^* - u_i)^2 \right)^{\frac{1}{2}} \qquad (2)$$

where u_i is the 1970 Census unemployment estimate for county i, and M is the number of counties with a specified unemployment rate in the 1970 Census (e.g., counties with unemployment rate from 4.0 percent to 4.9 percent).

The root mean square error is smaller for the synthetic estimates based on occupation than for those based on age - marital status categories: 1.9 versus 2.2 percent. The smallest relative root mean square error corresponds to unemployment between 4.0 percent and 4.9 percent, which is also the category where the overall U.S. unemployment rate falls (4.4 percent). For counties with unemployment rate below 3 percent and those above 11 percent, for synthetic estimates based on occupation, the relative root mean square error was above 0.5. This results in a U-shaped distribution. Because of the smoothing characteristic of the synthetic estimates, the estimates corresponding to 1970 Census unemployment estimates further away from the average tend to be less accurate than those for counties with 1970 Census estimates closer to the average

TABLE 1

Distribution of the Root Mean Square Error of Synthetic Estimates by Counties by Size of 1970 Census Unemployment Rate

1970 Census Unemployment Rate	Counties[a]	Root Mean Square Error(%)		Relative Root Mean Square Error[b]	
		Occupation	Age-Marital Status	Occupation	Age-Marital Status
Less than 1.0%	21	2.8	2.8	5.52	5.56
1.0% - 1.9%	171	2.0	1.5	1.36	0.99
2.0% - 2.9%	493	1.4	1.2	0.57	0.50
3.0% - 3.9%	679	0.9	0.7	0.24	0.21
4.0% - 4.9%	580	0.6	0.8	0.14	0.18
5.0% - 5.9%	363	1.2	1.6	0.22	0.28
6.0% - 6.9%	232	1.8	2.3	0.28	0.36
7.0% - 7.9%	137	2.5	3.0	0.33	0.40
8.0% - 8.9%	88	3.4	4.1	0.40	0.48
9.0% - 9.9%	51	4.3	4.9	0.46	0.52
10.0% - 10.9%	30	4.8	5.5	0.46	0.52
11.0% - 11.9%	22	6.5	7.1	0.56	0.62
12.0% - 12.9%	23	7.2	7.9	0.58	0.63
13.0% - 13.9%	10	8.1	8.9	0.60	0.66
14.0% - 14.9%	2	8.4	9.1	0.58	0.62
15.0% - 16.9%	6	10.4	11.3	0.66	0.71
Average 4.4%	2908	1.9	2.2	0.43	0.50

a See footnote 2.

b The relative root mean square error was calculated by dividing the root mean square error by the mid-point of the unemployment interval.

unemployment rate. The results for synthetic estimates of unemployment based on age-marital status are similar to those based on occupation. Although the variance was not separately estimated, if it is relatively small, then the bias is not negligible.

Figure A plots the distribution of the relative method error for synthetic estimates based on occupation and those based on age - marital status. The relative method error for the unemployment rate is calculated as the difference between the synthetic estimate and the Census estimate divided by the Census estimate. For synthetic estimates based on occupation,48.3 percent of the counties had a negative relative method error and for synthetic estimates based on age - marital status, the corresponding percentage is 54.3. If we disregard the sign, a relative method error of 0.2 or less is obtained by 43.0 percent of the synthetic estimates based on occupation and by 38.3 percent of those based on age - marital status. Similarly, a relative method error of 0.5 or less is obtained by 79.9 percent of the occupation synthetic estimates and by 79.3 percent of the age - marital status synthetic estimates. About 95 percent of the counties for both distributions have a relative error of 1.0 or less. Approximately 1.1 percent of the 2908 counties tabulated had a relative method error over 2.0. The charts show quite similar distributions of the relative method error for both synthetic estimates; this result is expected since there is a very high correlation, 0.916, between the occupation and age - marital status synthetic estimates.

Fbr intercensal estimates of unemployment, we will consider regression estimtes for 122 Current Population Survey (CPS) primary sampling units (PSUs) (see Ericksen 1974). The CPS is a monthly survey which collects data on employment and unemployment. The data of the survey can be tabulated for individual PSU's, although the data are subject to a very high variance. The regressions use as dependent variables two summarizations of the CPS PSU data: (1) a one month summary based on the April 1970 data, Z, and (2) a summary of five months of CPS data centered in April 1970 and spaced at quarterly intervals, Y. The independent variables include 1970 Census estimates, U, and alternative estimates based on the unemployment insurance data, as well as synthetic estimates based on sex - race - occupation classifications, X_2.

The following regressions are obtained:

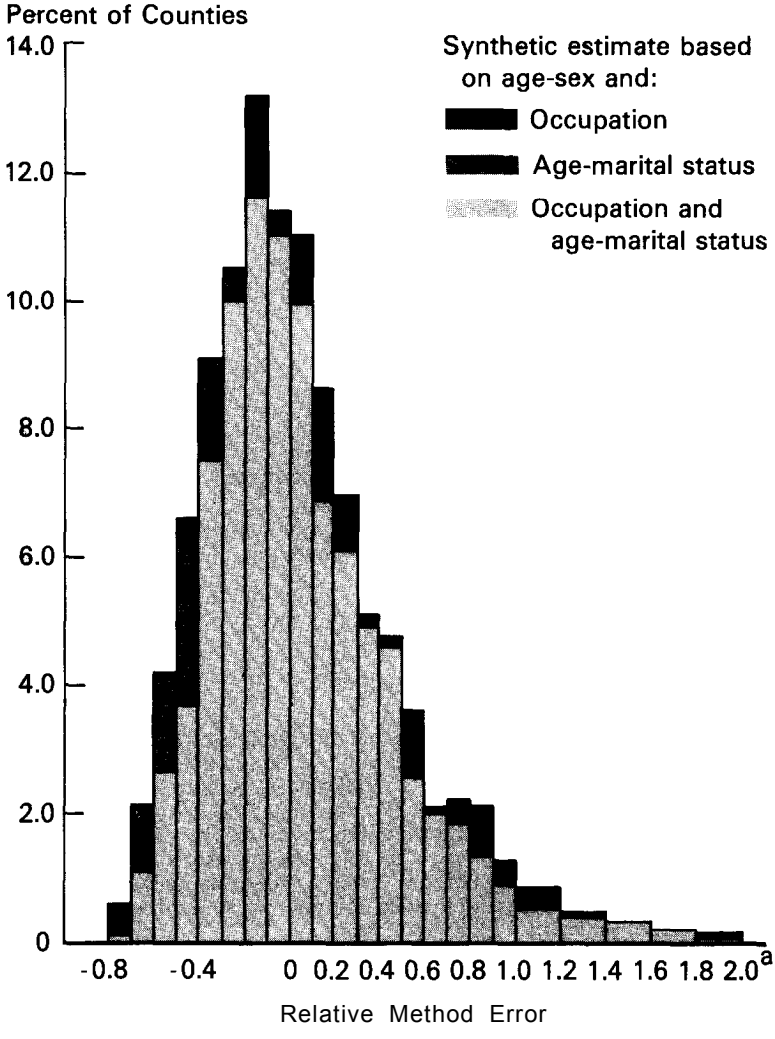
$$Y' = .016 + .884\,U - .080\,X_1 - .023\,X_2 \qquad (3)$$

$\bar{R}^2 = .540$

Residual mean square = $.881 \times 10^{-4}$

Standard error of estimate = $.938 \times 10^{-2}$

Figure A. Distribution of Relative Method Error for Alternative Synthetic Estimates of the Unemployment Rate for Counties in the United States, 1970



a 1.1% of the 2908 Counties Tabulated had a Relative Method Error over 2.0.

where X_1 is the insured unemployment as a percent of total unemployment.

$$Z' = .026 + 1.016\,U - .107\,X_1 - .396\,X_2 \quad (4)$$

$\bar{R}^2 = .263$

Residual mean square = 2.119×10^{-4}

Standard error of estimate = 1.456×10^{-2}

Because of the higher variance of Z, based on one month of CPS data, regression (4) shows a lower correlation than regression (3) which uses a dependent variable based on an accumulation of five months of CPS data.

Additional regressions follow:

$$Y' = .010 + .450\,U + .089\,X_2 + .326\,X_3 \quad (5)$$

$\bar{R}^2 = .563$

Residual mean square = $.835 \times 10^{-4}$

Standard error of estimate = $.914 \times 10^{-2}$

where X_3 is the new final annual "70-step" estimate[3] of unemployment before benchmarking the estimates by CPS data.

$$Z' = .019 + .442\,U - .247\,X_2 + .430\,X_3 \quad (6)$$

$\bar{R}^2 = .291$

Residual mean square = 2.040×10^{-4}

Standard error of estimate = 1.428×10^{-2}

The results show a slight improvement of the correlation in the regressions which use X_3 rather than X_1 as ah independent variable. However, in selecting the independent variables, the availability and timeliness of the variables must be taken into account. For the sample areas further improvements in the estimates could be achieved by combining the CPS PSU sample data with the regression estimates (Fay and Herriot 1978). Nevertheless, the regression methodology provides a feasible way of obtaining intercensal small area estimates of unemployment.

HOUSING STATISTICS: VACANCY RATES

After the initial completion of the enumeration for the 1970 Census of Population, a National Vacancy Check (NW) sample survey was

carried out (U.S. Bureau of the Census 1974b). Reinterviews were conducted for a sample of housing units initially reported as vacant by the enumerators to check whether they might have been occupied at the time of the census. The results of this survey showed that an estimated 11.4 percent of these vacant housing units actually occupied at the time of the census. This project was intended originally as an evaluation project of the 1970 Census, but when the extent of the problem became apparent, the project was converted into an operational census procedure. One possible reason why occupied housing units might have been erroneously classified as vacant was that the enumerator could not find anybody to report whether or not the unit could have been occupied at the time of the census. Based on the results of the NVC and the size of household found in the misclassified units, twelve conversion rates (4 regions x 3 types of census procedures) were used during the processing of the census to convert vacant housing units into occupied ones and to assign to the vacant units the nunber and characteristics of the persons in a neighboring unit. This is a type of synthetic estimate,andan analysis of the effects of this procedure on the population estimates for areas of different sizes is given in the paper by Gonzalez and Waksberq (1973). As a result of this procedure, 1,069,000 persons were added to the 1970 Census.

The main intent of this coverage improvement procedure was to adjust for population undercoverage. The percentage of housing units initially reported as vacant, but actually occupied (11.4 percent) was adjusted downward in determining the conversion rates (8.5 percent overall), because the average size of household for misclassified units was smaller than the average size of household reported in the 1970 Census. Therefore, fewer vacant housing units were converted into occupied than the estimate given by the NVC survey. In fact, the procedure used under-imputed population, because vacant housing units were neighbors of smaller than average households in the census. The vacancy rate, computed as the percent vacant of the total nonseasonal housing units, was affected by the imputation procedure used; the imputation procedure improved the initially reported vacancy rate, but additional housing units would have needed to be converted into occupied ones to improve further the estimates for 1970 vacancy rates.

Two main variables were measured in the NVC: misclassified vacant housing units, and persons living in these units. In specifying an improved mutation procedure, it would be necessary to control both variables: the number of housing units converted from vacant into occupied, as well as the total number of persons (and distribution by household size) to be imputed. For example, Figure B illustrates the needed controls to achieve specified housing unit and population control totals.

FIGURE B

Region 1	Size of household					
	Total	1	2	3	4	...
Type of enumeration 1						
Housing units to be converted from vacant to occupied	Σx_i	x_1	x_2	x_3	x_4	...
Type of enumeration 2						
Housing units to be converted from vacant to occupied	Σy_i	y_1	y_2	y_3	y_4	...
...						

The plans for the 1980 Census of Population and Housing include an independent reinterview of all housing units initially reported as vacant or deleted from the original list of addresses in order to be able to process a more correct count of persons and occupied and vacant housing units (U.S. Bureau of the Census 1978).

The possibility of estimating vacancy rates intercensally for small areas requires the use of the Annual Housing Survey (national and SMSA) sample data and the Quarterly Vacancy Survey as dependent variables and the use of regression techniques similar to those illustrated for estimating unemployment rates. Such aprojectneeds to determine the availability of local area data which might be used as independent variables, such as building permits issued or turnover in households.

HOUSING STATISTICS: DILAPIDATED HOUSING WITH ALL PLUMBING FACILITIES

Synthetic estimates were used in the 1970 Census of Housing (Vol. VI) to provide estimates of the component of substandard housing units which were dilapidated with all plumbing facilities (DWAPF). The 1970 census procedures did not provide for individual rating of structural condition, such as sound, deteriorating, and dilapidated, as was used in the 1960 Census of Housing. In 1970, census data on housing units with all plumbing facilities for specified areas and cells were multiplied by estimated proportions of dilapidated housing units which had all plumbing facilities, as derived from a post-census survey, Components of Inventory Change (CINCH) to obtain the synthetic estimates of DWAPF (Gonzalez 1973).

The estimate of accuracy used to evaluate the estimates of DWAPF was the root mean square error computed as follows:

$$(MSE_{D_i})^{\frac{1}{2}} = \left(\sum_{j=1}^{G} D_{ij}^2 \, var(q_j) + \frac{1}{M} \sum_{i=1}^{M} (D_i^* - D_i)^2 \right)^{\frac{1}{2}} \quad (7)$$

where

D_{ij} is the number of housing units with all plumbing facilities in area i for characteristic j (j=1,...,G) based on the 1970 Census of Housing

$var(q_j)$ is an estimate of the variance of the proportion of dilapidated housing with all plumbing facilities for characteristic j from CINCH

D_i^* is the synthetic estimate of DWAPF for area i based on the 1960 Census of Housing

D_i is the 1960 Census of Housing 25-percent sample estimate of DWAPF for area i

M is the number of areas being averaged.

The average of the squares of the 1960 biases for a group of areas was used as an approximation of the square bias for the 1970 DWAPF estimates for that same group (U.S. Bureau of the Census 1974a) . The relative size of the estimated root mean square error depends on the size of the area being estimated. Tables 2, 3, and 4 give relative root mean square error for geographic divisions, States and counties by size of estimate. These estimates provide only rough indications of the accuracy of the data, but in general for larger areas the relative root mean square error is smaller.

TABLE 2

Approximate Relative Root Mean Square Error of 1970 Estimates of Dilapidated Housing Units with all Plumbing Facilities for Division, by Inside and Outside SMSA's

	Relative root mean square error for division a/		
Size of estimate	Total	Inside SMSA	Outside SMSA
20,000 - 49,999	-	0.35	0.48
50,000 - 99,999	0.30	0.20	0.24
100,000 - 199,999	0.15	0.16	0.17
200,000 & over	0.15	0.12	

a The relative root mean square error was calculated by dividing the root mean square error by the lower limit of the size class as given in Table H of U.S. Bureau of the Census (1974a) .

TABLE 3

Approximate Relative Root Mean Square Error of 1970 Estimates of Dilapidated Housing Units with all Plumbing Facilities for States

Size of estimate	Relative root mean square error for States[a]
1,000 - 4,999	1.00
5,000 - 9,999	.42
10,000 - 19,999	.36
20,000 - 29,999	.26
30,000 - 49,999	.23
50,000 - 99,999	.20
100,000 - and over	.18

a The relative root mean square error was calculated by dividing the root mean square error by the lower limit of the size class as given in Table I of U.S. Bureau of the Census (1974a) .

TABLE 4

Approximate Relative Root Mean Sguare Error of 1970 Estimates of Dilapidated Housing Units with all Plumbing Facilities for Counties within SMSA's by Region

	Relative root mean square error for county[a]			
Size of Estimate	Norhteast	North Central	South	West
100 - 249	1.00	1.00	1.00	1.00
250 - 499	1.20	.80	.80	.80
500 - 999	.80	.60	.60	.60
1,000 - 4,999	.70	.90	.60	.80
5,000 and over	.34	.34	.32	.72

a The relative root mean square error was calculated by dividing the root mean square error by the lower limit of the size class as given in Table J of U.S. Bureau of the Census (1974a).

IMPLICATIONS FOR OTHER VARIABLE

The results presented here illustrate the uses and limitations of synthetic and regression estimates in the case of unemployment rates, housing vacancy rates, and housing units dilapidated with all plumbing facilities. However, the methods used could be applied to other subject-matter fields; the accuracy of the resultant data would probably depend on the specific data set used. In whatever context these methodologies would be applied, data relevant to the specific field are needed. For example, in the data shown on unemployment rate, the basic sources used were the 1970 Census of Population unemployment rates, as well as Current Population Survey data.

In the future, synthetic estimates will be used often. We need to recognize that at present synthetic estimates are sometimes used without being recognised as such; producers of data may not always be aware of the implications for the accuracy of the data of using synthetic estimates.

FOOTNOTES

1. The terminology was first used by the U.S. National Center for Health Statistics (U.S. Department of Health, Education and Welfare).

2. 2908 "counties" were analyzed (counties with population of less than 5,000 in the 1970 census were merged with a neighboring county). SMSA counties were never merged with non-SMSA counties; counties in the 1960 or 1970 CPS design were merged only with counties in the same PSU.

3. The Bureau of Employment Security (now Employment and Training Administration) of the Department of Labor published in 1960 a "Handbook on Estimating Unemployment" which describes the 70-step method. This Handbook specifies a series of computational steps (about 70) designed to produce unemployment estimates. These estimates are the sum of three components:

   a. Unemployed persons who were employed in an industry and Were covered by unemployment insurance immediately prior to their unemployment spell.

   b. Unemployed persons who were employed in an industry and Were not covered by unemployment insurance immediately prior to their unemployment spell.

   c. Unemployed persons who Were new entrants and reentrants into the labor force.

   The basic building block of these estimates of unemployment is the count of insured unemployed.

REFERENCES

Ericksen, E.P., A Regression Method for Estimating Population Changes of Local Areas, Journal of the American Statistical Association, Volume 69, 1974, pp. 867-875.

Fay, R.E., and Herriot, R., Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. Unpublished, 1978.

Gonzalez, M.E., Use and Evaluation of Synthetic Estimates, Proceedings of the Social Statistics Section of the American Statistical Association, 1973, pp. 33-36.

_____ and Boza, C., Small Area Estimation with Applications to Unemployment and Housing Estimates, Journal of the American Statistical Association, Volume 73, 1978, pp. 7-12.

_____ and Waksberg, J., Estimation of the Error of Synthetic Estimates, unpublished paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria, 1973, pp. 1-17.

U.S. Bureau of the Census, Census of Housing: 1970, Volume VI, Plumbing Facilities and Estimates of Dilapidated Housing, Addendum: Accuracy of Estimartes, Washington, D.C.: U.S. Government Printing Office, 1974a, pp. 1-7.

U.S. Bureau of the Census, 1970 Census of Population and Housing Effect of Special Procedures to Improve Coverage in the 1970 Census, Washington, D.C.; U.S. Government Printing Office, 1974b, pp. 11-14.

U.S. Bureau of the Census, Proposals for Coverage Evaluation of the 1980 Census, Presented at the March 2, 1978, Meeting of the Census Advisory Committee of the American Statistical Association. Unpublished, 1978.

U.S. Department of Health, Education and Welfare, Synthetic State Estimates of Disability, PBS Publication No. 1759, Washington, D.C.: U.S. Government Printing Office, 1968.

Some Recent Census Bureau Applications of Regression Techniques to Estimation

Robert E. Fay

INTRODUCTION

Adaptations and extensions of the classical theory of regression and linear models constitute one of the possible approaches to estimation for small areas. This paper will describe three recent applications of this theory to problems at the Census Bureau and indicate possible future directions. Much of what is presented here must be classified as simply exploratory research; yet, each of the three investigations has had tangible effects upon aspects of Bureau policy. Furthermore, with preliminary plans for evaluation of the 1980 census calling for use of regression and/or synthetic techniques to produce subnational estimates of undercount at particular levels of geography, the interest of the Bureau in these techniques may be expected to increase.

Because synthetic estimates are the principal topic of this workshop, the relation between regression and synthetic estimation serves as a natural point of departure. The two are linked by their common basis in linear models. For purposes of discussion here, we shall consider a linear model over any set of geographic units i to be a representation

$$c_i = \sum_{j=1}^{p} X_{ij} \beta_j + u_i \qquad (1)$$

of a characteristic c_i in terms of a linear transformation of the predictor variables X_{ij} plus a residual term u_i. The common vector representation for equation (1)

$$c = X\beta + u \qquad (2)$$

will also be employed in this paper.

Synthetic estimates may be expressed in the form of (1). In this instance, the X_{ij}'s become relative or absolute frequencies of population subgroups j in units i, while the β_j's become the rates of incidence of the characteristic in the subgroup j over the entire set of units. On the other hand, linear regression, or more specifically, weighted least squares, determinesthevector β through

$$\beta = (X^T W X)^{-1} X^T W c \qquad (3)$$

where W is a diagonal matrix of weights W_i. (In some applications, not included among those presented here, W may be other than diagonal.) This second approach, unlike synthetic estimation, does not impose structural restrictions upon X. In a sense, a synthetic estimate models relationships in the population at a micro-level, while a regression estimate models only at a macro-level.

The preceding description of the linear model departs somewhat from the usual. Here, equation (2) stands by itself as a mathematical relation between the terms. The practice in most linear theory directly links this equation to a stochastic model for u, and occasionally for X or β as well. In so doing, the statistical issues in linear theory are typically grounded in the properties of infinite populations. The conceptual standard for the evaluation of small area estimates, on the other hand, is generally the complete census (whether this census is actual or hypothetical), and this standard casts the problem in the context of the finite population. Equations (2) and (3) will, therefore, represent definitions of finite population parameters, although we shall at points consider implications of stochastic assumptions.

POST-CENSAL ESTIMATION OF POPULATION

The Census Bureau currently employs (2) and (3) in one of its methods of post-censal estimation, the ratio-correlation method, at the levels of both States and counties. (In what follows, simplifications will represent the nature of the statistical problem without fully detailing the implementation. A complete description is given in U.S. Bureau of the Census (1976).) TheX_{ij}'s are taken to represent the ratio of change in indicator variable j in unit i to the change at the national level (or in the case of counties, State level) in this manner:

$$X_{ij} \quad \frac{\dfrac{\text{Value of j at current year, unit i}}{\text{Value of j at census year, unit i}}}{\dfrac{\text{Value of j at current year, total}}{\text{Value of j at census year, total}}} \qquad (4)$$

Examples of indicator variables are data on school enrollment, automobile registration, tax returns, and labor force size. The c_i's are the corresponding rates of change in population and are defined analogously to (4). For example, if school enrollment decreases by 5 percent nationally but increases by 14 percent in a particular State, the value for the corresponding X_{ij} would be 1.20 (=1.14/.95). If the same State's population grew by 32 percent during a period in which the national growth was 10 percent, the value of c_i would be 1.20 also (=1.32/1.10). In a sense, therefore, each of the indicator variables is expressed in a form to indicate directly the relative change in population compared to the national rate of change. The β_j's act as weights to combine the various changes implied by the indicator variables. The current practice is not to force the weights to sum to unity but to include a constant term in the model as well, equivalent to setting $X_{ij} = 1$ for all i and some j.

Current estimates of population are computed as $X\beta$, where the X_{ij}'s are defined according to (4) for the current year relative to 1970. The Census Bureau derives β as $\beta_{60\text{-}70}$, the application of (3) to the 1960-1970 decade (that is, with X_{ij} and c_i defined as in (4) with 1970 as the current year and 1960 as the census year). W has been taken to be the identity matrix, thus giving equal weights to the geographic units.

Ericksen (1973, 1974) first outlined and investigated a technique, the regression-sample method, to estimate the current coefficients, β_c, that would result from (3) if a census were taken to determine the true values of c_i. He proposed the use of Y_i, sample estimates of the relative growth since 1970 in each sampled primary sampling unit (PSU, a county or group of counties), in the Current Population Survey (CPS). Using the X_{ij}'s for the current year relative to 1970,

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y \qquad (5)$$

estimates β_c. Because of considerations of sampling variance in Y, he employed weights W_i approximately inversely proportional to the estimated sampling variance of Y_i.

Ericksen delineated three sources of error in the estimates:

1. The random error not explained by the indicators.
2. The error due to structural changes in regression.
3. The sampling errors in the CPS estimates.

He noted that the ratio-correlation method and regression sample method are equally subject to the first source of error, whereas ratio-correlation is affected by the second and regression-sample by the third.

Another fundamental idea appears in these papers by Ericksen, namely that the sample data may provide an estimate of an average mean square error for the current estimates. In this computation, the average square of bias is defined as

$$\sigma_v^2 = \frac{u^T W u}{1^T W 1} \qquad (6)$$

where u was defined by (Z), and 1 is a column vector of 1's. With W_i^{-1} taken as the sampling variance of Y_i,

$$E((Y-X\hat{\beta})^T W (Y-X\hat{\beta})) = n - p + \sigma_v^2 1^T W 1 \qquad (7)$$

where n is the number of Y_i's and p is the rank of X. (The notation and some constants here have been altered from Ericksen's original paper in order to set the problem in the finite population context, although neither this nor his paper fully attacks the exact constants required to represent the effects of the first-stage selection in CPS. The practical consequences are trivial, however.) In this manner, the sample data may be used to measure the magnitude of error from the changes not explained by the indicators; classical regression theory gives the error due to sampling error in Y. Consequently, both components of the error may be estimated.

William Madow first noted [in a seminar given at the Census Bureau) that a judiciously selected weighted combination of $X\hat{\beta}$ and Y would produce estimates with smaller average error than $X\hat{\beta}$. For example, the combination

$$E_i = (X\hat{\beta})_i + \left(1 - \frac{W_i^{-1}}{W_i^{-1} + \sigma_v^2}\right) (Y_i - (X\hat{\beta})_i)$$

where W_i^{-1} is again the sampling variance of Y_i, is related to the original James-Stein estimator. The application of (8) or similar combinations has insignificant effects in this instance because of the large sampling error in Y, but similar formulas play a central role in a third example to be discussed here.

If the finite population is the standard for evaluation, three other possible sources of error in the regression estimates deserve addition to Ericksen's list:

4. The error due to differences between the population regression equations for sampling units (PSU's) and for the units of analysis (States or counties).

5. The error arising from bias in the sample data.

6. The consequences of redistributing error among units by altering the weights in the regression.

All three factors are at issue in this application: the use of the PSU in substitution for direct analysis of States or counties, deficiencies and lags in the CPS sampling frame whose effects may be distributed unevenly across the country, and a possibly undue emphasis in the weighting on estimating the most populous units (efficient in terms of sampling error but possibly undesirable as a population parameter). Several questions thus remain unanswered as to the practical merit of Ericksen's suggestion in this case, although his idea may have significant effects elsewhere.

A separate section of this paper describes alternative statistical procedures that may be used to provide evidence on how the current indicators should be weighted to estimate population change. Ericksen had formulated the problem as a dichotomy between use of past relationships applied without evidence of their currency and sample-regression methods that make an effort to he current at the cost of substantial sampling error. Relationships between the indicator variables themselves may be examined. Since this approach is unrelated to the methods in the other two applications to be discussed here, this topic is deferred to the end of the paper.

CHILDREN IN POVERTY

The second example to be discussed here is a direct application of Ericksen's regression-sample method to the problem of estimating the proportion of school-age children living in poverty families by State. Congress has employed census counts of these children by county in apportioning approximately $2 billion annually under Title I of the Elementary and Secondary Education Act of 1965. Recognizing the potential for change since 1970 in the relative distribution of poor children among States, Congress included in the Educational Amendments of 1974 a directive to the Secretaries of Commerce and of Health, Education, and Welfare to conduct a survey to produce sample estimates of children in poverty families by State. In comPliance with this legislation, the Census Bureau carried out the Survey of Income and Education (SIE) in the Spring of 1976.

In 1975, prior to the SIE, research at the Census Bureau explored other techniques to estimate the proportion of children in poverty families by State: After initial investigations of regression models of the 1970 proportions of children in poverty using other 1970 data, it became apparent that these equations were unlikely to carry forward in time adequately. This problem with a fixed regression model based upon the preceding census is, of course, the second source of error listed earlier that had been identified by Ericksen, namely, "the error due to structural changes in regression." Consequently, an adaptation of the sample-regression method was attempted, again using the CPS to provide current sample estimates of the dependent variable, Y_i, this time the proportion of children 5 to 17 years old in poverty families in each State. Unlike Ericksen's experiments with predicting changes in population, the sample data were employed at the State, rather than PSU, level.

Experimental regressions, modeling 1970 poverty rates for families by State based upon 1960 census and other data available independently of the 1970 census, pointed to the fundamental importance of total income. Estimates of Per Capita Personal Income (PCI) published annually by the Bureau of Economic Analysis (BEA) are employed in the model. Other variables associated with poverty, including female headship, racial composition, unemployent, and region, did not appreciably add to the explanation afforded by the model.

The final model proposed for years after 1970 consists of five independent variables plus a constant term. The poverty rate for children from the 1970 census is the first, while two variables are formed from BEA PCI for the census year (income year 1969) by first finding the median of the 51 State (and D.C.) PCI figures, PCI_m, and computing

$$X_{i2} = \ln(PCI_i/PCI_m) \quad \text{if } PCI_i > PCI_m \tag{9}$$

$$= 0 \quad \text{otherwise}$$

$$X_{i3} = 0 \quad \text{if } PCI_i > PCI_m \tag{10}$$

$$= \ln(PCI_i/PCI_m) \quad \text{otherwise}$$

The variables X_{i4} and X_{i5} are formed similarly from BEA PCI for the current year (the year immediately preceding the survey date), and, finally, X_{i6} is taken to be identically 1, so that $\hat{\beta}_6$ is the constant term.

The assessment of this technique was originally based upon its performance in relation to the 1970 census. A parallel model was developed for the proportion of families in poverty, with 1960 as the base year and 1970 as the current year. The 1970 census values for the proportion of families in poverty were used in place of sample estimates as the dependent variable. Thus, the lack of fit in this case is the bias of the model. When this research was conducted in 1975, an effort was made to characterize the distribution of these biases. The principal determinant seemed to be size: when States were grouped into four strata by population, the largest States had errors averaging only four percent, while the second group averaged about six, and the smaller groups, ten. Other experiments suggested that the relative error for children in poverty was likely to be approximately the same as for families in poverty, so these relative errors were interpreted as rough indications of the level of error for children. (The lack of counts from the 1960 census of children in poverty by State necessitated this indirect evaluation.)

The sampling errors for CPS State estimates of the proportion of children in poverty are simply too large to support the estimation by (7) of the average error as suggested by Ericksen. It is possible, however, to compute the sampling variance of the regression estimate for each State and to add an allowance for bias based upon the 1960-1970 test regression for families in poverty. With these estimates of the components of error, it is also possible to weight the sample and regression estimates together, as in (8). In only two States, however, New York and California,

does the weight on the sample estimate exceed .2 in this computation.

As mentioned earlier, the legislative directive was for a survey sufficient to produce State estimates. The 1976 SIE was of adequate size and design for this purpose, and in fact the sampling variances for States were generally lower than the preceding research suggested could be obtained as mean square errors-for the-regression estimates from CPS. From the perspective of 1975. therefore. the SIE seemed to afford an opportunity for a definitive evaluation of the regression estimates. In particular, the computation (7) of the mean square error for the regression estimates could be performed with the expectation of interpretable results, unlike the situation with CPS. In point of fact, however, the relationship between the regression and SIE estimates turned out to be more complex. In two important respects to be described here, the regression results served the purposes of the survey, once in the design and later in the evaluation, whereas a precise assessment of the bias of the regression model itself could not be obtained.

Under an agreement with the respective legislative committees, a specification for a coefficient of variation of 10 percent on the SIE estimate of the number of poor children in each State was chosen. This specification created some difficulty, since an efficient and practicable survey design required prior estimates of the current poverty rates for children in each State. If a prior estimate in a given State was too high, an insufficient sample size would have resulted, and the specifications would not have been met. In order to provide some protection against this occurrence, both the 1970 census poverty rates and the regression estimates based upon the March 1975 CPS were considered, and the smaller of each pair was used for purposes of design. Thus, the regression estimates helped to target additional sample to States in which the poverty rate had decreased since the 1970 census.

The regression estimates proved even more valuable in evaluating the SIE. The whole question of evaluation was critical in the case of this survey: for the first time Congress specifically legislated that an evaluation be performed, by requiring a report on the outcome of the survey, "including analysis of its accuracy and the potential utility of the data derived therefrom . . . " In response to this directive, the Census Bureau conducted an extensive evaluation of the SIE results. The principal basis for the evaluation was a reinterview of an approximately five-percent sample of SIE and of CPS households by more intensive intcrviewing techniques. (This reinterview survey is described in Fay (1978) and in the U. S. Bureau of the Census report (1978). "Assessment of the Accuracy of the Survey of Income and Education:")

The SIE yielded results that appeared to require explanation; in particular, the SIE national estimate of children in poverty was 12 percent below the corresponding value obtained by the CPS, a result that could not be ascribed to sampling error alone. On this point the reinterview data supported the SIE: there was no significant change in the national estimate in the SIE reinterview, whereas the reinterview result for CPS lowered the CPS estimate by about 20 percent. The CPS reinterview estimate consequently stood within sampling error of the original SIE result but not within sampling error of the CPS result. The SIE reinterview also detected no statistically significant bias by region or division.

Other questions could not be answered by the reinterview alone. The significance of the difference between the SIE and CPS national estimates is compounded by the fact that the 1970 CPS produced an estimate for children in poverty about 10 percent lower than the 1970 census. By combining these differences, it could be argued that had a national census been taken in 1976, the result for children in poverty might have exceeded the SIE by over 20 percent. Others suggested that, because of this potentially large difference in level, the SIE results for the distribution of poverty among States would be essentially incompatible with the census measurement of poverty. (See, for example, Ginsberg and Grob (1977).) The CPS regression estimates provided the most direct evidence on this question, since they linked 1970 to 1976 by an annual series obtained from a consistent methodology. Figures 1 to 4 show the trends in the series by division over this period, expressing the estimates in terms of the percent of the total number of poor children residing in each region. In essentially every case, the direction of change in the proportion of the total number of children in poverty agrees with the conclusions obtained in comparing the census and SIE; the Northeast, East North Central, and Pacific States have increased their share of the total, while a substantial decline has occurred throughout the South. This evidence implies that the SIE and census procedures would measure essentially the same distribution of poverty among States even though their national levels may differ markedly.

When the regression equation is fitted to the SIE data, there is a relatively strong agreement between the regression and sample estimates for the proportion of children in poverty by State. Table 1 shows these results. The average difference between the two sets is 14 percent (root mean square), whereas the average difference between the SIE and 1970 census values is 23 percent. Since the sampling error in the SIE estimates was approximately 10 percent, (7) gives an average bias in the regression of about 10 percent ($14^2 \doteq 10^2 + 10^2$).

The most remarkable outcome, however, comes from the comparison of the regression and reinterview. When each is classified by the direction of difference from the SIE, Table 2a results. Thus, there is an apparent statistical agreement between the two. A covariance adjustment to the SIE estimates, which did not change the reinterview measures of shift, produces Table 2b, which shows a highly significant relation. (The nature of the covariance adjustment and other specifics of the analysis are described in the report.) Consequently, the reinterview, which had not otherwise been noted to demonstrate any consistent pattern of shift, actually does measure a component of non-sampling error in the SIE State **estimates.** Analysis indicated that the magnitude of the non-sampling error was roughly 7 percent, although this result is measured to limited precision because of large sampling error in the reinterview estimates. Since the non-sampling error in the SIE is included in the preceding estimate from (7) of a 10 percent average bias for the regression, it is difficult to establish precisely the actual level of bias for the regression if the non-sampling error in the SIE were excluded, except to say that it is less than 10 percent, perhaps 7 percent.

The last finding represents possibly the first application of a technique to measure non-sampling error. Whether other applications are possible will depend upon the availability of both a successful model and independent estimates of net survey error that are obtained by a more controlled process than the original survey.

FIGURE 1

MODEL ESTIMATES BASED ON CPS OF THE PERCENT OF TOTAL POOR CHILDREN IN THE NORTHEAST REGION, BY INCOME YEAR AND DIVISION (1970 Census and 1976 SIE Estimates Shown Circled)

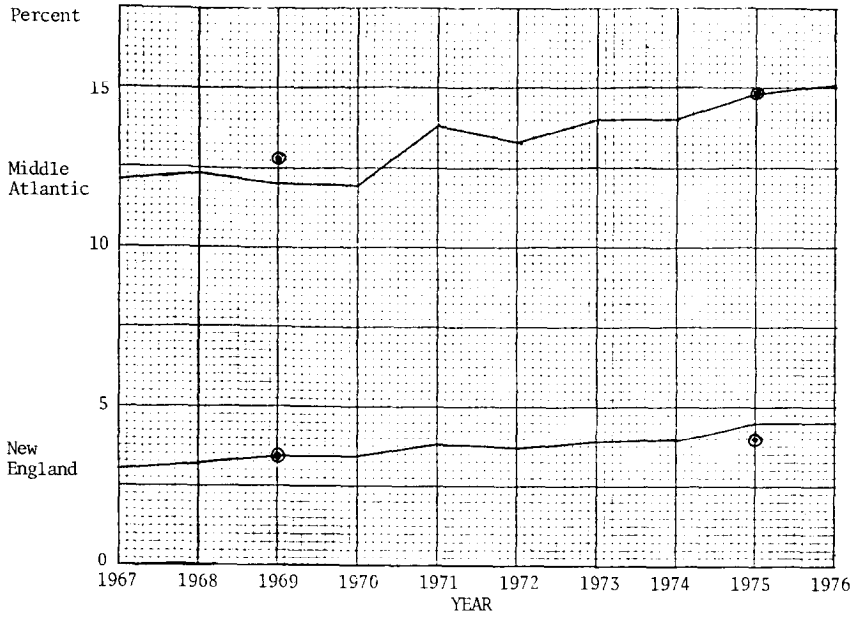


FIGURE 2

MODEL ESTIMATES BASED ON CPS OF THE PERCENT OF TOTAL POOR CHILDREN IN THE NORTH CENTRAL REGION, BY INCOME YEAR AND DIVISION (1970 Census and 1976 SIE Estimates Shown Circled)

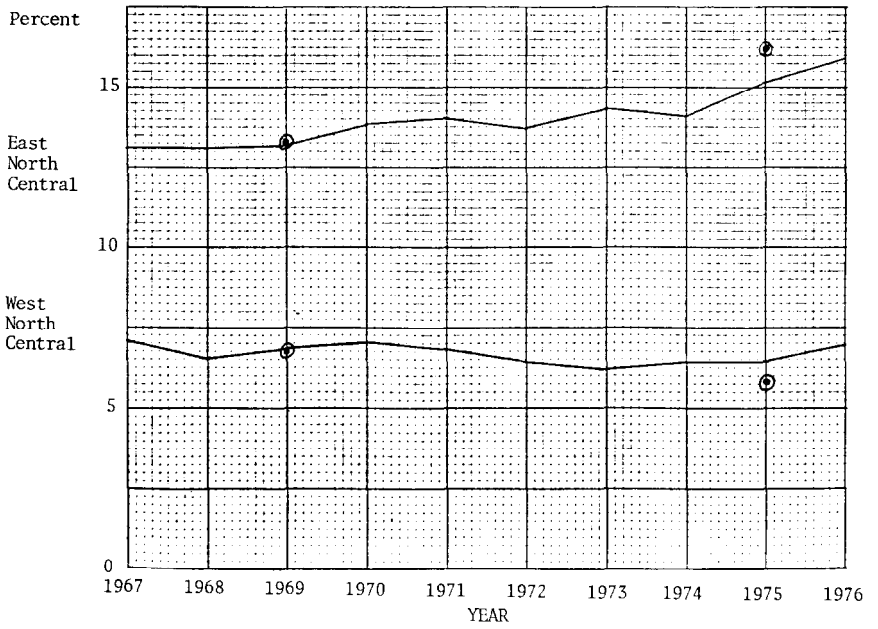


FIGURE 3

MODEL ESTIMATES BASED ON CPS OF THE PERCENT OF TOTAL POOR CHILDREN IN THE SOUTH REGION, BY INCOME YEAR AND DIVISION (1970 Census and 1976 SIE Estimates Shown Circled)

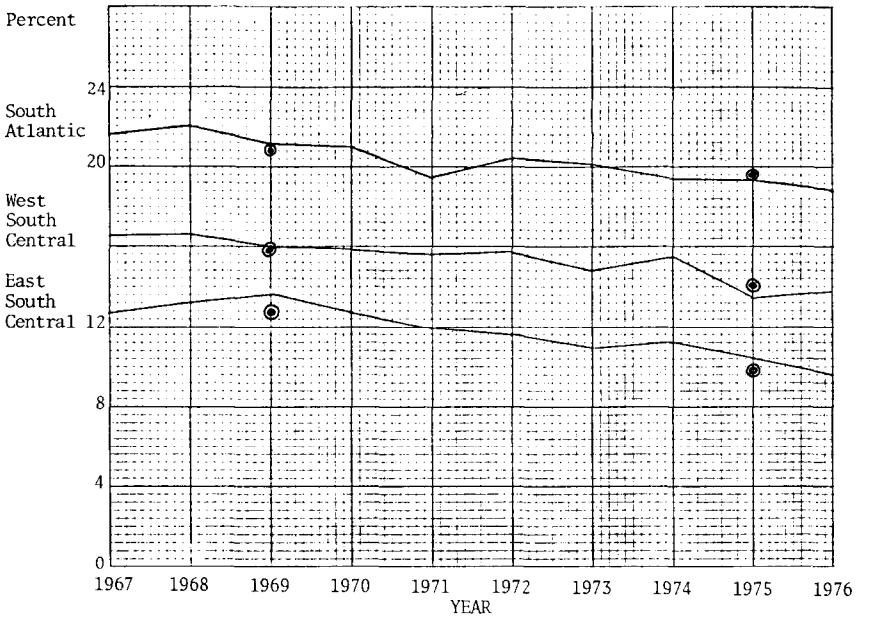


FIGURE 4

MODEL ESTIMATES BASED ON CPS OF THE PERCENT OF TOTAL POOR CHILDREN IN THE WEST REGION, BY INCOME YEAR AND DIVISION (1970 Census and 1976 SIE Estimates Shown Circled)

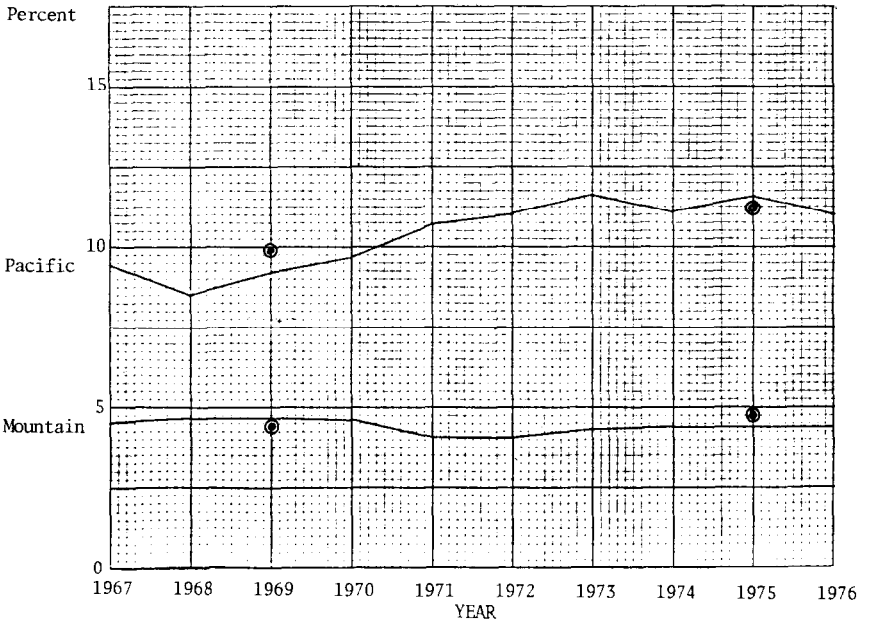


TABLE 1

Percent of Children Age 5-17 in Poverty Families According to 1970 Census, SIE, and Regression Modeln

Divisions, Regions, and States	1969 Estimates 970 Censu	975 Estimate SIE	Regression Model
UNITED STATES, TOTAL			
NORTHEAST			
New England			
Maine	14.2	15.3	14.2
New Hampshire	7.7	10.3	10.5
Vermont	11.4	17.8	11.9
Massachusetts	8.4	9.3	10.6
Rhode Island	11.0	10.5	11.8
Connecticut	7.2	8.4	9.6
Middle Atlantic			
New York	12.2	13.1	13.8
New Jersey	8.7	11.6	10.2
Pennsylvania	10.6	12.6	10.9
NORTH CENTRAL			
East North Central			
Ohio	9.8	11.6	11.8
Indiana	9.0	9.6	10.8
Illinois	10.7	15.1	10.8
Michigan	9.1	11.3	11.2
Wisconsin	8.7	9.4	9.6
West North Central			
Minnesota	9.5	9.1	9.7
Iowa	9.8	7.9	8.2
Missouri	14.8	14.7	14.8
North Dakota	15.7	11.5	10.4
South Dakota	18.3	13.1	15.3
Nebraska	12.0	10.1	10.3
Kansas	11.5	8.6	10.2
SOUTH			
South Atlantic			
Delaware	12.0	10.4	12.3
Maryland	11.5	10.7	11.2
District of Columbia	23.2	15.7	17.8
Virginia	18.2	13.7	15.0
West Virginia	24.3	18.9	18.2
North Carolina	24.0	17.8	20.2
South Carolina	29.1	23.9	23.4
Georgia	24.4	21.3	20.9
Florida	18.9	21.6	16.6

TABLE 1
(Continued)

Percent of Children Age 5-17 in Poverty Families According to 1970 Census, SIE, and Regression Model

Divisions, Regions, and States	1969 Estimates 1970 Census	1975 Estimates	
		SIE	Regression Model
UNITED STATES, TOTAL (continued)			
SOUTH CENTRAL			
East South Central			
Kentucky	25.1	21.4	20.2
Tennessee	24.8	20.5	20.2
Alabama	29.5	15.9	23.1
Mississippi	41.5	32.6	32.2
West South Central			
Arkansas	31.6	21.4	23.8
Louisiana	30.1	22.9	23.8
Oklahoma	19.5	14.6	16.2
Texas	21.5	20.5	17.7
WEST			
Mountain			
Montana	12.9	12.5	10.8
Idaho	12.0	11.0	10.5
Wyoming	11.2	8.6	8.2
Colorado	12.3	10.7	10.7
New Mexico	26.3	26.0	21.2
Arizona	17.5	16.8	16.1
Utah	10.0	8.0	9.4
Nevada	8.8	11.0	9.8
Pacific			
Washington	9.3	10.0	10.2
Oregon	10.3	8.4	10.2
California	12.1	13.8	12.5
Alaska	14.6	6.4	6.9
Hawaii	9.7	9.6	9.8

TABLE 2a

Comparison of Reinterview, Model, and SIE Estimates of Children 5-17 Years Old In Poverty Families by State (see text for explanation)

Comparison of Reinterview to SIE	Comparison of Model to SIE	
	States with Model Estimate Less than SIE	States with Model Estimate Greater than SIE
States with reinterview less than SIE	12	10
States with reinterview greater than SIE	10	18

NOTE: One State is omitted because of an estimate of no change in reinterview.

TABLE 2b

Comparison of Reinterview, Model, and Adjusted SIE Estimates of Children 5-17 Years Old In Poverty Families by State (see text for explanation)

Comparison of Reinterview to SIE	Comparison of Model to Adjusted SIE	
	States with Model Estimate Less than Adjusted SIE	States with Model Estimate Greater than Adjusted SIE
States with reinterview less than SIE	15	7
States with reinterview greater than SIE	8	19

NOTE: Two States are omitted: one with an estimate of no change in reinterview, and the other with an estimate of no difference (within 0.5 percent) between the model and SIE.

ESTIMATES OF INCOME FOR SMALL PLACES

The third application combines elements of the regression-sample method with the James-Stein estimator, mentioned earlier in relation to (8). Although the techniques again belong to those associated with small area estimation, their use in this case actually resulted in a greater reliance upon sample data than the procedures originally followed.

The Census Bureau provides the Department of the Treasury with current estimates of per capita income and population for approximately 39,500 units of local government participating in the Revenue Sharing Program. In general, these estimates represent an updating of census values by factors derived from administrative data. A significant exception occurred for the roughly 15,000 places of size under 500 persons, where the 1970 census values for county PCI were substituted as base figures for these places in preparing the first sets of estimates for income year 1972. The rationale for this substitution arose from the magnitude of sampling error in the 1970 census 20-percent sample estimates; for example, the coefficient of variation for PCI in the 1970 census was about 30 percent for places with population of 100 persons.

This situation falls rather easily into the framework constructed by Ericksen: sample estimates (from the census) are available for the variable of interest, and there is a presumed relationship to a predictor variable, the county PCI. Two other variables could also be added to the analysis: the value of owner-occupied housing obtained in the 1970 census (a 100-percent housing item) and the adjusted gross income per exemption from Internal Revenue Service data for 1969, although usable data were available for only a subset of the places in each case.

The other notion incorporated into the estimation, that of combining the sample and regression estimates, appeared in the two preceding examples, but in either instance the CPS data were unable to reduce appreciably the error of the estimates. In the case at hand, however, the contribution of the sample data was potentially significant. For example, a cursory examination of sample estimates for these places compared to the county values of PCI revealed a considerable number outside the usual range of sampling error, some by large multiples of the standard error. In consideration of this, the James-Stein estimator was adapted to this problem to provide a means to combine the sample and regression estimates.

Efron and Morris (for example, (1972), (1973), and (1975)) have argued and illustrated the potential utility of the James-Stein estimator to diverse problems in multivariate estimation. The estimator can be motivated by the observation that for k sample estimates Y_i with equal variances D and means θ_i, and for any set of fixed constants P_i, the estimator Z_a of θ_i,

$$Z_a = P + a\,(Y - P) \qquad (11)$$

for fixed a has its expected square error

$$R(\theta, Z_a) = E_\theta((\theta - Z_a)^T(\theta - Z_a)) \tag{12}$$

minimized by the choice

$$a = \frac{A}{A + D} \tag{13}$$

for

$$A = (\theta - P)^T(\theta - P)/k \quad . \tag{14}$$

With this a, the value of (12) is kaD, less than the value of (12), kD, for Y itself. The James-Stein estimator for $k \geq 3$, is simply (11) with $\hat{a}$ estimated from the data as

$$\hat{a} = 1 - (k-2)D/S \tag{15}$$

for

$$S = (Y - P)^T(Y - P) \quad . \tag{16}$$

Thus, differences between the sample estimates Y and prior estimates P are assessed to determine how much weight the sample data should receive: if P fits poorly, the sample estimates receive more weight than when differences are small relative to sampling error.

Efron and Morris have extended and refined the estimator. One suggestion of theirs, critically important in this application, effects a compromise between overall error, as in (12), and the error of individual components. The modification is to use the sample data to limit the reliance upon the prior estimates by constraining the final estimates to lie within some specified distance, usually a fixed multiple of the standard error, of the sample estimate for each component of 8. Thus, the estimator shrinks the data toward the prior estimates and maintains most of the resulting overall advantage, while guarding against unacceptably large risk to any individual component.

The program of estimation in this application may be outlined as follows:

1. Fitting a regression equation to the census sample estimates.

2. Measuring the goodness of fit between the regression equation and the sample data, taking into account the contribution of sampling error to the observed differences.

3. Forming a weighted estimate of the sample and regression estimates, letting the weights reflect the relative fit of the

regression and the sampling error of the sample estimate.

4. Constraining each weighted combination to lie within one standard error of the sample estimate.

For purposes of estimation, Y_i was expressed as the logarithm of the sample estimate. (Since the sample estimates have approximately a constant coefficient of variation for a given sample size, the logarithm of the sample estimate has approximately a constant variance for a given sample size.) In turn, all independent variables were similarly converted into logarithmic form. Separate regressions for each State and each of the two groups of places under 500 population and of 500-999 were fitted; reduced equations were employed for places lacking housing or IRS data. The strategy was to estimate A as in (13) and to reflect this value both in combining the regression and sample data and in weighting the regression.

The regression estimates were

$$\hat{Y} = X (X^T W X)^{-1} X^T W Y \qquad (17)$$

with $W_{ii} = (D_i + \hat{A})^{-1}$, where D_i is the sampling variance of Y_i and $\hat{A} \geq 0$ was determined iteratively as the unique solution to

$$(Y-\hat{Y})^T W (Y-\hat{Y}) = n - p \qquad (18)$$

for p, the rank of X, and n, the number of Y_i's. (If no positive solution existed, $\hat{A}$ was set to 0.) Each value was then estimated as

$$\delta_i' = \delta_i + (D_i)^{\frac{1}{2}} \quad \text{if } \delta_i > Y_i + (D_i)^{\frac{1}{2}} \qquad (19)$$

$$\delta_i' = \delta_i \quad \text{if } \left| \delta_i - Y_i \right| < (D_i)^{\frac{1}{2}} \qquad (20)$$

$$\delta_i' = \delta_i - (D_i)^{\frac{1}{2}} \quad \text{if } \delta_i < Y_i - (D_i)^{\frac{1}{2}} \qquad (21)$$

where

$$\delta_i = \hat{Y}_i + \frac{\hat{A}}{\hat{A}+D_i} (Y_i - \hat{Y}_i) \quad . \qquad (22)$$

The $\hat{A}$ obtained through the solution of these equations measures an average lack of fit between the regression and true values. Table 3 gives values of $\hat{A}$ from the estimation for places of population under 500 in States with the largest number of such places, and, similarly, Table 4 shows results for places of population 500-999. Roughly, $\hat{A}$ is in units

TABLE 3

Estimated $\hat{A}$ for Places with 20-Percent Sample Estimates of Population Less than 500

STATES		Regression Equation			
		County	County and Tax	County and Housing	County, Tax, and Housing
	a.	States with More than 500 Places in Class			
Illinois		.036	.032	.019	.017
Iowa		.029	.011	.017	.000
Kansas		.064	.048	.016	.020
Minnesota		.063	.055	.014	.019
Missouri		.061	.033	.034	.017
Nebraska		.065	.041	.019	.000
North Dakota		.072	.081	.020	.004
South Dakota		.138	.138	.014	--
Wisconsin		.042	.025	.025	.004
	b.	States with 200-500 Places in Class			
Arkansas		.074	.036	.039	.018
Georgia		.056	.081	.067	.114
Indiana		.040	.012	.003	.000
Maine		.052	.015	- -	- -
Michigan		.040	.032	.028	.023
Ohio		.034	.015	.004	.004
Oklahoma		.063	.027	.049	.036
Pennsylvania		.020	.018	.016	.011
Texas		.092	.048	.056	.040

NOTE: A dash (--) indicates that the regression was not fitted because of too few observations.

TABLE 4

Estimated $\hat{A}$ for Places with 20-Percent Sample Estimates of Population 500-999

STATES		Regression Equation			
		County	County and Tax	County and Housing	County, Tax, and Housing
	a.	States with More than 250 Places in Class			
Illinois		.032	.023	.012	.008
Indiana		.017	.014	.007	.009
Michigan		.019	.014	.005	.008
Minnesota		.056	.040	.021	.007
New York		.052	.015	.028	.006
Ohio		.024	.010	.005	.000
Pennsylvania		.035	.025	.015	.026
Wisconsin		.039	.030	.014	--
	b.	States with 100-250 Places in Class			
Iowa		.017	.005	.016	.004
Kansas		.025	.010	.014	.008
Maine		.022	.021	--	--
Missouri		.042	.019	.011	.013
Nebraska		.027	.007	.008	.008
Texas		.050	.017	.013	.012

NOTE: A dash (--) indicates that the regression was not fitted because of too few observations.

equivalent to squared relative error, so that .040 corresponds to about a 20 percent average error. A place of 225 persons has a c.v. of about 20 percent also; thus, Table 3 indicates that, for places of this size, (22) weights the sample data more heavily than the regression estimate in the majority of cases for the county-only equation. When other variables were available for inclusion, the values of A were generally considerably lower, indicating a substantially better fit.

Two further investigations of the performance of the James-Stein estimator were made in this application. In 1973, the Bureau of the Census conducted special censuses of a random sample of places, some of which had 1970 populations under 1000. These censuses collected 1972 income on a 100-percent, rather than sample, basis. Table 5 displays the comparison between the special census results for places falling into this category and alternative estimates based upon updating county or place sample estimates from the 1970 census or the James-Stein estimates. Thus, the table offers only an indirect assessment of the relative merits of the three base figures, as the resulting estimates for 1972 were equally affected by error in the common updating factor. Of the three, the set based upon the James-Stein estimates shows smaller average error (measured as absolute percent difference) and appears considerably better than the county values. (The tendency for the 1972 special census estimates to appear lower than the other estimates also occurs for the remaining special censuses for larger places and probably reflects principally the consequences of not imputing income for non-response in the processing of the special census returns.)

A second investigation served to demonstrate that the true values for places of this size differed in general from their respective county values, and that the James-Stein estimator was a useful mechanism to achieve a reduction in sampling error while preserving much of the actual variation. A sample of places with usable IRS estimates was sorted by adjusted gross income per exemption and then aggregated in order into groups of ten. The census sample estimate for per capita income of the groups as a whole was thus considerably more accurate than for the individual components and could be taken as an accurate estimate for the group. Table 6 displays comparisons of the sample estimates for these groups with aggregated estimates using the James-Stein or the county estimates. According to each measure of spread considered in the table, the aggregated values of the James-Stein estimates more closely matched the sample estimates than did the county values, by a substantial margin, in fact.

The Census Bureau has incorporated the James-Stein estimates as base figures into its computation of per capita income for 1974 and subsequent years. This represents perhaps one of the largest, if not the largest, formal applications of this estimator in a Federal statistical series.

TABLE 5

Comparison of Selected 1972 PCI Estimates to 1972 Special Census PCI Values

SPECIAL CENSUS AREAS	1972 Special Census PCI	1972 PCI Estimates and Percent Difference from Special Census PCI					
		Census Base		James-Stein Base		County or MCD Base	
		1972 Estimate	Percent Difference[d]	1972 Estimate	Percent Diffemnce[d]	1972 Estimate	Percent Difference[d]
a. 1970 Census Weighted Sample Population Less than 500							
Newington, GA	2,019	2,225	10.2	2,302	14.0	2,279	12.9
Foosland Village, IL	2,899	2,771	4.4	3,199	10.3	3,796	30.9
Bonaparte, IO	2,331	3,126	34.1	2,942	26.2	2,542	9.1
McNAry, LA	2,333	2,303	1.3	2,527	8.3	2,908	24.6
Freeborn Village, MN	2,741	3,693	34.7	3,338	21.5	2,922	6.6
Spruce Valley Twp, MN	2,430	1,894	22.1	1,949	19.8	2,076	14.6
Jacksonville, MO	2,723	2,338	14.1	2,611	4.1	3,233	18.7
Thayer, NE	2,742	2,245	18.1	2,870	4.7	3,452	25.9
Benton Town, NH	1,788	2,874	60.7	3,284	78.7	3,570	99.7
Nora Township, ND	1,780	2,629	47.7	2,754	54.7	3,476	95.3
Riga Township, ND	1,454	2,749	89.1	2,411	65.8	2,711	86.5
Deer Creek, OK	2,451	2,493	1.7	2,673	9.1	2,762	12.7
Dudley Borough, PA	2,446	2,168	11.4	2,411	1.4	2,608	6.6
Brookings Township, SD	3,132	3,400	8.6	3.309	5.7	2,395	23.5
Valley Township, SD	1,574	1,946	23.6	1.972	25.3	2,114	34.3
Bryant Township, SD	2,412	1,120	53.6	2,158	10.5	2,695	11.7
Parrish Town, WI	3,567	5,399	51.4	4,079	14.4	2,721	23.7
Average, all areas	--	--	28.6	--	22.0	--	31.6
b. 1970 Census Weighted Sample Population Between 500 and 999							
Caswell Plantation, ME	1,946	2,656	36.5	2,490	28.0	2,646	36.0
Sugar Creek Township, MO	2,224	2,035	8.5	2,315	4.1	2,018	9.3
Jeromesville, OH	3,329	3,081	7.4	3,418	2.7	3,072	7.7
Rush Township, OH	2,241	2,545	13.6	2,619	16.9	2,546	13.6
Dennison Township, PA	3,521	4.411	25.3	4.095	16.3	4,430	25.8
Manor,Tx --	2,062	2,746	33.2	2,765	34.1	2.740	32.9
Derby Center, VT	2,968	2,694	9.2	2,754	7.2	2.675	9.9
Average, all areas	--	--	19.1	--	15.6	--	19.3

NOTE: "d" = absolute percent difference. "Average, all areas," is average of absolute percent differences.

TABLE 6

Relation of 1969 Revised Estimates and 1969 County Averages to 1970 Census Sample Estimates for Groups of Ten

(for places with the ratio of 1969 IRS exemptions to 1970 census population between .8 and 1.1)

Relation to 1969 Sample Estimates	1969 Revised Estimates		1969 County Averages	
	Number	Percent	Number	Percent
Total Groups	212	100.0	212	100.0
Within 10% of Sample PCI	172	81.1	111	52.4
Outside 10% of Sample PCI	40	18.9	101	47.6
Within One Standard Error	149	70.3	61	28.8
Between 1 and 2 Standard Errors	28	13.2	60	28.3
Outside 2 Standard Errors	35	16.5	91	42.9
Closer to Sample PCI	154	72.6	58	27.4

THE PROBLEM OF TWO REGRESSIONS

The regression paradox or the problem of two regressions appears in most texts on linear regression. If we restrict the problem temporarily to univariate regression, including a constant term, the least squares estimate of the regression of Y on Z is based on the coefficient

$$\hat{b}_1 = \frac{\sum_i (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_i (Z_i - \bar{Z})^2} \tag{23}$$

for

$$\bar{Z} = \sum_i Z_i/n \tag{24}$$

$$\bar{Y} = \sum_i Y_i/n \tag{25}$$

whereas the regression of Z on Y gives the coefficient

$$\hat{b}_2 = \frac{\sum_i (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_i (Y_i - \bar{Y})^2} \tag{26}$$

when there is a perfect linear relationship between Y and Z, $\hat{b}_1\hat{b}_2 = 1$, as logic might seem to dictate. In all other situations, however, the product $\hat{b}_1\hat{b}_2$ is less than 1, which is the root of the so-called "regression paradox." In the presence of residual error, (23) and (26) determine two distinct regression lines intersecting at the joint means, and their different interpretation requires care.

To illustrate the implications of this problem to small area estimation, consider the case where Z is a sample estimate of X, and Y is an indicator for X. One approach to determine X on the basis of Y is to follow Ericksen's suggestion to form the regression of Z on Y, computing a coefficient for Y using (26). Our attitude toward this procedure might change, however, if we were to learn that Y was in fact a sample estimate for X. We would find generally that the coefficients estimated from (26) would not tend toward the value 1, as the principal of unbiased estimation would require, but in fact to a lesser value. (We would obtain an expected value of 1 if we could substitute the actual X for Z in (23).) To see what this lesser value is, suppose that we let the sampling error of Z go to zero, for the sake of argument. We would find a convergence of (26) to approximately the value of "a" given earlier in (13) as the optimal weight to combine sample and prior information (in this case, the mean) to minimize mean square error. (In formula (13), A assumes the role of the true variability of X and D the sampling error of Y.) Thus, the regression approach leads to a shrinkage of the sample estimates Y-toward the mean very much in the spirit of the James-Stein estimator, although by an entirely different route.

As an illustration of this phenomenon of shrinkage, let us return to the first example of population estimation. For the values of c, the growth of State population-relative to the national as in (4) for the decade 1960-1970. the regression coefficients for the 51 States and District of Columbia are .324, . 374, and .177, for school enrollment, labor force size, and number of tax returns, respectively. This set of coefficients is employed in the most recent revision of the ratio-correlation method (to a greater precision than shown here, however). Their sum, .875, is less than unity. Consider the consequences of reweighting the regression: using the square root of 1960 population as a weight,the coefficients become .334, .435, and .124; weighting proportional to population (included in Ericksen's proposal) gives .371, .483, and .058. The sum of the second set is .893; that of the third, .912. Thus, the shrinkage effect, the summation of the coefficients to a value less than one, is reduced somewhat as larger States receive increased weight. An interpretation of this effect is that the better fit of the regression to the larger States supports less shrinkage than for smaller.

This last example was chosen only to suggest that linear regression includes a shrinkage effect that works to reduce mean square error and runs counter to the notion of unbiasedness. Furthermore, if some specific subsets of units favor less shrinkage than others, the regression equation will express a compromise between the different degrees of shrinkage. In these cases, the question of weighting must be considered carefully. The possibility exists, moreover, for estimators that would explicitly accomplish varying degrees of shrinkage for different groups.

POST-CENSAL ESTIMATION OF POPULATION (REVISITED)

As described earlier, Ericksen proffered the regression-sample method as a means to counter possible obsolescence of past relationships applied to measure the present. This section will illustrate that multivariate methods in some applications may enable the study of the structure of the same past relationships and permit inferences about the approximate degree of their persistence. (The following discussion addresses the actual merit of Ericksen's proposal only obliquely, however, since the models will be analyzed on the level of States rather than PSU's. Furthermore, the computations carried out here are for the purposes of exploration only and are insufficient to constitute a complete methodology.)

Subsequent to Ericksen's original work on population, circumstances have limited the field of possible indicators of population change to statistics on school enrollment, labor force size, and number of tax returns. Recent instability due to changes in abortion laws has virtually eliminated the utility of births as an indicator of general population change, although this variable had been demonstrably effective in predicting-change during the 1960-1970 decade. Similarly, fluctuations in the data on automobile registrations, never a strong predictor, have also resulted in its exclusion from current estimates. The Census Bureau has altered the methodology in another important respect: Medicare data are now used directly to estimate the component of the population age 65 and over, and consequently the ratio-correlation method is now used only to predict the population under 65 years old.

School enrollment, labor force size, and number of tax returns correlate almost identically with population change (for the component of the population under 65) for the 1960-1970 decade, with values .955, .952, and .954, respectively. Rather than weighting the three equally, however, the regression coefficients for the decade are .324 for school enrollment, .374 for labor force, and .177 for tax returns. As mentioned in the preceding section, weighting the regression by the square root of population or by population further reduces the coefficient on tax returns. A general explanation for unusual coefficients is near-colinearity among the variables, which can lead to instability in the estimated coefficients. In this case, however, colinearity has a relatively mild effect upon the stability of the coefficients computed from the census data, and the differences between the resulting coefficients and an equal weighting cannot be ascribed to this factor alone. The analysis that follows suggests why the coefficients take this form.

The linear regression of population change on the three variables constitutes one measure of their interrelationship. Other multivariate techniques, in particular principal component analysis, can be useful for exploring the structure of the independent variables apart from their relationship to the dependent variable. The three-dimensional space determined by the three independent variables may have its points specified by the values of the individual variables. Equivalently, the points of this space may be measured in relation to other component dimensions arising as linear combinations of the original variables. One such representation, the principal components, establishes dimensions that are uncorrelated according to the sample covariance matrix. In addition, these dimensions may be specified to represent progressively the largest remaining component of variation subject to the constraint of zero correlation with the preceding principal components. Hence, in a three dimensional space, the first principal component represents the direction of maximum variation, and the third corresponds to the least variation. Algebraically, the principal components are the eigenvectors of the sample covariance matrix, and the corresponding eigenvalues measure the variance of the original variables along the dimension of the space determined by the respective eigenvector.

The top half of Table 7 gives the principal components for the 1960-1970 decade for the three predictor variables. The first component represents effectively an average of the three variables, suggesting its origin in their common relation to population change. The second, with an eigenvalue only a twenty-fifth of the first. contrasts labor force and school enrollment, with tax returns playing a'minor part. The third component, the dimension of least variation, has an eigenvalue only about half of the second, measures the tax return variable against the average of the other two.

This description of the variables, together with the tendency of the regression to favorthecombination of labor force and school enrollment over tax returns, suggests the following interpretation: the second component reflects a possible demographic phenomenon, that the labor force and school enrollment variables are indicators of two separate elements of the population, and their combination is able to represent the entire population efficaciously. The small eigenvalue of the third component indicates that the tax variable represents generally an average of the other two, although the regression clearly favors the combination of school enrollment and labor force as a prediction of population change.

TABLE 7

Principal Components of Indicators

Indicators	Principal Components		
	1st	2nd	3rd
	1960-1970		
School enrollment	.61	-.62	-.48
Labor force	.53	.78	-.32
Tax returns	.58	-.06	.81
Eigenvalue	.0541	.0021	.0011
	1970-1976		
School enrollment	.33	-.81	-.48
Labor force	.72	.54	-.43
Tax returns	.61	-.20	.77
Eigenvalue	.0221	.0018	.0006

Table 8 presents evidence in support of this interpretation. In the upper section of the table, the two-variable regressions of population change on school enrollment and labor force size indicate that school enrollment dominates the prediction of age 5-17 and contributes equally with labor force for 0-4, while being less effective for 18-44 and entirely negligible, once labor force is considered, for 45-64. The three-variable regressions in the lower part of the table show the potential of tax returns as a general indicator but also its inability to dominate both labor force and school enrollment for any age group. (The shrinkage effect described in the preceding section is apparent in these separate regressions, but to the least extent for the age group 5-17. The small shrinkage applied for this group may be attributed to the excellent fit of the regression here.) (The computations for age groups are only illustrative and are based simply upon published census counts without the necessary adjustments for the institutional population, etc., in the ratio-correlation method.)

To address the issue of possible change in the regression relationships since the 1970 census, the lower half of Table 7 gives the principal components of the 1970-1976 variables. The reduced coefficient on school enrollment in the first principal component is a direct consequence of the smaller variation among States for this indicator. (During the 1960-1970 decade, the average variation among States ranged from 13 percent for labor force to 15 percent for school enrollment. For the period 1970-1976, however, the average variation in school enrollment is only 6 percent, whereas tax returns vary by 9 percent and labor force by 11 percent.) We find substantially the same alignment of components as for the 1960-1970 decade. The second principal component still may be understood to represent the difference in relative growth between the school-age population and the labor force. The second eigenvalue here is now larger relative to the first eigenvalue than previously; it is now almost a tenth of the first. The third eigenvector, which still contrasts tax returns with the average of the other two, has remained relatively small, with an eigenvalue only 1/40th of the first. close to the ratio between these two eigenvalues for 1960-1970.

These last observations provide a limited assurance that the relationships established during the 1960-1970 decade have largely continued to hold. If either labor force or school enrollment were to have deteriorated substantially in its ability to predict their respective components of population change, this would be reflected in a larger third eigenvalue. Hence, the tax data as a general indicator suggest the demographic relations observed earlier have persisted. (Some adjustment to the weights might be argued, however, in terms of the declining proportion of the total population under age 17.)

Should the tax variable, which serves to confirm the relationship between school enrollment and labor force, receive increased weight? The analysis based upon principal components does not fully resolve this question. Unfortunately, a linear regression incorporating CPS data would also be quite unsuccessful in answering this, since the extremely small variation in the third component, which represents the dimension at issue, forces an extremely high variance on the estimated coefficient from sample data. At best, the sample-regression method represents a tool of possible future use for this question, but other techniques appear to be required as well.

TABLE 8

Regression Coefficients for Population Growth, 1960-1970, for States

Indicators	Age				
	Total	0-4	5-17	18-44	45-64
	Two-Variable Regression				
School enrollment	.421	.390	.925	.251	.005
Labor force	.449	.442	-.019	.508	.851
	Three-Variable Regression				
School enrollment	.324	.374	.856	.231	-.241
Labor force	.374	.429	-.071	.492	.663
Tax returns	.177	.030	.124	.036	.446

NOTE: Computations for age groups for illustration only and not consistent with current methodology.

REFERENCES

Efron, Bradley, and Morris, Carl (1972), "Limiting the Risk of Bayes and Empirical Bayes Estimators -- Part II: The Empirical Bayes Case," Journal of the American Statistical Association, 67, 130-9.

_____ and Morris, Carl (1973), "Stein's Estimation Rule and Its Competitors -- an Empirical Bayes Approach," Journal of the American Statistical Association, 68, 117-30.

_____ and Morris, Carl (1975), "Data Analysis Using Stein's Estimator and Its Generalizations," Journal of the American Statistical Association, 70, 311-9.

Ericksen, Eugene P. (1973), "A Method of Combining Sample Survey Data and Symptomatic Indicators to Obtain Populatjon Estimates for Local Areas," Demography, 10, 137-60.

_____(1974), "A Regression Method for Estimating Population Change for Local Areas," Journal of the American Statistical Association, 69, 867-75.

Fay, Robert E. (1978), "Problems of Nonsampling Error in the Survey of Income and Education: Content Analysis," in Proceedings of the Social Statistics Section, 1977, Part I, American Statistical Association, Washington, DC.

Ginsberg, Alan, and Grob, George (1977), "Uses of Data from the Survey of Income and Education for Policy Analysis," in Proceedings of the Social Statistics Section, 1977, Part I, American Statistical Association, Washington, DC.

U.S. Bureau of the Census (1976), Current Population Reports, P-25, No. 640 "Estimates of the Population of States with Components of Change: 1970 to 1975," U.S. Government Printing Office, Washington, DC.

U.S. Bureau of the Census (1978), "Assessment of the Accuracy of the Survey of Income and Education," submitted to Congress on April 25, 1978, by the Secretaries of Commerce and of Health, Education, and Welfare.

Discussion

Eugene P. Ericksen

DEFINING CRITERIA FOR EVALUATING LOCAL ESTIMATES

The selection of criteria for evaluating local estimates is at once a statistical and political issue. The statistician first of all wants a methodology for evaluating errors and then wants to verify that the selected set of estimates has a smaller average error than any competitive set and that there are no indications of systematic bias for particular subgroups of local areas. The policy-maker naturally wishes to have statistically satisfactory estimates, but also must value presentability since s/he will need to defend the estimates before legislative groups, local critics, and the general public. Unfortunately, the best statistical estimates are sometimes difficult to present to a nonstatistical audience. More often, the policy-maker is forced by legislative demands or other requirements to produce and use "the best available estimate" which either does not meet accepted statistical standards or has not been subjected to statistical evaluation. The Federal estimates of population growth since 1970 which are used to allocate revenue sharing funds to local jurisdictions are an example of this. Congress specified that estimates be computed for about 39,500 localities, and the Census Bureau had to produce the estimates, even though it had not developed and tested a method for doing so.

The procedure of synthetic estimation provides a method of computing local estimates which would not otherwise be available. It has been used to give local estimates of dilapidated housing, unemployment, drug-taking behavior, and vacant housing. The alternative to these estimates was either nothing or a set of estimates already shown to be fallible. Unfortunately, the accuracy of synthetic estimates has not usually been assessed and we don't have a systematic method which could tell us how inaccurate or biased the estimates might be. On the other hand, for the regression-sample data method there are already usable, though imperfect, methods of evaluating errors. Although these methods can usually tell us which of several sets of estimates are better, they cannot specify the level of error precisely. Moreover, the methods are complex and sometimes require assumptions which are statistically acceptable but difficult to sell politically. There seems to be a belief that a good local estimate incorporates information collected from that jurisdiction only and does not make use of information borrowed fromm other local areas as is done in the regression-sample

data estimates (Ericksen 1974). Nonetheless, it seems to me that synthetic estimates could be made more acceptable and more complex estimates salable if statisticians emphasized the assessment of errors as the most important criterion to evaluate the methodology of a set of local estimates. Top priority should be given to research strategies designed to improve the methodology of error estimation. Fortunately, Bob Fay has made steps toward that goal.

I feel that the synthetic procedure is of questionable validity. The estimates have the unfortunate characteristic of "shrinking" estimates toward the mean of all areas. For a variable where characteristics of local areas are important, synthetic estimates might be very poor. Such a variable might be usage of a drug which is available in some areas but not others. This is because individual level characteristics like age, race, and sex are typically used to compute synthetic estimates, and these characteristics are weakly related or unrelated to the volume of drugs on a local market. Moreover, if a synthetic estimate is to be used to identify extreme cases like local areas with particularly high unemployment rates? the shrinking is a decisive liability. While there may be estimating situations where the synthetic procedure gives accurate results, there are usually also reasons to disbelieve their accuracy. Therefore the acceptability of a set of synthetic estimates should be based on an evaluation of errors. I suggest that this can often be done using the sample data on which the synthetic estimate is based.

Maria Gonzalez has presented an overview of some of the better known applications of synthetic estimates. Sane of these applications have been important to users, such as the set of estimates correcting the numbers of housing units classified as vacant in the 1970 Census. Her paper indicates the versatility of synthetic estimation, and I think it is clear that the methodology will be used in important ways in the years to come. While she did not indicate a method by which the accuracy of estimates can be ascertained without resorting to census counts of the variable in question, she and I have worked on the problem. We did this for the set of unemployment estimates for 122 large metropolitan areas which she has reported here and given more extensive information about elsewhere (Gonzalez and Hoza 1978).

Many synthetic estimates, particularly those derived from Census or CPS data, are based on large sample calculations. For these, unbiased estimates of the characteristic in question can be computed from the survey data for the sample psu's. These estimates have large variances, but unless the number of psu's is small, the estimates can be used as a standard for accuracy. The series of synthetic and competitive estimates can be compared to the psu sample estimates. The set of estimates most highly correlated to the sample estimate is judged most accurate. This assumes that the sample estimates have only random errors.

In the unemployment application discussed by Gonzalez, the main competitor to the synthetic estimates was the set of "70-step" esti-

mates computed by the Department of Labor. We correlated various sets of synthetic estimates and the 70-step estimates with the 122 sample estimates and found that the 70-step estimates were consistently more strongly related to the sample estimates of unemployment. We then used the occupation-race-sex synthetic estimate, thought to be the best synthetic estimate, and the 70-step estimate as independent variables in regression with the sample estimates as the dependent variable, following the methodology of the regression-sample data technique. There we found the regression weights of the 70-step estimates to be considerably larger than those of the synthetic estimates. Fortunately, the synthetic estimates contained some independent information. The regression estimates computed with 70-step and synthetic estimates as the two independent variables were more accurate than either the 70-step or synthetic estimates, particularly when outliers due to large sampling errors were removed (Ericksen 1975; Gonzalez and Hoza 1978).

With hindsight, we can see why the synthetic estimates of unemployment should be so poor. The variance of the synthetic estimates was very small, considerably smaller than either the variance of the 70-step estimates, the sample estimates, or the sample estimates after an estimate of the within-psu variance had been removed. This should have been an indicator of the shrinking problem. The synthetic procedure assumed that the unemployment rate was the same for all members of a given sex-race-occupational group in a region. For example, if the unemployment rate for steelworkers was high, this high rate was applied to all local areas. This unemployment rate was the result of economic problems in the steel industry which have led to the selective closing of plants. Bethlehem Steel, for example, is closing only some of its plants. A number of other steel plants have been closed in Youngstown, Ohio, but more are still working in Gary, Indiana. As a result, synthetic estimates computed for 1978 would give a misleading result indicating the unemployment rates to be overly similar in Gary and Youngstown. Because the 70-step estimates were sensitive to local fluctuations, they would again prove superior.

A key issue, then, is the accurate estimation of the within-psu error of the sample survey estimates. This is needed to establish the magnitude of errors of synthetic and other estimates as well as to evaluate the errors of estimates computed by the regression-sample data method. This estimation problem has been difficult, and its lack of solution prevents us from specifying a definitive answer to the important problem of assessing the errors of local estimates. Using only the synthetic and 70-step estimates and the sample data, we were unable to give accurate estimates of the mean squared errors of the various unemployment estimates. We were only able to rank order them in terms of accuracy.

It can be seen from Fay's discussion of the SIE estimates of the number of children in poverty that the accurate estimation of the within-psu variance is a continuing problem. In this case, Fay was unable to compute a direct estimate of the errors of regression

in 1975, although a complex and ingenious assessment of errors was eventually carried out. We faced a similar variance estimation problem in our work on 1960-70 population growth. We found that a few local units with extraordinarily large errors upset the stability of our within-psu variance estimates. These large errors appeared to be due to nonsampling errors, to the inclusion of special strata important nationally but found in only a few sample psu's, to poor estimates of the location of new construction, and in some cases, to pure chance. We found some improvement through a rejection of outliers routine (Ericksen 1975) but more research needs to be done on the estimation of within-psu error and its components.

Among the many issues usefully discussed in Fay's paper, there are two which deserve special attention. One is his delineation of sources of error in regression-sample data estimates, and the second is his application of the James-Stein technique. Both of these points suggest that the most fruitful applications of the regression-sample data technique will occur in estimating situations where sample estimates are available for all local units and explicit use can be made of the unbiased nature of the sample estimates.

It is recognized that errors in regression-sample data estimates arise due to structural errors in regression and to the presence of within-psu error. Fay correctly points out that errors also arise due to (1) differences between population regression equations for sampling units (psu's) and for- the units of-analysis (states or counties), (2) biases in the sample data. and (3) the weights used in the regression equation. I would like to underscore his argument by giving an example of how the first and third sources contributed to error in one application. The job was to compute estimates of 1960-70 population growth for 2,586 counties in 42 states. Symptomatic information was available for all counties and for psu's in the CPS sample. We estimated a regression equation using 444 CPS psu estimates as the dependent variable. Because some of the self-representing psu's were very large, much larger than the typical nonself-representing stratum, they were given larger weights. These weights were directly proportional to population size and hence to the sample sizes in the psu's. In this way, the weights were proportional to the expected accuracy of the psu sample estimates and we hoped to reduce the within-psu component of error by giving greater weight to the more reliable estimates. When the regression equation was applied to the 2,586 counties,we found the mean error to be 4.54 percent and 221 of the errors were 10 percent or greater. We then, as an experiment, proposed to eliminate the within-psu source of error entirely by usine 1960-70 Census figures for the 444 psu's as the denendent variable in the calculation of the regression equation. When we applied this regression equation to the 2,586 counties, we found to our surprise that the mean error was now 4.55 percent and that the number of errors of ten percent or greater had, in fact, risen to 234. How was this possible? We compared errors by size of county. We found that where the county population was 25,000 or greater, the errors were consistently and substantially reduced by the second equation.

For smaller counties, the large majority of all counties, the errors had increased. and these increases offset the decreases in the larger counties. It'should be clear that psu's are more similar to the larger counties, particularly those psu's given greater weights. As a result, our weighted equation based on psu's, when improved, increased the accuracy for psu's but decreased the strengths of the inferences to counties. We had made better estimates with less information.

A second point to be made is that we cannot directly assess the errors for local areas not included in the sample. More importantly, the presence of the sample survey information, as Fay has shown, can lead to further reductions in error. By applying the Stein-James methodology, he was able to compute optimal weights for regression and sample estimates and to reduce the errors below those obtained from either method. There are two quibbles I would like to make. The first concerns the assumption that the sample observations are drawn from a population with equal means and variances. Since our objective is to estimate the differences among local units, how do we sustain this assumption? Is it necessary to subdivide local areas into categories with similar means, and just how robust is the assumption?

The second quibble concerns the constraint that final estimates must lie within a specified distance, perhaps one standard deviation, of the sample estimates. If we assume that the within-psu errors are totally random, then we would expect the errors to have mean zero and to be normally distributed. As a result, there would always be a small subset of local areas which would have particularly bad sample estimates due to chance alone. As a result, the constraint would be particularly bad in these areas. If a constraint is necessary it is probably better practice to use the regression estimates rather than the sample estimates as the standard and to remove bad sample estimates from the equation. In the three applications I have worked on, estimating population growth, unemployment, and income, the regression equations have been considerably more accurate on average than the sample estimates.

This leads to a final point about within-psu errors. As Hogg (1974) has pointed out, outliers can have drastic effects on the calculation of a regression estimate. For regression-sample data estimates, outliers due to measurement error can be particularly damaging, even when their number is small. We have found a suitable way to identify these outliers and thus remove them from the equation (Ericksen 1975). We first computed a regression equation based on all cases, and then compared the regression and sample estimates. Those sample observations at a specified distance from the regression estimate, usually two standard deviations, were identified and removed from the sample. A second regression equation was then computed from the remainder and this equation was used to calculate the final estimates. Sizable reductions in the mean squared error were obtained by this technique which does not seem incompatible with the general idea of the James-Stein methodology. Moreover, if outliers

due to large within-psu errors were excluded, a more optimal set of weights between the sample data and regression estimates could perhaps be computed.

To summarize, both the synthetic and regression-sample data methodologies promise good, though uneven, results. If the synthetic estimate has a reasonable competitor, it is likely that a more optima result could be obtained by using both synthetic and competitive estimates in a regression format using the sample estimates as the dependent variable. The most important point, though, is that we need a systematic way of evaluating and comparing errors. One way to do this is to make explicit use of the sample data on which the synthetic and regression-sample data estimates were computed. Given the difficulty of evaluating estimates for areas where there is no sample information, the most useful applications of the regression-sample data method are likely to occur in estimating situations where sample data are available for all local units.

Finally, let us hope that future research on synthetic estimation does not follow that of ratio-correlation estimates. This latter method is a technique for estimating population change which has been used extensively on the State and national level. There is a literature full of variations on the basic method which in a particular estimating situation gave an improvement. People have tried stratifying local units, using differences between ratios instead of ratios of ratios, dummy variables, and many other variations, and have shown that their particular variation worked for them. Unfortunately none of these papers ever provided a methodology for determining which variation or the basic method was optimal in a new situation, and statisticians and demographers have been left to make the same ad hoc judgments as before.

REFERENCES

Ericksen, Eugene P. (1974), "A Regression Method for Estimating Population Changes of Local Areas," Journal of the American Statistical Association, 69, 867-875.

Ericksen, Eugene P. (1975), "Outliers in Regression Analysis When Measurement Error is Large," Proceedings of the Social Statistics Section of the American Statistical Association, 412-417.

Gonzalez, Maria E. and Christine Hoza (1978), "Small-Area Estimation with Application to Unemployment and Housing Estimates," Journal of the American Statistical Association, 73, 7-15.

Hogg, R. V. (1974), "Adaptive Robust Procedures: A Partial Review and Some suggestions for Future Applications and Theory," Journal of the American Statistical Association, 69, 909-923.

General Discussion

* Some very challenging philosophical issues were raised at some of the sessions. It is important to continue to explore the questions concerning synthetic estimates: When is it and when isn't it safe? What are the conditions under which one could use the method? What are the criteria?

* One criterion for when a synthetic or any of the other types of estimates should be used would be a circumstance when one can evaluate the error and determine whether the estimates are sufficiently accurate. If we are not able to assess the error in any way, then this should be a strong indication that the estimate should not be used unless it is politically dictated that it has to be. As statisticians working either for or with the government, we don't always have the freedom to make the choice not to use a synthetic estimate. Sometimes we have to do things that statistically we don't necessarily agree with.

If we are going to talk about errors of estimates, the size of error is important and also the direction of error. Almost every error for a place with a high unemployment rate is negative. If the objective is to spot places with high unemployment rates, than a synthetic estimate is particularly bad for that and should not be used.

Competitors will arise if the agencies that have the responsibility to compute estimates don't give out estimates that seem plausible to groups that might object.

It is possible that you could use regression methods and get rid of some of the bad characteristics of the synthetic estimates. But regression does not get rid of these characteristics. All it does is dampen them. Places with high unemployment rates where the synthetic estimate is too low, if the data are used for regression estimates, come out too low once again.

One of the things that you learn about in sociological statistics is ecological correlation. You learn not to use the characteristics of aggregates to make inferences to individuals. It seems equally invalid to use characteristics of individuals to make inferences to aggregates.

That is where the synthetic type of estimate that uses regressions of aggregates could go wrong. It is likely to misorder the weights that would be applied to variables. For example, variables that would predict drug usage on an individual level, for example, age, would be the most important. Yet age distribution of the population would not really do very well compared to other factors in estimating whether drug use is very high. If you have the kind of local area sample data, like the number of drug treatment centers or the number of drug arrests or the FBI's best guess as to the rate of drug traffic, they would turn out to be much better predictors and that would be the data to use.

(Contributing to the general discussion during this period were: Eugene Ericksen and Joseph Steinberg.)

Part IV

Drug Abuse Applications: Some Regression Explorations with National Survey Data

Reuben Cohen

ABSTRACT

Personal interview surveys in recent years have provided national estimates of use of marihuana, heroin, and other substances. Over a number of national surveys, consistent relationships have been observed between drug abuse and demographic variables such as age, education, and sex. Where one lives has also been found to be significantly related to level of drug abuse. This is observed in survey data in relationships between experience with drugs and geographic region of residence and community size and type.

Regression and other multivariate analyses have been used to help understand the prevalence of drug abuse among various segments of the general population and have provided a means to explore relationships between drug use and a number of additional factors related to location of residence. Regression procedures have also been used in an exploratory way to provide drug abuse estimates for States.

NATIONAL SURVEY RESULTS

A number of sample surveys in recent years have provided national estimates of use of marihuana, heroin, and other substances. Data collection and analyses for five such surveys have been carried out by Response Analysis Corporation, starting in 1971 and 1972 for the National Commission on Marihuana and Drug Abuse, and continuing in later years in cooperation with the Social Research Group, George Washington University, under sponsorship of the National Institute on Drug Abuse.

This paper aims to provide a flavor of the findings and something of the methodology of these surveys and invites the reader to think about the ways that results could be made more useful by appropriate use of small area estimating techniques.

Typically, the surveys have been based on national probability samples in the range of 3000 to 4500 personal interviews. They have included special samples of youth age 12-17, and have oversampled young adults in the 18-25 age range. Something more about the methodology is described further on, but first a few findings from the 1977 survey are presented to suggest the range of content and types of data available for additional analysis (Abelson, Fishbume, and Cisin 1977).

All of the surveys included a variety of measures of use and frequency of use of a range of substances, including illicit drugs as well as nonmedical use of drugs legally obtainable only under a doctor's prescription. Table 1 shows the range of substances and figures on lifetime experience reported in the 1977 survey by youth, young adults, and older adults. As a quick summary, each group is more likely to have had experience with marihuana and/or hashish than with any of the other psychoactive drugs studied. Clearly also, marihuana use is strongly associated with age, and the highest prevalence rate is found among young adults age 18-25.

TABLE 1

NATIONAL SURVEY ESTIMATES FOR 1977
LIFETIME EXPERIENCE*

	YOUTH 12-17	YOUNG ADULTS 18-25	OLDER ADULTS 26+
MARIHUANA AND/OR HASHISH	28.2	60.1	15,4
INHALANTS	9.0	11.2	1.8
HALLUCINOGENS	4.6	19.8	2.6
COCAINE	4.0	19,1	2.6
HEROIN	1.1	3.6	.8
OTHER OPIATES	6.1	13.5	2.8
STIMULANTS (Rx)	5.2	21.2	4.7
SEDATIVES (Rx)	3.1	18.4	2.8
TRANQUILIZERS (Rx)	3.8	13.4	2.6

*PERCENT EVER USED

Lifetime experience (ever used) is considerably higher than current use (use in the month prior to interview). For youth and young adults, the figures on current use of marihuana and/or hashish are roughly half as large as those reported for lifetime experience. For other substances, reported levels of current use fall off much more sharply from the figures for lifetime experience.

The national surveys have also shown substantial differences in reported levels of drug use among population subgroups other than age, and these have been generally consistent across the five points in time. Table 2 shows lifetime experience with marihuana for sex, race, and educational level. Males are more likely than females to report experience with marihuana, and reported marihuana experience also increases with educational level. Differences by race are smaller and less consistent.

TABLE 2

LIFETIME EXPERIENCE WITH MARIHUANA AND/OR HASHISH*
1977 SURVEY

	YOUTH	YOUNG ADULTS	OLDER ADULTS
TOTAL	28	60	15
SEX			
MALE	33	66	21
FEMALE	23	55	10
EDUCATION			
NOT HIGH SCHOOL GRAD	- -	52	6
HIGH SCHOOL GRAD	- -	60	16
COLLEGE	- -	65	26
RACE			
WHITE	29	61	15
NONWHITE	26	54	20

*PERCENT EVER USED

Patterns of use by geographic region and community type (Table 3) are of more specific interest to the topic of this workshop. For each of the three age groups, highest levels of experience are reported in the Northwest and West. and lowest levels in the South. For each age group also, more lifetime experience with marihuana is reported by residents of metropolitan areas than by residents of nonmetropolitan areas, with at least a suggestion of more experience in large metropolitan areas than in small metropolitan areas.

Lifetime experience with marihuana has increased significantly over the period covered by the five national surveys, as shown by figures for age groups in Table 4. With some allowance for sampling variability from one time period to the next, the figures also show a reasonably consistent pattern for sex and education (Table 5) and for geographic region and community type (Table 6).

TABLE 3

LIFETIME EXPERIENCE WITH MARIHUANA AND/OR HASHISH* 1977 SURVEY

	YOUTH	YOUNG ADULTS	OLDER ADULTS
GEOGRAPHIC REGION			
NORTHEAST	35	66	20
NORTH CENTRAL	29	61	14
SOUTH	19	50	9
WEST	36	67	23
COMMUNITY TYPE			
LARGE METROPOLITAN	37	63	20
SMALL METROPOLITAN	28	64	16
NONMETROPOLITAN	18	48	9

*PERCENT EVER USED

TABLE 4

LIFETIME EXPERIENCE WITH MARIHUANA AND/OR HASHISH*

	1971	1972	1974	1976	1977
12 - 13	6	4	6	6	8
14 - 15	10	10	22	21	29
16 - 17	27	29	39	40	47
18-25	39	48	53	53	60
26-34	19	20	30	36	44
35+	7	3	4	6	7

*PERCENT EVER USED

TABLE 5

LIFETIME EXPERIENCE WITH MARIHUANA AND/OR HASHISH*
ALL ADULTS

	1971	1972	1974	1976	1977
SEX					
MALE	21	22	24	29	30
FEMALE	10	10	14	15	19
EDUCATION					
NOT HIGH SCHOOL GRADUATE	8	5	9	12	12
HIGH SCHOOL GRAD	14	13	20	22	26
COLLEGE	23	32	28	30	35

*PERCENT EVER USED

TABLE 6

LIFETIME EXPERIENCE WITH MARIHUANA AND/OR HASHISH*
ALL ADULTS

	1971	1972	1974	1976	1977
REGION					
NORTHEAST	20	14	22	24	29
NORTH CENTRAL	19	15	17	19	24
SOUTH	5	8	13	17	17
WEST	21	33	29	29	32
POPULATION DENSITY					
LARGE METRO	20	21	24	26	30
OTHER METRO	18	20	20	24	26
NONMETRO	7	6	12	13	16

*PERCENT EVER USED

SURVEY METHODS

So much for the summary of national survey results. The starting point is a multi-stage area probability sample of the cotenninous United States, stratified by Census geographic divisions, metropolitan/nonmetropolitan place of residence, and other demographic factors. Primary sampling units were counties and groups of counties, with 103 such units selected for the Response Analysis national sample. Interviews for the series of studies described have typically been carried out in approximately 400 segments within the 103 PSU's.

Reasonably careful probability sampling and field interviewing procedures have been used at each step in the data collection process. Rough field counts are used to divide census enumeration districts and block groups into small segments, and field listings of specific housing units are completed in advance of interviewing. Letters are then written to households selected as part of the survey sample to announce the interviewer's visit and to urge cooperation with an inportant national survey.

In most cases, interviewers were trained on procedures for these surveys in regional meetings scheduled just before the start of field interviewing for each study.

The interviewer's first task at the sample household is to list residents of the household. Although the details of the procedure have varied somewhat over the period covered by the five surveys, the listings of residents have been divided into age groups for youth, young adults, and older adults, in order to provide for oversampling of the two younger groups.

In effect, two independent sampling procedures have been carried out at each household -- one for the youth sample, one for the adult sample. In households which include one or more elegible youth age 12 to 17, one such person is always randomly selected for the youth sample regardless of whether an adult is selected from that household.

The adult sampling procedure is somewhat more complex and depends on whether the household includes only young adults, only older adults, or both. No more than one adult is selected, and younger adults are favored by the probability selection procedures. Weights are used in processing survey results to compensate for the disproportionate nature of the sampling procedure.

Interviewers make repeated visits to sample households, as necessary, in an effort to complete interviews with each designated respondent -- sometimes up to ten visits or more. Additional efforts are made to solicit the cooperation of persons who initially refuse or who are

reluctant to participate. Interview completion experience for the series of surveys has generally been in the range of 80 percent of designated respondents; in the most recent survey, interviews were completed with 82 percent of the youth sample and 81 percent of the adult sample.

As one might expect for a survey on a sensitive issue such as use of illicit drugs, special efforts are made to protect the privacy of the respondent and to insure the confidentiality of data. A combination of procedures is used in the interview. Part of the questionnaire is a standard interview instrument with answers recorded by the interviewer, and techniques to afford greater privacy for the respondent are used in other phases of the interview. In those sections of the interview on illicit drug use, the respondent marks his or her own answers to questions read aloud by the interviewer. This procedure permits respondents to conceal potentially sensitive answers, while allowing the interviewer to maintain control of the interview. The answer sheets were designed so that, whether or not the respondent had ever used illicit drugs, the same amount of time would be required to fill out the forms.

Codes were used to identify completed questionnaires and answer sheets but neither names nor addresses were used. As each answer sheet was completed, the respondent was instructed to place it directly in a return envelope. At the conclusion of the interview, the main questionnaire was also placed in the envelope, and then, in the presence of the respondent, the envelope was sealed. The respondent, who had been told of these procedures in advance, was invited to accompany the interviewer to a mailbox. The interview materials did not contain the respondent's name or address anywhere on the questionnaires or envelope and were mailed directly to the central office. Interviewers were not permitted to review or to edit questionnaires.

REGRESSION ESTIMATES FOR STATES

Now that we have these kinds of data, how can we use them to assist in the development of estimates for States or smaller areas?

First we might consider the possibility of extracting estimates by looking into the survey data for interviews conducted within specific states. But sample surveys of adequate size to provide reasonably stable estimates for the total U.S. population are rarely large enough to provide direct estimates for specific States. A survey intended to provide estimates for the State of New Jersey, for example, would require about as large a sample for that State as for the U.S. as a whole in order to yield estimates of similar accuracy. Within the national sample, the number of locations and the number of persons in the sample in any given state are too small to provide a useful estimate. Indeed, the national sample used for the series of surveys described in this paper does not include interviews in every State.

Synthetic estimates of a type which require dividing the total population into a large number of specific cells based on a set of factors believed to be associated with drug abuse were not seriously considered because of the relatively small size of our national samples. Much larger samples would be needed than those on which this series of studies is based.

The specific procedure chosen for the work that is discussed next is a dummy variable multiple regression analysis. One portion of the analysis was carried parallel form, using a multiple classification analysis, with almost identical results.

In each case, a number of independent variables, or predictors, are identified. Each of these techniques deals adequately with the general problem of intercorrelated predictors provided that certain other assumptions are met.

One assumption of the classic multiple regression approach is that the variables used in the analysis are continuous and normally distributed. However, the technique has been adapted to deal with classifications (e.g., geographic regions) by using dummy variables in the regression equation. The multiple classification analysis (MCA) technique was developed specifically for classification data and is generally equivalent to the dummy variable multiple regression used for the complete series of analyses (Andrews, Morgan, and Sonquist 1969).

An important assumption of both the regression and MCA techniques is that relationships between the predictors and the dependent variable are additive -- that is, that the effect of each class of each predictor is not dependent on the values of any of the other predictors. In the case of the present analysis, multiple regression and MCA models would assume that a person's likelihood of having experience with a substance is composed of a series of additive coefficients, corresponding to the particular category or class in which he or she stands on each predictor. Thus, for example, separate effects could be calculated for age, sex, education, region of the country, and so on, and summed to obtain an estimated probability which takes all of those factors into account.

While the assumption of additivity is often taken to be a good initial approximation to reality, it poses some obvious difficulties in the analysis of drug abuse. An alternative assumption which must be considered is that the predictors interact -- i.e., two or more predictors have an effect in combination which is different from-the sum of their effects computed separately. Some parts of this general problem of interaction have been dealt with in the way that variables have been combined for the analysis. Additional work on the general problem of interaction would be a useful aspect of any further effort to develop a drug abuse index from survey research data.

Two sets of analyses have been done using these procedures. The first used data from the 1972 national survey; the second combined data from the 1974 and 1976 surveys.

In the 1972 survey analysis we created dependent variables for each of eight substances, for both lifetime experience and current use. Each of these was coded as yes/no. Before going on to discuss the predictor variables, Table 7 shows the proportion of variance we were able to explain in the analyses. The figures in the chart are the multiple R^2 for each analysis. At least a small proportion of variance in use is explained for each of the substances. The squares of the multiple correlation coefficients are highest for marihuana, and are higher for lifetime experience than for current use. This suggests, of course, that in likelihood of use, marihuana is more predictable than other substances -- and lifetime experience more predictable than current use. The sizes of the coefficients are probably at least in part a function of the overall levels of reported use. For drugs with very low levels of reported use, errors of various types, including reporting errors, are larger relative to reported frequency of use and thus are likely to reduce the amount of variance that might otherwise be attributed to the predictor variables in the equation.

TABLE 7

MULTIPLE R^2
1972 SURVEY ANALYSIS

	LIFETIME EXPERIENCE	CURRENT USE
MARIHUANA	.27	.18
HEROIN	.05	.03
COCAINE	.06	.05
HALLUCINOGENS	.13	.08
INHALANTS	.05	.02
SEDATIVES	.05	.05
TRANQUILIZERS	.02	.02
STIMULANTS	.06	.06

A number of different versions of the regression analyses were carried out with the 1972 survey data, using different numbers of predictor variables. The figures shown in Table 7 were based on an analysis using seven sets of dummy variables. With some differences in the group of dummy variables, the analysis was repeated for selected drugs with the combined 1974-76 survey data. Tables 8A through 8D compare results of analyses of the two sets of survey data for lifetime experience with marihuana. The youth and adult samples were combined in these analyses. In Table 8A we note that the multiple correlation coefficients were identical in the two analyses. Table 8A also shows "index numbers" for a combined age/education set of dummy variables, and for sex. The index numbers created for ease of interpretation are simply multiple regression coefficients multiplied by 100 and rescaled with the lowest valued coefficient set equal to zero.

TABLE 8A

MULTIPLE REGRESSION INDEX, 1972 AND 1974-6
LIFETIME EXPERIENCE WITH MARIHUANA

	1972	1974-6
Multiple R^2	.27	.27
AGE/EDUCATION		
12 - 13	4	3
14 - 15	12	18
16 - 17	28	37
18 - 20/COLLEGE	50	52
18 - 20/NONCOLLEGE	37	52
21 - 24/COLLEGE	48	52
21 - 24/NONCOLLEGE	35	45
25 - 34/COLLEGE	29	37
25 - 34/NONCOLLEGE	14	26
35 - 49/COLLEGE	4	10
35 - 49/NONCOLLEGE	4	5
50 AND OVER	0	0
SEX		
MALE	8	10
FEMALE	0	0

The same kinds of index numbers are shown in Table 8B for family income groups, used only in the 1972 survey analysis, and for race/ethnic group dummy variables. A question on family income has been included in interviews with adults but not in youth interviews. In order to include income in the 1972 survey analysis we used that part of the youth sample for which an adult had been interviewed in the same household, and assigned the income reported by the adult to the youth interview also. In the 1974-76 analysis we used the full youth sample and did not use the income variable.

It is possible that inclusion of family income in the analysis for 1972 but not for 1974-76 also has affected the results for race/ethnic group for the two years, but we have not tried to unravel these effects.

TABLE 8B

MULTIPLE REGRESSION INDEX, 1972 AND 1974-6
LIFETIME EXPERIENCE WITH MARIHUANA

	1972	1974-6
FAMILY INCOME		*
UNDER $5,000	9	
$5,000 - $9,999	4	
$10,000 - $14,999	0	
$15,000 AND OVER	4	
RACE/ETHNIC GROUP		
WHITE	4	6
BLACK	0	9
HISPANIC	0	0

* Family income not included in 1974-76 analysis

Table 8C shows results for the two principal sets of geographic variables we have used in the analyses. These show generally consistent results in terms of direction of differences between geographic groupings, but the differences are generally smaller in the 1974-76 analysis than in the 1972 analysis. There is a clear relationship between community type and reported lifetime experience with marihuana, and similarly between geographic region and marihuana use.

TABLE 8C

MULTIPLE REGRESSION INDEX, 1972 AND 1974-76
LIFETIME EXPERIENCE WITH MARIHUANA

	1972	1974-6
COMMUNITY TYPE		
LARGE METRO/CENTRAL CITY	19	12
LARGE METRO/SUBURBAN	14	7
SMALL METRO/CENTRAL CITY	19	6
SMALL METRO/SUBURBAN	6	2
NONMETRO/URBAN	4	2
NONMETRO/RURAL	0	0
REGION		
NORTHEAST	6	4
NORTH CENTRAL	2	2
SOUTH	0	0
WEST	14	9

Finally, in this series of findings, Table 8D shows results of one of our side excursions. In the analysis of the 1972 survey data, we coded a number of additional geographic variables based on county of residence of survey respondents. For example, each county in the national sample was coded as high, middle, or low in terms of percent of population living in college dorms, and similarly in terms of percent of population enrolled in college. For the 1972 analysis, percent in college dorms was selected for inclusion based on an early informal inspection of regression and correlation data for a large number of variables. In the 1974-76 analysis, both sets of dummy variables were originally incorporated in the analysis and stepwise regression procedures were permitted to select one set. The suggestion in both cases is that some proportion of experience with marihuana is explained by the presence of large numbers of college students in the community relative to total population.

TABLE 8D

MULTIPLE REGRESSION INDEX, 1972 AND 1974-6
LIFETIME EXPERIENCE WITH MARIHUANA

	1972	1974-6
% POPULATION IN COLLEGE DORMITORIES		
LOW	0	
MIDDLE	0	
HIGH	13	
% POPULATION ENROLLED IN COLLEGE		
LOW		0
MIDDLE		2
HIGH		7

If for no more than their curiosity value, the complete list of additional variables coded for the 1972 survey analysis is shown in Table 9. They have not been very useful so far, but they may suggest additional possibilities to the reader.

TABLE 9

POPULATION CHARACTERISTICS USED IN REGRESSION ANALYSES OF 1972 SURVEY DATA

CODED HIGH, MIDDLE OR LOW FOR COUNTY OF RESIDENCE OF SURVEY RESPONDENTS

POPULATION PER SQUARE MILE

PERCENT POPULATION CHANGE, 1960-1970

MEDIAN NUMBER OF PERSONS/HOUSEHOLD

PERCENT POPULATION IN ONE-PERSON HOUSEHOLDS

PERCENT FOREIGN BORN

PERCENT FOREIGN BORN AND NATIVE BORN OF MIXED OR FOREIGN PARENTAGE

PERCENT POPULATION IN GROUP QUARTERS

PERCENT POPULATION IN MILITARY BARRACKS

PERCENT POPULATION IN COLLEGE DORMITORIES

PERCENT OF CIVILIAN LABOR FORCE THAT IS UNEMPLOYED

PERCENT OF HOUSEHOLDS WITH INCOME LESS THAN POVERTY LEVEL

PERCENT BLACK POPULATION

LOCATION NEAR INTERSTATE HIGHWAY

LOCATION NEAR MAJOR POPULATION CENTER

To illustrate the possible application of regression estimates for specific States, indexes were computed from the 1974-76 analysis. Table 10 shows figures for the three highest and three lowest estimates.

TABLE 10

MARIHUANA INDEX
LIFETIME EXPERIENCE
1974-76 SURVEYS

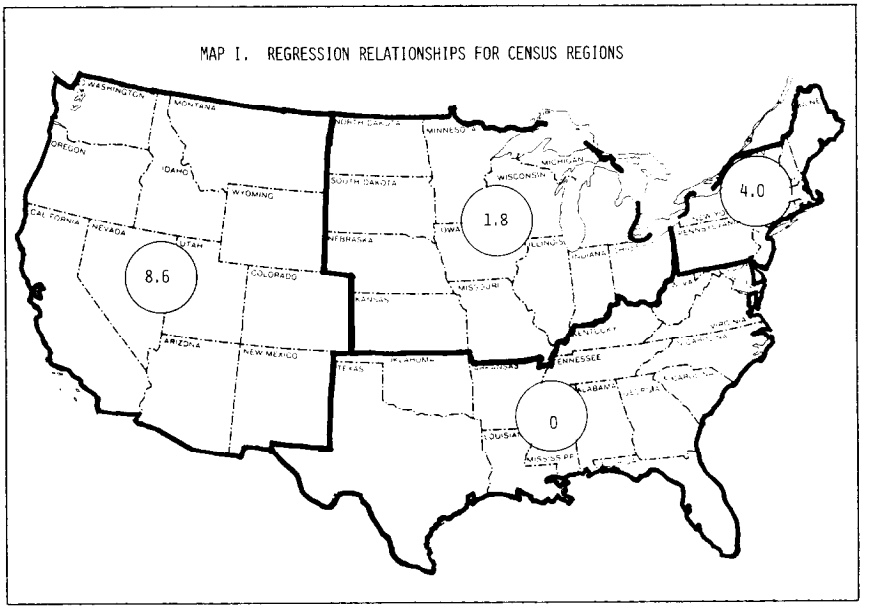
	INDEX*
HIGHEST	
DISTRICT OF COLUMBIA	155
CALIFORNIA	142
COLORADO	137
LOWEST	
ALABAMA	57
KENTUCKY	53
MISSISSIPPI	51

*AVERAGE FOR ALL STATES = 100.

EXAMINATION OF REGRESSION RESIDUALS

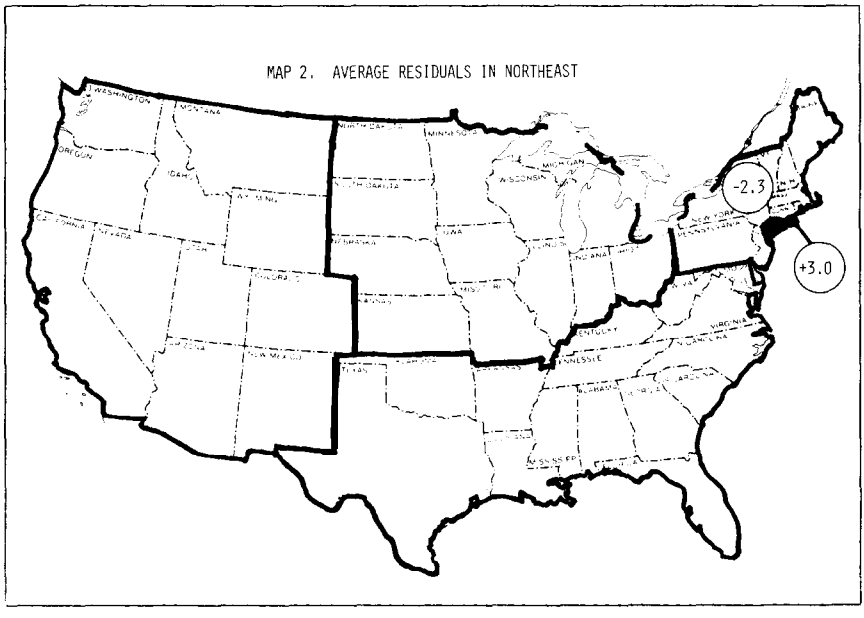
The final step in the exploratory work that is included in this paper was an examination of regression residuals from the 1974-76 analysis. The research started with a hypothesis, but most statistical cautions were thrown aside in looking at residuals for areas in the national sample figuratively plotted on a map of the United States. The implication of the regression coefficients shown earlier is that the United States consists of four large plateaus, at four different heights with respect to reported experience with marihuana, represented by regression coefficients for the four census regions shown earlier.

The plateaus would be at the relative heights shown in Map #1. There would, of course, be sharp elevations wherever metropolitan concentrations occurred, with peaks represented by central cities.

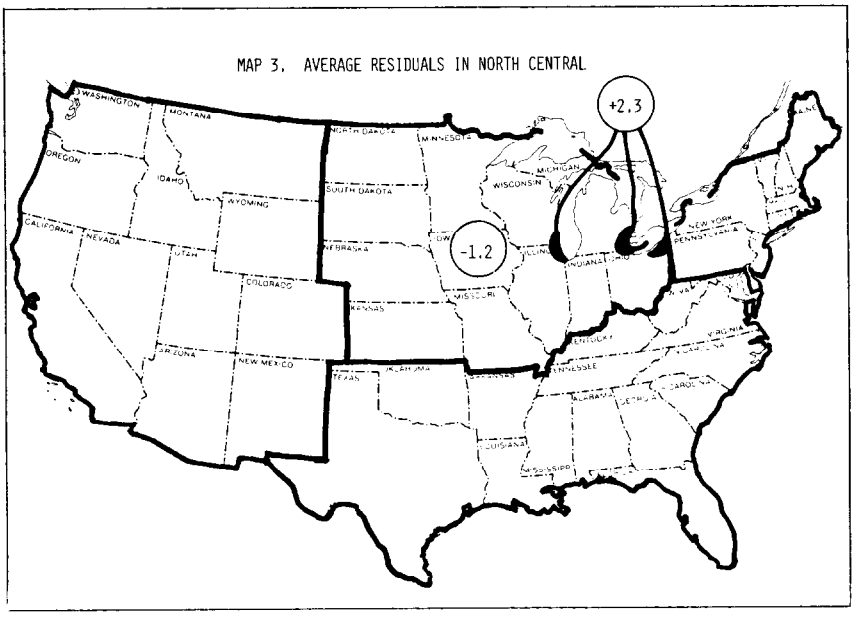


My own mental map of the United States suggests something quite different -- perhaps rolling hills and valleys corresponding to points of entry and avenues of diffusion of drug experience. With this in mind, I looked at residuals which are in effect deviations from the plateaus, after taking into account metropolitan/nonmetropolitan community type and variations in demographic features such as age, sex, and education.

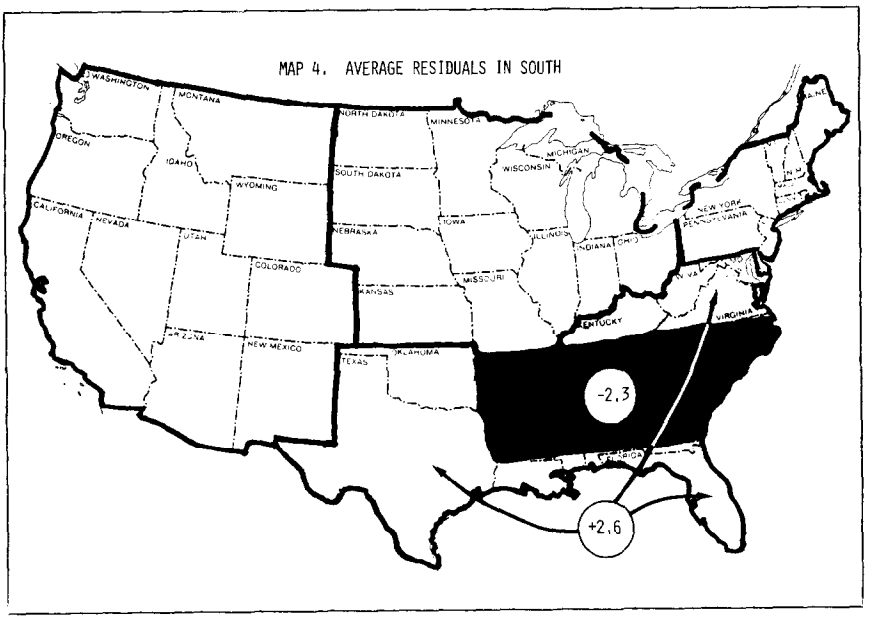
The number of PSU's in our national sample poses obvious limitations for this type of examination of residuals, but let me share with you the terrain features that emerged for me. Starting with the Northeast region (Map #2) there seems to be a difference between an area included within a broad arc drawn around New York City and the rest of the region. The arc extends into Connecticut and into Northern New Jersey. Residuals for sample locations within the arc average plus 3 percentage points.[1] In other words, even after taking community type and demographic features into account, New York City and the surrounding area average about three percentage points higher than the region as a whole, or about 5 percentage points higher than the rest of the region.



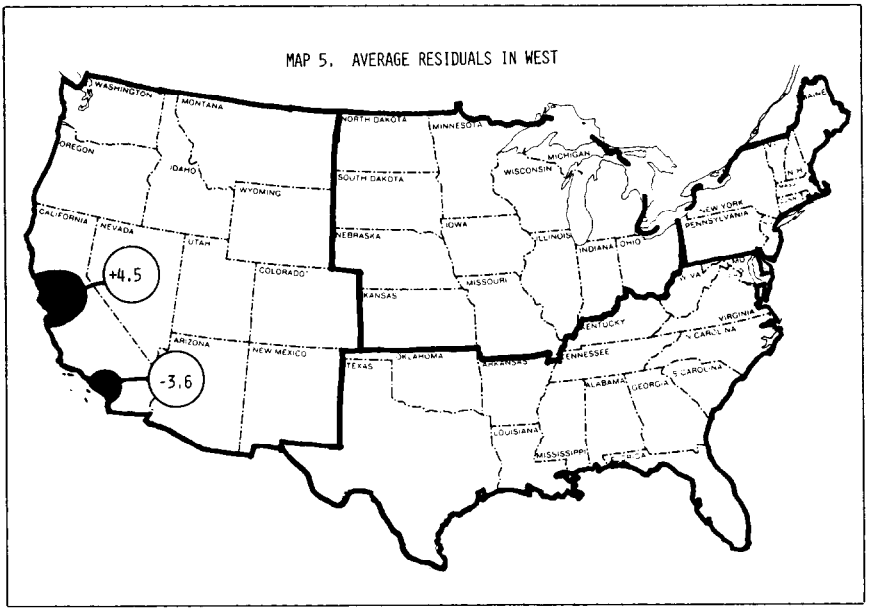
For the North Central region (Map #3), the specific features don't exactly pop off the map but there does seem to be something different about the metropolitan regions near the Great Lakes and the rest of the region. The Great Lakes metropolitan group embraces the areas of Chicago-Milwaukee, Detroit-Ann Arbor, and Cleveland-Akron-Youngstown. This grouping averages plus 2 percentage points, and the rest of the region minus 1.



For the South (Map #4), the picture is different. There is a depression in the terrain that runs across the States of the deep South. The band extends from Georgia and the Carolinas across to Arkansas and Louisiana. Residuals in these states average minus 2 percent compared to about plus 3 percentage points in the rest of the region.



The West is more complex (Map #5). The most noticeable features are highs in Northern California and lows in the Los Angeles area. The Northern California grouping of locations extends from the Bay Area to Sacramento; residuals average about plus 5 percentage points. For the Los Angeles area, including Orange County, residuals average minus 4 percentage points. Locations in the rest of the region average about the same as the entire region.



Examination of the residuals has been an interesting exercise. I suspect that careful study will suggest new approaches to meaningful estimating procedures for small areas.

FOOTNOTE

1. For this analysis, residuals were averaged for four or more primary sampling units. For any grouping examined and reported separately, the minimum number of interviews is 280.

REFERENCES

Abelson, H.I.; Fishburne, P.M.; and Cisin, I. National Survey on Drug Abuse: 1977. Princeton, N. J.: Response Analysis Corp., 1977, and (Volume 1, Main Findings) National Institute on Drug Abuse. DHEW Pub. No. (ADM)78-618. Washington, D. C.: Superintendent of Documents, U. S. Govt. Printing Office, 1977.

Andrews, A.; Morgan, J.; and Sonquist, J. Multiple Classification Analysis. Ann Arbor, Michigan: Survey Research Center, University of Michigan, 1969.

Discussion

Monroe G. Sirken

Reuben Cohen proposes and illustrates a multiple regression model for producing State and local area synthetic estimates of drug use. He suggests that the designs of the national surveys conducted by NIDA favor the regression estimator over a synthetic estimator because the sample size of NIDA's national survey is too small to be divided into a large number of population subdomains. In other words, the sampling errors would be larger based on the synthetic estimator. However, Reuben does not present any empirical or theoretical evidence to substantiate this view. Personally, I doubt that much is gained by dividing the population into a large number of subdomains. For instance, synthetic estimates of health service utilization are changed very little by increasing the number of subdomains beyond those by age and sex.

In his continued work with the drug use data, I suggest that Reuben undertake two types of studies - one theoretical, the other empirical. First, it would be very helpful if he would indicate the relationship between the multiple regression estimator and the synthetic estimator. Second, that he use the NIDA data to compare the State estimates of drug use and their sampling errors for the two estimators.

One of Reuben's observations deserves underscoring. He notes that although drug use varies greatly by demographic variables, like age and sex, these variables account for only a small fraction of the total variance in the populations's use of drugs. He shows this to be particularly true for the rarer drugs. Does this imply that we should be wary of synthetic estimates of drug use, particularly for rarer drugs?

Discussion

Ira Cisin

The scope of this workshop is considerably broader than I had expected; we are scheduled to discuss a wide variety of estimating procedures, both direct and indirect, and perhaps to discuss a hierarchy of utility within the indirect domain. As far as I can tell, our vocabulary in this field is not sufficiently differentiated, so that when a term like "synthetic estimates" is used, we are not all necessarily thinking about the same thing. Even the term "synthetic" is a little unfortunate, since the connotation it evokes suggests "imitation" or "ersatz" --not quite the genuine article. My intent is to demonstrate that synthetic estimates are indeed genuine and potentially important; to make explicit some obvious conditions and assumptions under which synthetic estimates can be most useful; and to make a couple of modest proposals on how their utility can be increased.

Our procedures are "synthetic" in that they synthesize information from more than one data set. In the case of the drug use estimates, we have the results of a national sample survey; we search these results for an explanatory model--that is, we seek a set of "predictor" variables or "independent" variables which will maximally account for the variance in some particular "criterion" or "dependent" variable. Fundamentally, this is a regression procedure, whether the results are expressed in terms of regression coefficients or whether they are expressed in differential probabilities for defined subgroups. Then, armed with our survey results, we apply our model to a geographic segment of the population which is a part of the total population but which was not sampled intensively. Usually the geographic segment is a State or a city or a county. The result is a synthesis of the national sample survey data with available Census information about the geographic segment or segments of particular interest.

Three observations on the procedure are appropriate at this point:

1. Obviously, the procedure is not very useful if the explanatory power of the regression model derived from the national survey results is weak. If the best we can do is a very small R^2 in explaining the variance in the criterion, and/or if that model is based on variables whose distribution does not differ much from State to State or county to county, then the exercise will

inevitably lead to estimates which differ only minutely from estimates based simply on population size. On the other hand, if a powerful regression model can be generated and if it uses components which differ considerably from smsall unit to small unit, then the outcome will be quite different. Several speakers have mentioned that synthetic estimates for small areas usually do not differ much from area to area. The reason is obvious: practitioners have concentrated on predictors which maximally differentiated in terms of the criterion behavior; only as an afterthought have they remembered that such variables as sex and age do not differ very much from one small area to another. So the net outcome is disappointingly nondiscriminating.

2. In applying this procedure, we assume that the influential factors which apply to large aggregates apply equally well to small aggregates; that is, we assume that there are no significant interactive effects which are unique to the individual States or other entities for which estimates are generated.

3. We must also keep reminding ourselves that the search procedures we use in generating our explanatory model are themselves **maximiz**ing procedures. Regression statistics as applied in search procedures are descriptive statistics, fine tuned so as to take every advantage of the idiosyncracies of the particular sample in which they are calculated. In psychometrics, we know that cross-validation of a regression equation is expected to yield a lower R^2 on a new sample than it did on the sample from which it was derived. In exactly the same way, we are undoubtedly overestimating our explanatory power.

Recognizing these limitations, I want to comment briefly on the importance of these procedures in various research applications, in addition to their applications to geographic estimates.

Exactly the same synthesizing procedures are widely used in generating regression estimates of missing data because of item nonresponse in surveys. Given that we have some information on the nonrespondents, we search the respondent data set for an explanatory model, seeking the correlates of the responses to an item that is missing among the nonrespondents, again seeking those correlates which differentiate among the responses within the respondent group and at the same time differentiate the respondent group from the nonrespondent group.

Similarly, but less widely recognized, we have applied this technique to the standardization of samples in natural quasi-experiments. Morris Rosenberg first suggested this tactic in his work on test-factor standardization. The paradigm is simple. Let's say we are studying the relationship between TV viewing and aggressive behavior; we do not have a controlled experiment; we have a survey, and we can compare the aggressive behavior of heavy viewers with that of light viewers; but heavy viewers and light viewers are self-selected and the two groups differ markedly in various other ways. Obviously we should standardize the two viewing-level groups with respect to their other differing characteristics, using the Rosenberg procedures, but Rosenberg (1968) does not suggest systematic ways for choosing the variables

on which to standardize. William Belson (1959), an English psychologist, gets credit for the first attempt, however crude, to select systematically the standardization variables which would in this instance differentiate between the viewing groups and, at the same time, differentiate on the criterion behavior--in other words, to select standardization variables which would do the most work.

Two constructive suggestions arise from consideration of these applications:

First, it seems obvious that the search procedures could be improved by use of an interactive tactic like AID rather than linear multiple regression. Certainly interactions can be built into linear multiple regression, but this has to be done artistically, as Reuben Cohen did it. The AID disadvantage of dichotomization of predictors is easily overcome and the interactions among the predictors can be detected objectively.

Second, and most important, we should continue to explore techniques for systematic selection of predictor variables which provide maximum power; that is, predictor variables which contribute to explanatory powver and at the same time differentiate among the small geographic units. The trick, of course, is to select standardization variables with optimum relationship to the two criteria. To start, we can follow Belson's lead: he developed a search technique which would make a stepwise selection among the candidate predictors this way: he invented a summary statistic to express the candidate variable's relationship with one of the criteria and separately its relationship with the second criterion. Then the basis for selection would be the product of the two summary statistics. Subsequent selections are accomplished stepwise in a manner that has become familiar in the AID adaptation. Although Belson's invented statistic is statistically questionable, we at the Social Research Croup have been working with both correlation coefficients and analogs of chi-square to achieve the same objective in a statistically defensible manner.

The symbolic representation is simple:

Let variable 1 be a drug use criterion; and variable 2 be State of residence; then we are seeking a set of variables "3" that will maximize the absolute value of the product:

$R_{13}\,R_{23}$, not merely maximize the absolute value of R_{13}. The correlation product is recognizable as the right-hand term of the numerator of the familiar formula for the partial correlation coefficient.

There are minor technical difficulties in our dual criteria technique. Since residence in the 51 States is a nominal variable, and we are using it as a criterion, we have some trouble with nominal variables as candidate predictors. Ideally, we could use correlation coefficients for some of our calculations and non-parametric chi-square analogs for others. But we have qualms about equivalence.

In any case, we now have a solution for the dual criteria problem in simple cases like item nonresponse estimation; and we are confident

that the approach can be generalized to more difficult practical problems.

REFERENCES

Belson, W.A. Matching and prediction on the principle of biological classification. Applied Statistics, 8:65-75, 1959.

Rosenberg, M. The Logic of Survey Analysis. Nav York: Basic Books, 1968.

General Discussion

* It is useful to note that there is a relationship between the regression-based estimates using dummy variables and the covering (nearly unbiased) estimates that Paul Levy discussed. when you use a regression procedure instead of using a cell mean in the covering estimate equation, you are using a predicted value of a cell mean from a linear combination of data. One advantage is that you can account for more variables because you are building up your degrees of freedom; you might be able to include six or eight variables (or however many you might want to use). Whereas, if you are using the covering estimator, then six or eight variables would involve a multiway crossclassification with 400 cells and would become awkward to use. Another advantage of the regression procedure is that by taking into account more variables you could probably get ones that are better (given that you have measured them and have them available). The difficulty is that unless you carry out an assessment of the regression relationship you run the risk of leaving out variables. If you leave out variables, that causes estimates to have properties that may be misleading.

If you did a statistical test that demonstrated that the interactions were unimportant, then the estimates based on the regression would be essentially the same as the estimates based upon the ordinary means, and they would probably have smaller standard errors. The dilemma is that the bigger you make the table, the poorer your ability to do the test. And then you have to start assuming that the model you are producing is useful on certain kinds of a priori considerations.

When you use these estimates you are adopting something called a "response error model" point of view. You are in essence saying: response errors dominate and sampling errors are less important. If it turns out that the assumption that there is no sampling error is an appropriate one, the regression estimates may be very satisfying. If it turns out that each particular unit in a population has unique characteristics, so that the sampling error is indeed important, then the prediction model may not work out very well. The dilemma is that most situations are a mixture of the two and we don't necessarily know how to deal with the mixture.

* One problem that exists is the multiple use of the same word: regression. When you take the regression approach in the sense of trying to find alternative indicators for geographic units you're interested in, one of the properties of that approach is that it allows you to make use of any information. One could use information which has nothing to do with the variables in the survey. For example, a practical suggestion would be to consider a regression estimate using the number of drug treatment centers in an area as a predictor variable. Of course, sometimes after trying a predictor variable, it becomes necessary to throw it out as having a poor predictor ability.

* Another way of considering the problem is to use available data for changing the strategy of the structure of the basic survey design. In this approach one would aim towards the use of the data not only for a national survey result but also for the basic needs of synthetic estimate purposes.

* The dilemma for the user is that while the technique discussed can be implemented, there seem to be problems of lack of variation among areas, between proportions of useful demographic variables, and a lack of explanatory power of predictor variables.

* (Joan Rittenhouse) I'd like to follow up on that point because I'm deeply involved in a data set, that is, the National Survey, which gives us very respectable estimates for drugs of wide prevalence, particularly marihuana. But our office gets calls constantly from States and localities, and they really need, not only for treatment purposes, but also for public health purposes, good estimates for heroin. In the unidentified (i.e., nonclinical) population we have little, very little to help them. So when we got into the Levy discussion I began to feel like that bumper sticker which said "I found it!' because it seemed like the answer to States and localities: synthetic' estimates. We can give them this technology and they can put it to work to come up with the estimates they need.

But a little later on in the Levy presentation, when he talked about the power to discriminate one area from another given equal distribution or powerful predictors such as age, I began to get a feeling more like the other bumper sticker which says "I lost it." All these small areas have people in these age groups; so there it goes. You get a very nondiscriminating estimate. Reuben was suggesting a number of other non-age variables which contribute to the prediction of drug abuse less significantly than age, but which contribute something. They also discriminate one area from another: for example, race, and density of the population. Since these factors have been associated in the past with different rates of drug abuse, they would seem ideal for incorporation into the synthetic estimates procedure and for the generation of discriminating prevalence estimates by locality. So there may be a second chance to say "I found it."

But--the National Surveys have shown in the past two years or so that population density and race, to persist with these variables! are losing their meaning so far as drug abuse is concerned. The 1977 findings made the point even stronger; the differences are disappearing.

So now I really feel that "I've lost it."

* The situation may not be that bleak, although you've dramatized the issues quite a bit. It may be useful to focus on the variance components--the between and the within components. The heart of the issue is how things vary not in the population as a whole, but area by area. One could suppose it is possible to get a moderately low R^2 and find that most of it is accounted for by within area variance. It would be necessary to investigate the between and within variance aspects to know whether the synthetic procedure would be useful.

* You want to look at two things. One is the R^2 for the national data; the other is the variability between areas in the composition of the population. Perhaps some statistical work could be done. It may be useful to determine and to define the combination of the two criteria under which it might be fruitful to try to use a synthetic estimator and the conditions under which it might not be.

* Another question: Is there a cutting point for R^2 before we should become serious about using the regression estimator? It may be worth noting that sometimes the R^2 can be increased considerably by taking into account other variables (e.g. lifestyle variables in a drug use application). These may be considered soft types of variables, and some data collecting agencies may prefer not to collect them. However, these types of variables may be worth obtaining.

* It is necessary to consider whether there is a systematic way to get synthetic estimates which are as different as possible from simply applying the national data to the small areas. The answer may lie in selecting predictors-- independent variables--such that the product $R_{13}\ R_{23}$ is maximized. This implies that you can determine a small set of predictors which maximally explain the criterion variables and are maximally different among the States or among the small areas. Setting up the dual criteria answers that question. If either one of the two relationships is zero, it doesn't matter how big the other one is--it is not going-to make a difference. You might as well apply the national estimates. You can think of it as a continuum rather than a cutting point.

If you're going to predict a phenomenon temporally, you have to use demonstrably antecedent variables. However, if somebody else has done a survey in which questions concerning soft variables have been asked, there is no reason why the soft variables cannot be used in a synthetic estimate. The objective is not temporal prediction. The objective is estimation, and for estimation anything goes. They can be used for this purpose.

* The heart of the problem is not whether variables are soft or hard but what is the likelihood of being able to get reliable data at the local area level.

* One should consider using available data (e.g., the existence of treatment facilities for a specified disease) if the data are very reliable on a small area basis, and different from area to area, and demonstrably

correlated with criteria. There is nothing that restricts you to using only your own sample survey results.

* It would be useful if there existed an archive of national sample data that have been collected, giving the nature of the variables that have just been referred to, and if the information would be available so that you could assume certain relationships were preserved over time. But the point is worth recognizing that you are in a prediction mode. There may be something uncomfortable with the notion of maximizing variation between local areas, particularly at the State level, because a number of States are relatively homogeneous with respect to each other but very heterogeneous within. They are comprised of individual units which may be quite different county by county or for the metropolitan area versus the rural area. If you are not careful with respect to $R_{13}R_{23}$, you could get into some difficulty; you may start out thinking about States but really want counties; and you probably should be pretty sure as to exactly why you are choosing a particular criterion. It is an interesting concept. However, it has to be used fairly carefully relative to where you want to produce the estimate.

* To summarize, if an analysis shows the demographic variables do not explain much of the variance of the dependent variables, then there may not be any point in going ahead and using a synthetic estimate for local areas with these variables. Even if there is a reasonable degree of explanation, if there is little variability in the distribution of the demographic variables among areas, the synthetic estimate approach may not be very useful. Political subdivisions are not necessarily going to be the areas that one wants to use for synthetic estimates. It may be better to produce estimates for classes of local areas that are likely to show better results and then recombine the results into the areas of interest for use. In our discussion the question arose whether the multiple regression synthetic estimator is better than the demographic synthetic estimator. It might be interesting to set up a test where sample size is varied to get some idea of how variance and bias of the two types of synthetic estimates vary by sample size.

(Contributing to the general discussion during this period were: Ira Cisin, Reuben Cohen, Eugene Ericksen, Gary Koch, Fred Oeltjen, Louise Richards, Joan Rittenhouse, Monroe Sirken, Joseph Steinberg, and Joseph Waksberg.)

Applications of Synthetic Estimates to Alcoholism and Problem Drinking

David M. Promisel

ABSTRACT

This paper focuses on the application of synthetic estimation techniques to issues involving estimation of the prevalence of alcoholism and problem drinking. Demands for information led to the first use of synthetic estimation in this area. However, the experience of bringing that first application to fruition led to new uses where previously no attempt would have been made to develop information. Three examples are discussed briefly: estimating the relative prevalence among the States; identifying health manpower shortage areas; and calculating the need for service in a community.

BACKGROUND

The question "How many people are there with alcohol-related problems?" is a difficult one for two reasons: (1) defining what are alcohol-related problems; and (2) counting the number of people who have them.

Alcohol is associated with a multitude of problems, ranging from alcohol addiction and behavioral difficulties associated with intoxication to diseases such as liver cirrhosis and various cancers resulting from excessive alcohol consumption. The causal nature of the association has been established in some cases and is only suspected in others. Often, the individual's problem is the result of alcohol working in conjunction with other factors such as diet, genetic or familial conditions, psychological status, concomitant use of tobacco or other drugs, etc. And there is a reasonable degree of independence among all these factors, so that there is no small set of them that can be used as markers of the entire population with drinking problems.

The World Health Organization has summarized this situation (Edwards, et al. 1977) by &fining two concepts. The "alcohol dependence syndrome" is "a state, psychic and usually also physical, resulting from taking alcohol, characterized by behavioral and other responses that always include a compulsion to take alcohol

on a continuous or periodic basis in order to experience its psychic effects, and sometimes to avoid the discomfort of its absence; tolerance may or may not be present." In addition, an "alcohol related disability exists when there is an impairment in the physical, mental, or social functioning of such a nature that it may reasonably be inferred that alcohol is part of the causal nexus determining that disability,"

Historically, two approaches have dominated attempts to estimate the prevalence of alcohol problems: surveys and indirect estimation. A useful review of this topic is provided by Keller:

> In recent years numerous efforts have been made to identify by survey methods populations exhibiting drinking problems. For the most part these surveys have sought primarily to describe the drinkers and abstainers in general or particular populations, and secondarily to identify the kinds of motivations and problems associated with the drinking by some people, and the kinds of people who experience those problems.
>
> One important culmination of these efforts is the work of Cahalan and his associates. Improving on prior methods they have developed a description of drinking that takes account of quantity, frequency and variability, and from the drinking thus delimited they have developed a classification of infrequent, light, moderate and heavy drinkers. Based further on reported reasons for drinking, they have extracted a class of "escape" drinkers. These are persons who reported two or more of the following motives: (a) helps them relax, (b) is needed when tense, (c) cheers up, (d) helps forget worries, (e) helps forget everything.
>
> Keller 1975. Reprinted by permission from *Journal of Studies on Alcohol,* Vol. 36, pp. 1442-1451, 1975. Copyright by Journal of Studies on Alcohol, Inc., New Brunswick, NJ 08903

Building on these techniques, the National Institute on Alcohol Abuse and Alcoholism, shortly after its founding in 1971, initiated a series of national surveys. Over a five-year period, seven surveys were conducted by Louis Harris and Associates (Harris and Associates, Inc. 1974) and Opinion Research Corporation (Rappeport, Labow and Williams 1975). It has proven quite difficult to merge all of the Cahalan and later surveys for analysis purposes. However, for illustration, table 1 shows the results of an analysis of data on problem drinking from several of the NIAAA-sponsored surveys. These results suggest that of adults who drink, about 10 percent can be classified as problem drinkers, with women having a substantially lower rate than men. An example of the combined use of Cahalan's and these later surveys applied to synthetic estimation is provided later in this paper. A national survey commissioned by NIAAA is currently being designed which, among other things, will specifically establish the linkages among the alcohol problem indicators used in these various surveys.

Some of the difficulties in using survey methods for estimating prevalence were described briefly by Cahalan:

> However, survey methods have some inherent drawbacks, a few of which are worth noting here. They are relatively costly and time consuming. Area probability samples may miss people who are not in households-- and these may be people who are particularly relevant to alcohol studies. Thus the Armor report suggests that the clinic populations are more extreme in alcohol use than survey data indicate. Surveys depend upon the cooperation of respondents and thus in large part they collect respondents' estimates and recollections, which may of course be inaccurate: not only in the playing down of unflattering materials, but also the reconstruction of the past in terms of what "everyone knows" about alcohol use and alcohol problems. (Cahalan 1976, p. 17)

Jellinek's formula is the famous instance of application of indirect techniques to prevalence estimation. Jellinek hypothesized (Keller 1975) that there was a relatively constant relationship between alcoholism and mortality from cirrhosis which would permit an estimate to be made of the number of "alcoholics with complications." This led to the development of the formula $A = (PD/K)R$. In this formula, the number of reported deaths from cirrhosis in a given year, D, is multiplied by P, the presumed constant percentage of such deaths attributable to alcoholism (different for men and women), and divided by K, another constant, representing the percentage of alcoholics with complications who die of cirrhosis. The result is then multiplied by R, the ratio of all alcoholics to alcoholics with complications in the given place and time.

Over time, many including Jellinek expressed doubt about the reliability of this formula and the constancy of its parameters. One proposed solution was a modified version of the formula. Keller argued that there was no evidence that the basic rates associated with alcoholism in the U.S.A. had undergone any substantial change since the early 1940's. If then the average basic rate of the years 1940-1945, when the formula appeared to yield reliable results, were applied to the current population, an approximation of the prevalence of alcoholism could be derived. This has been the method used in the Efron, Keller, and Gurioli series, Statistics on Consumption of Alcohol and on Alcoholism, published by the Rutgers Center of Alcohol Studies.

Even with these modifications, however, numerous questions remain regarding the adequacy of the formulation, estimation of parameter values, and the nature of the alcoholic population represented by this estimation procedure. Nevertheless, indirect techniques are believed to have large potential utility for prevalence estimation and are currently under active investigation by NIAAA.

As difficult as it may be to estimate, prevalence is central to innumerable program and policy decisions. These decisions range from the need to compare the numbers of people suffering from various health problems to the requirement for predicting the extent to which alcoholism treatment benefits will be utilized under national health insurance. The next section describes three examples of synthetic estimation techniques applied to alcoholism prevalence questions: estimating the relative prevalence among the States; identifying health manpower shortage areas; and calculating the need for service in a community.

CASE STUDIES

1. Relative Prevalence of Alcohol Problems Among the States

In the legislation establishing NIAAA in 1971, a requirement was stated that revenue sharing funds be alloted to the States "on the basis of the relative population, financial need, and need for more effective prevention, treatment and rehabilitation of alcohol abuse and alcoholism." For several years, need for more effective prevention, treatment and rehabilitation was expressed by the relationship of the population of each State to the total population of all the States. However, in the report of the Committee on Labor and Public Welfare, U.S. Senate, in 1976, it is stated that the Committee was distressed to learn that this "need" provision in the law had been totally disregarded. As a result, the legislation that was passed that year to continue the existence of NIAAA required that within 180 days the Secretary of HEW, by regulation, establish a methodology to assess and determine the incidence and prevalence of alcohol abuse to be applied in determining this 'need."

The NIAAA, with the help of the National Center for Health Statistics, undertook to respond to this congressional mandate. It was clear that the response needed to be quick and that it should be equitable to the States in that they should not be penalized for their reporting practices. It was decided that the best way to ensure equitability was to use national data sources such as national population surveys and data collected by the U.S. Census Bureau. In the time available the only mechanism for developing prevalence estimates was the use of synthetic estimation in conjunction with the data that were then available. It was not possible to initiate collection of new data. It should be noted that there was no necessity to estimate the actual number of alcoholic people in each State but only the relative numbers from State to State.

The problem became one of defining an index of alcohol problems and then establishing on a national basis the relationship of various demographic variables to this index. There were no single measures felt to be sufficiently indicative of all alcohol problems. Furthermore, there did not exist a single survey considered to be definitive for the purposes of establishing the necessary relationships. Accordingly, two surveys were used, with a different index of problem drinking from each. These were selected strictly on a judgmental basis. The first survey was carried out by the Social Research Group (SRG),

University of California at Berkeley (Cahalan 1970) in 1967. The other was the Harris Alcohol Survey of December 1971.

The two indices of problem drinking are:

(1) Frequent Heavy Drinking (FHD) - the number of times per week that a respondent drinks 5+ drinks on one occasion (coded in 4 categories). Based on Harris survey.

(2) Current Tangible Consequences (CTC) - an additive score concerning problems with spouse, relative, friends, job, police, finances; and health (coded to 10 categories). Based on SRG survey.

The first, FHD, was considered representative of chronic alcohol problems in need of treatment. The second, CTC, was associated more with intoxication and incipient alcoholism where prevention programs would be appropriate.

The eight individual characteristics used to "predict" problem drinking are: age, sex, residence (urban/rural), race, region of the U.S., marital status, education and income. The choice of these characteristics was based on their known relationship to alcohol problems and their availability on a State basis from the U.S. census.

The statistical technique used to establish the relationships is called the Automatic Interaction Detector (AID) (Sonquist, Baker and Morgan 1973; Sonquist and Morgan 1964). This approach is somewhat analogous to "stepwise regression" where the independent variables need not be quantitative nor even categorized into equal intervals or into ordinal categories.

The results of the AID analyses are shown in figures 1 and 2 and an example of the use of this information is provided in table 2. It can be seen in figure 1 that the best single predictor with the FHD index is sex. The only other significant split for females was marital status. The FHD factors for males included age, marital status, region of the country, education, and income. For the CTC index (in addition to sex) race, age, marital status, and geographic region were also significant.

The final "need" index, or index of relative prevalence, proposed in response to the congressional mandate was as follows: the total FHD and CTC scores for the State were divided bv the national average scores to produce relative scores for the State; the mean of the resulting FHD and CTC scores was the relative measure of alcohol abuse and alcoholism in each State or the "need for more effective prevention, treatment, and rehabilitation." The index of relative prevalence is combined with population data and financial need in a formula which computes for each State its allotment from the Federal revenue sharing fund established for use with alcohol programs.

This formula was presented in a Notice of Proposed Rulemaking published in the Federal Register (Vol. 42, No. 21, pp. 6066-6069) in February of 1977. In that notice, comments on the formula were requested and 46 letters were received by NIAAA. Summaries of these letters and the NIAAA responses to them were published in the Federal Register (Vol. 42; No. 227, pp. 60398:60403) in November of 1977. The dominant theme of the responses was obiections that some States would get reduced funds as a result-of the formula. To resolve that issue legislation was passed specifying essentially that no State shall receive an allotment less than it would have received using the formula in its prior version.

Several comments pertained more specifically to the needs index derived from the synthetic estimates. Objections were made that the estimates were based on survey data gathered in 1967 and 1971 and were unreliable because of their age. There were complaints that the indices used were unreliable and proposals were made to replace them with others considered to be more suitable such as per capita consumption of alcohol? deaths from cirrhosis of the liver. or alcohol-related fatalities. Others pointed out that the indices used did not reflect specific geographic factors such as those that occur in rural areas or States with special problems, such as Florida; and some objected to the relative weight assigned to need compared to the other factors in the formula.

The general response by NIAAA to these concerns was to point out that NIAAA planned to undertake a new national survey to get current data; that the regulations did not require that the same indices be used each year so that better indices could be implemented after they became available; that there were restrictions on the use of indices resulting from the need to be both comprehensive regarding alcohol problems and thoughtful of the reporting capabilities-of each of the States; and that some valid issues could not be resolved with the knowledge available at the moment.

2. Identification of Health Manpower Shortage Areas

The Health Professions Educational Assistance Act of 1976 contains a number of provisions providing support for the education and training of individuals working in health services. Certain geographic areas with shortages of health services will be eligible to request National Health Service Corporation personnel. They will also constitute areas of service for those receiving aid from Public Health Service scholarships and loan repayment programs. This concept of manpower shortage areas will also be used in connection with other Public Health Service programs. In late 1976, the NIAAA was given the opportunity to recommend criteria for use in determining which geographic areas had a shortage of alcoholism treatment personnel. At that time manpower in the alcoholism context referred solely to psychiatrists.

Conceptually, identification of manpower shortage areas is a function of estimates of the prevalence of problems in given areas, specifications of model staffing patterns and desirable staff to

client ratios, and inventories of available manpower. None of this was available for use in identifying alcoholism manpower shortages. Nevertheless, it was considered important that the alcoholism factor play some role in connection with implementation of the various programs in the Educational Assistance Act.

The work that was then going on in developing relative prevalence estimates among States offered a feasible approach to this problem. Accordingly, it was argued that individuals with alcohol problems consumed a substantial portion of total health care resources. For example, estimates were available indicating that 20 to 25 percent of all hospital beds are occupied by alcoholics and that 17 percent of the physician's practice involves alcoholics. In addition, alcohol admissions in one study represented 47 percent of all male additions to State and county mental health hospitals during a one-year period.

Thus, treatment of alcohol-related problems pervades the service of all primary health care physicians and psychiatrists. It was proposed that alcohol-related health manpower shortage areas be identified in terms of added numbers of psychiatrists required to provide alcohol-related treatment in communities with a relative excess prevalence of alcoholism. This assumed that requirements for numbers of psychiatrists to treat the mean level of alcohol problems were included in the general manpower requirements enunciated by the Public Health Service.

This proposal was generally accepted. The "interim final" regulations for designation of areas having shortages of psychiatric manpower states that one criterion for eligibility is that an area has an unusually high need for mental health services. One such unusually high need is stated as follows:

> A high prevalence of alcoholism in the population, as indicated by a relative prevalence of alcoholism problems which exceeds that in 75 percent of all catchment areas (or other complete set of areas for which the prevalence index is computed), using the index of relative alcoholism oprevalence developed by the National Institute on Alcohol Abuse and Alcoholism for the purposes of allotting funds under 42 U.S.C. 4571. (Federal Register, Vol, 43, No. 6, Jan. 10, 1978, p. 1592).

The index of relative alcoholism prevalence had been developed on a State basis. However, these manpower shortage areas had to be defined for much smaller geographic units. Psychiatric manpower requirements were being calculated for Community Mental Health Center (CMHC) catchment areas, so that the same units had to be used for alcoholism purposes. The National Institute of Mental Health maintains a Mental Health Demographic Profile System on a catchment area basis. These data were used for calculating the FHD and CTC indices. The same categories of the population were used as had been identified by the AID procedure for the States.

However, no education or income information was available, so that these categories were dropped from the calculation.

There are approximately 1,500 CMHC catchment areas, the top 25 percent of which are to be considered as representing alcoholism manpower shortages. Table 3 shows a comparison between the States represented in this top 25 percent compared to the top 13 States identified in the State calculations. It can be seen that 10 States appear on both lists and that there is some degree of correspondence of their rank order (the catchment areas list is based on the numbers of catchment areas in the top 25 percent by State, so that the State with the largest number of designated catchment areas, California, is first on the list).

Again the regulations specify only the methodology to be used and not the specific data. The currently available list of shortage areas has not been subjected to thorough analysis for its reasonableness. Neither are the comments available made in response to publication of the proposed regulations. However, as new data become available and as greater understanding is achieved of the relationships among the demographic variables at the local level and indices of alcohol problems, new calculations will be made.

3. Estimating the Number of Persons Needing Alcoholism Treatment Services

One last example will be discussed briefly to illustrate use of synthetic estimation of alcoholism prevalence in yet another area of application. Increasingly, at all levels of government, pressure is being brought to bear on service providers to estimate the number of people who might need and could use their services. Marden reviewed this situation at the request of NIAAA and proposed a solution based on the use of synthetic estimation.

A review was made of 385 proposals for grant funds to provide direct services. Forty-three percent included no estimate of the number of alcoholics in the service area; another 18 percent provided estimates with no indication of their origin. Table 4 describes the remainder. As can be seen, a diversity of techniques are used, many of them quite crude.

It was argued that any proposed remedy to this situation should take into account several considerations. Prescribed procedures for developing the estimate would have to be appropriate for use by service individuals lacking in experience with research techniques. The procedures should be flexible and easily modified as additional pertinent information became available. And data requirements should reflect the availability of data in local areas.

A Population Matrix and a "Problem Drinker" Matrix were developed. The Population Matrix had dimensions of sex, age, and occupation. The cells of this matrix were to contain the size of the population

in that geographic area that corresponded to the designated demographic characteristics (e.g. , the number of male sales workers aged 20-29 living in that area). The "Problem Drinker" Matrix had the same dimensions. However, the cells contained the proportions of the various subpopulation groups whose score in the Cahalan problem scale exceeded a threshold value. Estimates of these proportions were obtained from the national household surveys conducted by Cahalan. To estimate the number of people in a given area with alcohol problems one had only to get the local population breakdown, multiply it by the "Problem Drinker" Matrix and add up the cells.

This application of synthetic estimation is similar to the preceding two in that primarily the method rather than the specific data is being prescribed. It differs in producing an estimate of the actual prevalence of alcohol problems. The other applications provided only relative estimates, a somewhat easier task. Marden's approach has been widely used but the results of this use have not been carefully studied.

CONCLUSION

The use of synthetic estimation techniques has permitted the NIAAA to respond to congressional mandates and take initiatives not otherwise possible. The methodology seems to have been accepted by government policy makers, the general public, and to some extent, at least, the technical community. It could be argued that synthetic estimation is an elegant stopgap measure either to be used until more direct information can be obtained or to replace more expensive direct estimation whose added value is questionable.

TABLE 1

RATES OF PROBLEM DRINKING AMONG
U.S. DRINKERS, BY DRINKING POPULATION 1973-1975

Drinking Population	Percentages For Each Survey			
	Mar. 1973	Jan. 1974	Jan. 1975	June 1975
All Drinkers				
No problems	64	700	65	63
Potential problems	26	24	24	26
Problem drinkers	11	6	10	10
Males				
No problems	57	66	62	57
Potential problems	29	27	23	31
Problem drinkers	14	8	15	13
Females				
No problems	74	77	70	73
Potential problems	21	19	27	21
Problem drinkers	5	4	3	6

SOURCE: Paula Johnson, David Armor, Susan Polich and Harriet Stambul, U.S. adult drinking practices: time trends, social correlates, and sex roles. Draft report prepared for National Institute on Alcohol Abuse and Alcoholism under Contract No. ADM 281-76-0020 July, 1977.

NOTE: A problem drinker experienced four or more of sixteen problem drinking symptoms frequently or eight or more symptoms sometimes;

a potential problem drinker experienced two or three of sixteen problem drinking symptoms frequently or four to seven symptoms sometimes.

TABLE 2

HYPOTHETICAL SYNTHETIC ESTIMATES FOR CTC

	Subgroup	Mean CTC Index	Proportion of State Population in Each Subgroup
1.	Black males 35+	.602	.046
2.	Black males 21-35	2.034	.012
3.	White males 65+	.200	.048
4.	White males under 65 who are married or were never married	.450	.378
5.	White males under 65 who were previously married	.980	.010
6.	Black females	.490	.063
7.	White females living in Pacific region	.423	0 *
8.	White females 65+ living outside Pacific region	.035	.069
9.	Previously married white females under 65 living outside Pacific region	.395	.369
10.	Married or single white females under 65 living outside Pacific region	.151	.005
			1.000

Synthetic estimate:
CTC = .602 x .046 + 2.034 x .012 + .200 x .048 + .450 x .378 + .980 x .010 + .490 x .063 + .423 x 0 + .035 x .069 + .395 x .369 + .151 x .005 = .421

*This value is zero since the hypothetical State is not in the Pacific region. If the State is in the Pacific region, this value would be the proportion of white females in the State's population and the proportions in subgroups 8, 9, and 10 would all be zero.

TABLE 3

LISTING OF STATES IN ORDER
OF DECREASING RELATIVE PREVALENCE
DOWN TO THE 75TH PERCENTILE

Based on Manpower Shortage Calculations*	Based on "Needs" Estimate Calculations
California	Alaska
New York	District of Columbia
Washington	Hawaii
Oregon	New Jersey
Illinois	California
Louisiana	New York
Pennsylvania	Pennsylvania
Alabama	Washington
New Jersey	Louisiana
Texas	Mississippi
Alaska	Oregon
Mississippi	Alabama
Michigan	Nevada

*Catchment areas in the top 25% of relative prevalence were tallied by State. States were then ranked in order of the number of catchment areas listed.

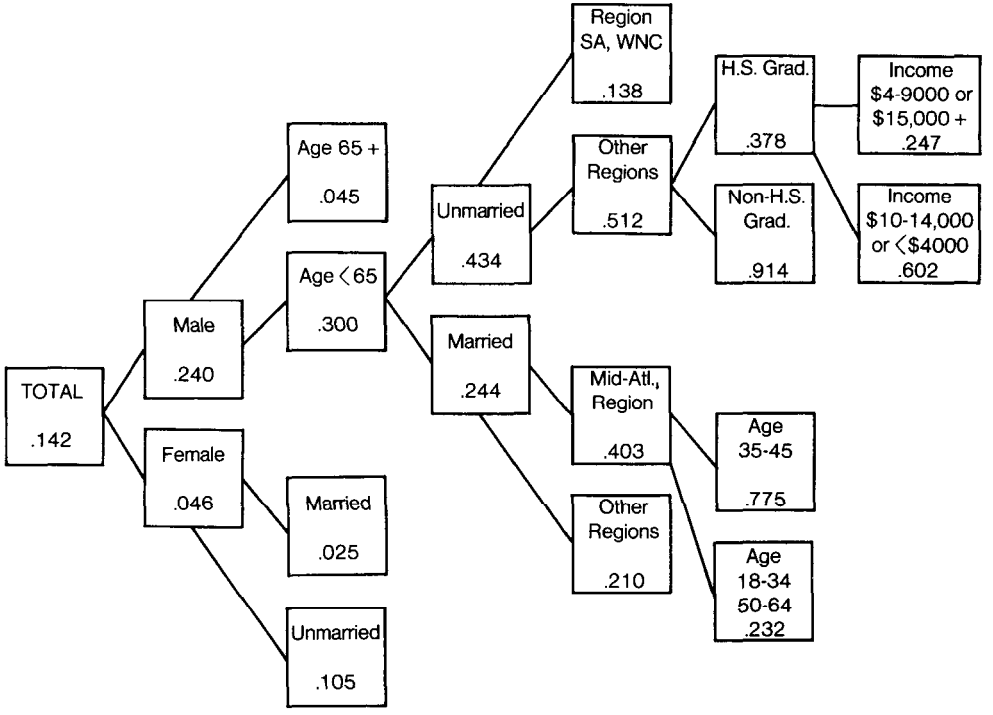
TABLE 4

METHODS OF ESTIMATING THE NUMBER OF ALCOHOLICS USED BY FUNDED PROPOSALS

	Number of Proposals	Percent of Proposals
Jellinek Formula	40	23.3
Agency or Other Records	55	32.2
Arrest Records	29	
Unemployment Figures	17	
State Mental Health Statistics	9	
Percentage of Population	61	35.7
Percentage of Adult Population		
10.0	14	
8.0	5	
5.3	3	
5.2	5	
	6	
5.0	11	
3.8	6	
Percentage of Total Population		
6.0	2	
4.4	3	
2.5	5	
Percentage of Low Income Population		
6.5	1	
Sample of Population	15	8.8
	1712	100.0

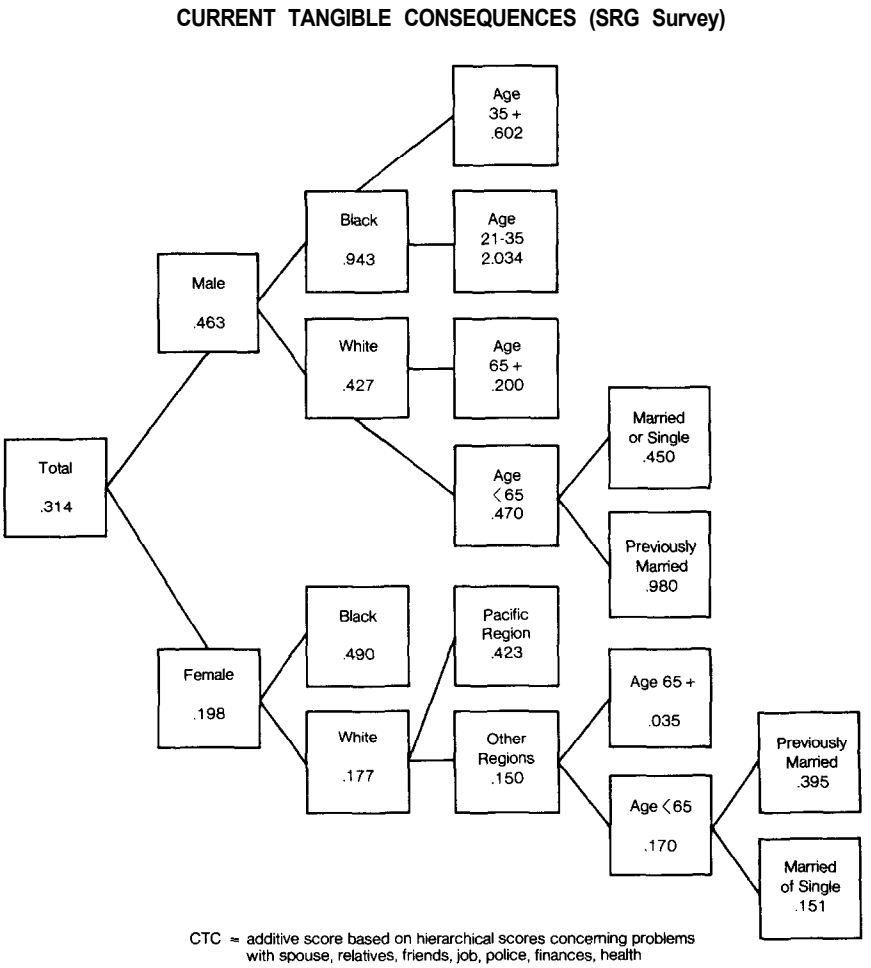
FIGURE 1

FREQUENT HEAVY DRINKING (Harris Survey)



FHD = number of times per week a respondent drinks 5 + drinks on one occasion

FIGURE 2


CURRENT TANGIBLE CONSEQUENCES (SRG Survey)
Total
.314
Male
.463
Female
.198
Black
.943
White
.427
Black
.490
White
.177
Age
35 +
.602
Age
21-35
2.034
Age
65 +
.200
Age
< 65
.470
Pacific
Region
.423
Other
Regions
.150
Married
or Single
.450
Previously
Married
.980
Age 65 +
.035
Age < 65
.170
Previously
Married
.395
Married
of Single
.151
CTC = additive score based on hierarchical scores concerning problems
with spouse, relatives, friends, job, police, finances, health

REFERENCES

Cahalan, D. Problem Drinkers. San Francisco, Jossey-Bass, 1970.

Cahalan, D. "Some Background Considerations in Estimating Needs for States' Services Dealing with Alcohol-Related Problems," paper prepared for presentation at Conference on "Need" Methodology for Formula Grants, HEW, 1976.

Edwards, G.; Gross, M.M.; Keller, M.; Moser, J.; and Room, R., ed. Alcohol-Related Disabilities. Offset Publication No. 32, Geneva: World Health Organization, 1977.

Harris, L., and Associates, Inc. Public Awareness of the National Institute on Alcohol Abuse and Alcoholism Advertising Campaign and Public Attitudes Toward Drinking and Alcohol Abuse. Phase 0: Fall, 1971. Study No. 2138; Phase One: Fall, 1972. Study No. 2224; Phase Two: Spring, 1973. Study No. 2318; Phase Three: Fall, 1973. Study No. 2342; and Phase Four: Winter, 1974 and Overall Study. Study No. 2355. Reports prepared for the National Institute on Alcohol Abuse and Alcoholism.

Keller, Mark. "Problems of Epidemiology in Alcohol Problems," Journal of Studies on Alcohol, Vol. 36, No. 11, 1975.

Marden, Parker G. "A Procedure for Estimating the Potential Clientele of Alcoholism Service Programs," Prepared for the Division of Special Treatment and Rehabilitation Programs, National hstitute on Alcohol Abuse and Alcoholism.

Rappeport, M.; Labow, P.; and Williams, J. for Opinion Research Corporation. The Public Evaluates the National Institute on Alcohol Abuse and Alcoholism Public Education Campaign, Vols. I and II, July 1975.

Sonquist, J.A., Morgan, J.N. The Detection of Interaction Effects, Monograph No. 35, Michigan: Survey Research Center, Institute for Social Research, 1964.

Sonquist, J.A.; Baker, E.L.; and Morgan, J.N. Searching Structure, Michigan: Survey Research Center, Institute for Social Research, Revised ed., 1973.

Discussion

Donna O. Farley

Following a full day of discussion of the statistical design, characteristics, and limitations of synthetic estimation, I am finding that some of my own questions and concerns about the method have been reinforced by the experts of the field. Therefore, my discussion will address somewhat philosophically several of my questions, while focusing on the need for an estimation method by many people in the health field, and on the growing tendency to use synthetic estimation regardless of its limitations.

I am trained in environmental health, and my perspective reflects that training. The work I have done with synthetic estimation, which I will summarize briefly, was for the purpose of developing an instrument that could be used as part of environmental health impact assessments,

But first there are several points that have already been made by many of the speakers, which I would like to reiterate, with a slightly different viewpoint:

1. There is, as we know, a growing demand for small area estimates. That demand is coming from local sources as well as national, and the areas involved are often of very small geographical popopulation size. I can cite the health planning agency for which I am presently working as an example of that local initiative. There are at least three different demands for local estimates within the agency. These include a) needs assessments for review of projects under Certificate of Need or for grant applications, b) the internalagencyneed for morbidity estimates, and c) estimates for use in problem identification as part of our planning process.

2. The appeal of synthetic estimation will probably tend to make it a preferred technique in the field. It can be easily conceptualized, is adaptable to many different data sets, and can be used readily by practitioners without extensive statistical training. 'The latter characteristic is one I feel should be emphasized here, because expertise such as that around this table is not always readily available to assure judicious local use of this uncertain method.

3. Research findings have demonstrated the wide variation in

the validity of synthetic estimates. This variation indicates that the method should not be used casually, but with perhaps a conservative approach, recognizing that while synthetic estimation may estimate some variables well, for others it will be much less effective. The people in the field need to be kept aware of that fact.

In reviewing David Promisel's paper I observed that the case studies he describes offer excellent examples of these three points. All three applications in his paper resulted from governmental mandates tied to the distribution of dollars. The very different approaches used demonstrated the flexibility in application of synthetic estimation; but because local direct estimates were not available for validation, the estimates themselves must be considered to be at the least uncertain. They filled a need, however, and quite possibly are the best estimates around at this time.

My own work with synthetic estimates also filled a need, although not one that involved financial allocations. In order to estimate the potential health impacts of air pollution, one needs to know the number of people exposed to the pollution, the severity of the exposure (measured as dosage if possible), and a dose response relationship which will convert the dosage to estimated health effect, People with certain chronic health conditions are at high risk to such exposures, and therefore should be included in a health impact assessment. Among the important high risk groups are those with chronic bronchitis, emphysema, and asthma (chronic obstructive lung disease). We needed a method to estimte the number of people in those groups for local areas.

Using national prevalence rate estimates from the 1970 Health Interview Survey and national mortality data from the vital statistics, synthetic estimates were calculated for death rates from these three conditions for about thirty (30) local urban areas. These were small areas of populations between 78,000 and 400,000. Validation of the estimates with local mortality data showed they were biased and imprecise, and the variables of age, race, and sex accounted for only a small portion of local differences in rates. However, when compared to estimates based on local application of State level crude death rates for the diseases, the synthetic estimates were the better estimates of local death rates.

The local estimates needed for our work, though, were prevalence rates rather than death rates. Yet the validity of synthetic estimates of prevalence could not be evaluated without local direct estimates of prevalence. Although we recognize that limitation, synthetic estimates of the local prevalence of the three conditions have been used in subsequent work,with the intuitive expectation thattheyare better estimates than those based on the local application of national level crude prevalence rate estimates.

Another phenomenon was observed during the validation work with the death rate estimates. The synthetic death rate estimates tended to

cluster around the mean, not showing the same local variation as the actual local death rates. This characteristic has also been mentioned several times in this workshop. In order to take advantage of the available local mortality data, another approach to prevalence estimation was developed. A synthetic estimate of the ratio of cases to deaths for a disease was calculated for an area, then to be applied to the actual local death rate, yielding what was called a "death rate conversion" estimate of the local prevalence rate.

The assumptions underlying this approach are 1) that the cell specific case fatality experience among those people with a disease is at least as consistent as the prevalence or death rates for the same cell, and 2) that building the estimates from the actual local death rates would bring into the estimate the influence of local variables that are not reflected in the regular synthetic estimates of prevalence. It is an appealing approach intuitively, and I ask your comments on it. This method has been used also, whenever mortality data were available, for chronic bronchitis, emphysema, and asthma. It consistently yields smaller prevalence estimates for these conditions than does synthetic estimation of prevalence.

In summary, I would like to address an underlying issue of the workshop, which already has been discussed at length. The studies described in David Promisel's paper show that approaches using synthetic estimates of either relative or absolute values can be and are being used quite freely for various demographic data bases. Similarly, his paper and my own efforts show that a variety of approaches can be designed for producing local synthetic estimates. If the user can expect that those estimates will be better than those from more crude methods, synthetic estimation will probably be used -- for better or for worse. The use of synthetic estimation will probably increase, with people of various levels of training in diverse disciplines applying it to their own specific problem. Recognizing this, we need an answer to a very practical question:

> How freely can synthetic estimation be used for different variables and for different geographical areas; and perhaps more importantly, what modifications or adjustments can be made in its application to enhance the validity of the local estimates?

This is not a new question, nor by any means a simple one, but I ask it with the perspective ofauser of the method who is aware it has limits. There are growing numbers ofuserswho needtobekeptaware of its practical limits, its capabilities, and the ways in which its use can be optimized. Those of you here who are the collective "parents and guardians" of synthetic estimation are the ones who can help provide that guidance.

General Discussion

* Donna Farley's use of synthetic estimates raises an interesting point. Her problem was to try to devise synthetic estimates for prevalence of chronic obstructive lung disease. She used death data to estimate the deaths and then had to convert that to an estimate of prevalence. One of the problems is: How good is NCHS data from two sources: HIS or HES estimates of the prevalence of, say, chronic obstructive lung disease? Can they be used with data on deaths that are also diagnosis-specific but from a different data system? I would like to pose the further question: Is this a useful area on which to put more emphasis for estimating case fatalities? This is an area of extreme importance to epidemiologists.

* We don't feel that we know as much as we should about the validity of classification of causes of death, particularly as it varies from one place to another. NCHS is, in fact, doing some studies now. Some work has been done in the past for selected diseases, but the thought is to have a more systematic attempt to evaluate the quality of the classification of causes of death. We are thinking of it primarily in terms of national statistics. Measuring validity for local areas will be even tougher than producing prevalence estimates for those local areas.

In the broader perspective, what we're talking about is: What kind of data do we have at the local area level in addition to complete count data from the decennial census? There are vital statistics on a complete count basis for local areas. The registration system provides the advantage that vital registration is a continuous system. The statistics are available on an annual basis. It might be interesting to compile a listing of the kinds of statistics that are available for local areas with some consistency and therefore are potentially useful for synthetic estimates.

Measuring the prevalence of disease is one that interests NCHS very much. The primary instrument that has been used is the Health Interview Survey. Securing diagnostic information in a personal interview is subject to serious quality limitations. Now a completely different kind of survey approach is being explored--a survey of medical sources.

The hope is by that means to get diagnosed cases of disease. Some fairly large studies are being done now trying to estimate disease prevalence by means of surveys of medical sources (including physicians and hospitals and other places that provide care). The area of collecting data on prevalence, whether you're talking of drug use or alcohol use, or chronic diseases, is a very difficult one. Over the next five or ten years, perhaps, survey methodology will be developed. For local area data, one part of the system is to develop hospital discharge statistics within each of the States and then build up to the national level. If that kind of approach is productive, eventually there will be much greater information at the subnational, State, and perhaps the local level.

* As we how there are some administrative programs that have operating data in the same area, e.g., the Medicare program. There also are abstracted data from various hospital-based systems. There are now two reports by the Institute of Medicine (1977a,1977b) that deal with the quality of coded diagnostic data. One is for several abstracting services collectively, and the other deals with the Medicare system.

* Yesterday there was some discussion about the desirability of statisticians providing comments to Congress regarding the feasibility of compiling certain types of data. I'd like to reinforce the need for such activity. It extends beyond Congress. Congress imposes demands on the Executive Branch. Within the Executive Branch we impose demands on State and local governments. There are two kinds of problems. One is: Is the question reasonable? For example, we request estimates of relative prevalence, but that is not what the law asked for. The law asked for need, not prevalence. Someone arbitrarily equated the two (and probably had difficulty in defining the term, let alone estimating it). So perhaps that is not a reasonable question. Is the request to identify need a reasonable one? If it is reasonable, it has to be couched in very careful terms. For example, there may be quite a difference between estimating the number of people who have a certain ailment and what is the need for a particular service as a result. Even if the question is reasonably posed, there is the question whether it can be answered. There is no bulwark against the flood of demands for information. Hopefully, it doesn't do too much harm; but are we sure?

* Perhaps the following idea is responsive to the concerns that have been raised. There are two basic issues that we have talked about from time to time. One is, how to produce different kinds of local estimates given certain kinds of data sets. The other issue is, how to provide some sort of advice to policymakers who would like us to help them make a decision. To some extent there are certain limits on the latter issue depending on the data that is available.

Let us consider a design which a consulting statistician might suggest that possibly would assist a policymaker trying to make a decision. Consider splitting resources available among three different kinds of research designs. In the first design, a national survey would be conducted to obtain data by personal interview, but of only moderate depth, say, a one-hour interview. Second, consider the use of a selected set of observational studies (similar to the types of multiclinic

clinical trial type designs that are used in a lot of experimental situations) where you would pick selected sites in local areas of interest. You would try to do in depth studies of a lot of variables, trying to produce for any given local area the best estimate for that area that significant expenditures would buy. You would try to identify variables which were good correlates to the variables that you were most interested in --variables that were easy to measure, or variables that you could perhaps obtain by some sort of a telephone interview survey. Third, you would follow that up with a survey that would encompass every local area, either a telephone survey or collection by any other approach that could be quick and easy. If you could spread the resources among these three things, that would be something which possibly would be better than putting all of your resources into any single one of them and having the limitations that would apply to any one of them, whether it would be cost, ability to estimate, or feasibility. It appears to represent a statistically cost effective way of trying to address a policymaking question. Knowing the overall budget one can experiment with different design strategies.

* A subcommittee of the Federal Committee on Statistical Methodology has prepared a "Report on Statistics for the Allocation of Funds." This report, issued by the Office of Federal Statistical Policy and Standards (1978) looks at five specific case studies of distribution of Federal funds to local areas and then tries to generalize on the problem.

* The previous proposal for a three activity research design is similar to some thinking that the NCHS has been doing. First, there is under consideration a telephone survey capability, using random digit dialing. This would eliminate listing and other expenses that go with selecting an area sample. If you consider that the areas of interest are States, NCHS is thinking of a dual frame survey using the existing HIS as the first frame of the dual frame survey. The other frame would be based on telephone random digit dialing. There is some work that has to be done to decide what the sample design of the telephone survey would have to be in each State to adequately supplement the existing interview survey. This will vary by State because the PSUs and sample sizes in HIS vary by State. For local area statistics the strategy is twofold. One, a telephone survey random digit dialing manual is being prepared that will be available to local areas. This manual should facilitate efforts of those who want to do such surveys on their own. In the field of health, there isn't any mandate for annually produced statistics for as many areas as for revenue sharing. Some areas seem to be much more advanced in their thinking than others. In addition, there is the possibility of having NCHS have the capability of conducting local area-surveys from Washington. Thus, if a particular area could-not do its local area telephone survey. NCHS would have the capabilitv of doing it. There are a number of problems that have to be worked but.

*What is the role of OMB in regard to our discussions in terms of the approaches, the level of interviewing that would be permitted, and so forth?

* OMB has a role whenever funding gets involved. In regard to design and issues of statistics needed, and how you're going to get them,

there is some involvement with the responsibility of review and statistical clearance.

* Agencies are questioned about the extent of survey work. OMR needs to concur with the approach to obtain data by survey.

* A lot of the decisions are made by a Department clearance office and are reviewed at OMB.

* We should note another point concerning the recent work on random digit dialing. If it turns out that random digit dialing is going to lower the costs of surveys quite a bit and if there are manuals available, will it be a problem of local area survey proliferation?

* We'll have to wait and see what the savings really are.

* It's likely to be, say, three to one.

* Are populations without telephones covered by the estimated reduction factor of three to one?

* It depends. You would want to cover nontelephone households at a lower sampling rate. Therefore, the reduction in overall costs depends on whether the rate of subsampling of nontelephone households is one in three or one in five. So it's hard to provide a unique answer. In terms of one experience, lately, with telephone you can probably figure on one third or a half reduction, or something of that order of magnitude.

(Contributing to the general discussion during this period were: Maria Gonzalez, Gary Koch, Paul Levy, David Promisel, Monroe Sirken, Joseph Steinberg and Joseph Waksberg.)

REFERENCES

Institute of Medicine Reliability of Hospital Discharge Abstracts. Washington, D.C.: National Academy of Sciences, February 1977a.

Institute of Medicine, Reliability of Medicare Hospital Discharge Records. Washington, D.C.: National Academy of Sciences, November 1977b .

Office of Federal Statistical Policy and Standards, Statistical Policy Working Paper 1, Report on Statistics for Allocation of Funds. Washington, D.C.: U.S. Department of Commerce, 1978:

Synthetic Estimates as an Approach to Needs Assessment: Issues and Experience

Charles G. Froland

ABSTRACT

An overview of a study which applied the synthetic estimates technique to derive rates, numbers, types and characteristics of potential clientele for substance abuse related programs in the State and counties of Oregon is presented. A brief description is given of the methods utilized to obtain estimates as well as the means for examining their validity.

In as much as the objective of the study was to provide useful information to State and local program planners and adninistrators, the experience of utilizing the study's findings is presented. Several applications are highlighted to indicate the range of ways in which the study was utilized. The experience of applying the results in a program and policy context surfaced several issues concerning the requirements for validity and accuracy, specificity and, finally, the role of synthetic estimates in needs assessment. The experience suggests that the information derived by this technique will be most useful if integratedwitharange of other types of information, both quantitative and subjective.

INTRODUCTION

The value of quantitative information about a community's substance abuse problems has been well recognized by planners, providers, and policymakers. While such information is not often available in many communities, this has generally not been for lack of interest or expertise. Barriers to obtaining estimates of a population at risk for substance abuse treatment have generally involved the prohibitive technical and resource demands associated with producing accurate and timely information about the nature and extent of substance abuse within a given community, issues of confidentiality, and fundamental disagreement regarding the definition of abuse, dependency, or addiction. As one consequence of these basic limitations, decisions about programs and policies are often made without the benefit of quantitative statements of the

size or scope of a community's needs for substance abuse treatment resources. To be sure, information of a quantitative or "scientific" nature is clearly not the only input into the policymaking process if resulting plans are to be politically acceptable (Lindbloom 1973). Values, political interests, community norms and other considerations perhaps form a set of n-ore immediate policy determinants of which information must be seen as only one contender. On balance, the promise that information about substance abuse problems holds in this context is to provide a common and valid frame of reference for discussing values and interests. Without such information, policy is likely to be created solely in response to impressions, status quo and/or political interest groups, making it difficult to determine whether the needs of the community are being addressed or met.

In the abstract, the development of effective policies and programs must be based upon a clear understanding of: (a) the nunbers of individuals potentially needing substance-related services, i.e., potential clientele, (b) their characteristics: and (c) the types of substances being used. Given this information, policymakers planners, administrators and citizens may, for example, be guided in the allocation of resources to various types of services, the determination of the appropriateness of existing services in meeting a community's needs, and the identification of target populations needing services. however, the task of directly obtaining information about the nature and extent of substance abuse problems within a community is usually beyond the technical or financial capabilities of most local and State jurisdictions. As a result. most local planners have typically adopted a number of indirect strategies for developing needs assessment information including, for example, interviews with key city representatives, canmunity forums with local residents, or using available data concerning rates of arrest, emergency roan admissions, cirrhosis deaths, and other drug-related deaths. At best, such indirect and inexpensive approaches yield a global but useful picture of the needs in a community. Most often, these methods are not entirely satisfactory for deriving an evenhanded estimate of the size of substance-related problems and need for services.

The synthetic estimates method offers the promise of a useful alternative. To the extent that existing survey data are available which adequately reflect conditions in a planning area, reasonable estimates can be obtained of the number of problems that might be expected to occur given the geographical and sociodemographic mix of the area. Although not without a number of technical limitations, the technique was considered to have sufficient merit that it was applied as one part of a study of substance abuse needs in the State of Oregon. The study was conducted by a research arm of the Oregon Mental Wealth Division in 1976. The results of the study are reported elsewhere (Froland 1976). What is presented here is an overview of the approach taken in deriving synthetic estimates for the counties and State of Oregon as well as findings related to the appropriateness of the resulting information. Additionally, some discussion is given to the uses made of the information by program planners at the State and local levels. Finally, a number of issues regarding the utility and potential applicability of

synthetic estimates for needs assessment can be shared, based on the experience of Oregon.

STUDY OBJECTIVES

The study was conducted to serve a number of audiences. The prinary objective was to derive information that could satisfactorily address questions at both local and State levels of planning concerning the accessibility, appropriateness and adequacy of service efforts. The State Legislature wished to know whether too much or too little money was being spent on substance abuse treatment. State and regional planning specialists responsible for approving county plans and allocating legislative appropriations wanted to know which counties had the greatest need as well as the nature of local substance-related problems. County administrators and program directors not only were concerned with whether they were getting their fair share but also whether they were serving clients who were in some manner representative of the nature of their community's needs. Given this range of questions, three types of information were considered necessary. Estimates of the numbers of potential clientele for each county would address legislative and local concerns as to the adequacy of resources allocated to counties. Descriptions of the varying degrees of different classes of substance abuse within each county's population would permit State and local planners to assess the appropriateness of different mixes of service modalities in dealing with the communities' problems. A third type of information, estimates of the sociodemographic characteristics of the population of potential clientele, would allow local programs to assess the representativeness of the problems of clients actually served. Since the synthetic estimation technique could be based on a body of existing survey data that could provide rates of different classes of substance abuse specific to different demographic subgroups within a population sample, the method was capable of yielding these three types of information.

APPROACH

For use of the technique in Oregon, a search was undertaken to determine the best source of survey information. Several broad questions were considered in choosing among alternatives: (1) What population base is the survey information representing? (2) What are the technical attributes of the survey? (3) What kind of informtion does the survey provide? Several sources were consulted which consistently indicated that the survey of greatest utility would be one conducted for the National Institute on Drug Abuse by the Social Research Group at George Washington University (1975).

First, the survey could provide information on a general population sample. The survey was administered to a nationwide sample of youths (aged 12-17) and adults (aged 18 and over). Survey information was available for the Western region of the United States. which included Oregon. On these grounds, such a general population sample focus was felt to be appropriate to the general population of Oregon. Second, the survey was technically

acceptable. It was timely, having been conducted in 1974 and published in 1975. (The study which developed and used the synthetic estimates was conducted in 1976.) The survey protocol was administered by trained interviewers except for a self-report section. A reliability and validity study was conducted which demonstrated acceptable results. The sample size was sufficient to provide acceptable error rates in the survey information. Finally, the survey questionnaire items were appropriate for Oregon's purposes in that they covered a wide range of different types of substances; they were behaviorally focused and included a sufficient breadth of items to estimate current potentially abusive patterns of use. Beyond this, the information was tabulated for specific categories of several demographic and residence characteristics of the sample. Thus, the survey addressed all of our questions of acceptability to a satisfactory degree.

CASENESS

While the survey was concerned with identifying use patterns for many licit and illicit substances, it was not expressly concerned with identifying abusive patterns or individuals with a potential need for service by reason of their use of a substance. In general, the survey was simply concerned with providing information about varying degrees of frequency, duration and amount of use for a wide range of substances, some of which are illicit and/or potentially hamrful. No information was provided as to the extent to which such use patterns implicated reduced physical, interpersonal or social functioning. Thus, the first task was to identify that canbination of frequency, duration and amount of drug use which could be used to approximate "caseness," i.e., an individual with a potential need for substance-related services.

To some extent, the study relied upon common operational definitions used in sociobehavioral research (Elinson and Nurco 1975). Beyond this, a common decision rule was to define the use patterns of those individuals with a potential need for services on the basis of the most extreme patterns of use in terms of high frequency and amount with indications of problematic duration. Additional considerations involved adjusting for use of other types of drugs, i.e., polydrug abuse. Table 1 shows the resulting definitions.

COMPUTATIONAL OVERVIEW

Having identified and adjusted survey rates to reflect expected levels of potential clientele specific to various demographic and residence characteristics, the next step involved applying these expectations in respect to the population base of the 36 counties of Oregon. Essentially, the approach taken could be characterized as actuarial. In general, this involved weighting the rates provided by the survey according to the respective characteristics of each of the counties. Several general steps were followed: (1) First, adjustments were made on the basis of urban/rural mix for each county. (2) Next, area-adjusted rates for each category of a demographic characteristic were weighted by corresponding census distributions of such characteristics.

Four characteristics were employed: age, sex, race, and education. (3) Finally, area and demographic adjusted rates were weighted and summed to obtain an overall rate for a given class of distribution. To find numbers of potential clientele, rates were simply multiplied times the updated population count for each county.

Thus, the results yielded the three types of information desired: number, types of problems, and demographic distribution associated with potential clientele.

TABLE 1

Definition of Use Patterns

Alcohol	Drank average of nine or more drinks each time they drank in past month,* and/or drank every day in past month.
Opiates	Used three or more times in past month and/or used in past month and will use again.
Depressants	Used five or more times in past month and/or used in past month and will use again.
Stimulants	Used five or more times in past month and/or used in past month and will use again.
Other	Used cocaine, inhalants and/or LSD nine or more days in past month* and/or used in past month and will use again.

*Youth sample used the following: Drank five or more times in past month an average of five or more drinks each time.

**Youth sample used the following: Five or more times in past month.

Results

The resulting rates and numbers of users with a potential need for substance-related services for the State of Oregon are shown in Table 2 for alcogol, opiates (heroin, illegal methadone and other opiates), depressants (barbiturates and tranquilizers), stimulants (amphetamine and nonamphetamine stimilants), and other drugs (psychedelics, cocaine and inhalants). The rates for each substance may be interpreted as indicating populations whose use patterns leading to a potential need stem primarily from the specific substance. For example, the figure of 16.7 persons per 1000 for other drug use refers to those persons who use principally either cocaine, psychedelics or inhalants in a manner which is indicative of a potential need for drug-related services.

TABLE 2

Statewide Rate and Number of Potential Clientele

Substance	Rate as a Percent of Population	Number (1975 Population)
Total	.0870	199,320
Alcohol	.0565	129,510
Drugs	.0305	69,810
Opiates	.0020	4,590
Depressants	.0027	6,240
Stimulants	.0090	20,710
Other drugs	.0167	38,270

The total Drugs, which excludes alcohol users, refers to the total of opiate, depressant, stimulant and other drug users whose use of one of these drugs is indicative of a potential need for substance-related services. The overall total reflects the addition of all individual substance classes.

APPROPRIATENESS OF ESTIMATES

In extracting rates from the national survey and applying them to Oregon's population, the assumption was made that the survey's information was applicable, i.e., rates for Oregon would not differ markedly from other States in the Western region of the United States. Such an assumption was obviously open to question. Beyond this, the inability to precisely define a rate of substance abuse from the survey that would refer only to substance abusers who were clearly in need of treatment easily creates suspicion of the estimates that had been developed. These potential objections and others demanded that some means be developed to determine the extent to which the estimates approximated actual conditions. Since the prime reason for estimating numbers of potential clientele by the synthetic estimates technique was the absence of such information, an indirect method for examining the appropriateness of the estimates had to be explored. The approach relied upon the substance abuse problem indicators shown in Table 3. Composite indicators were developed within each class of substance-related problems and correlated with the corresponding synthetic estimate for that class. The results, shown in Table 4, demonstrate that with the exception of depressant-related problems, the synthetic estimates are significantly correlated with the distribution of the level of problems observed in the 36 counties of Oregon.

One aspect of the appropriateness of the estimates could not be addressed definitively. While the estimates of potential clientele appeared to be distributed across counties in a manner that was reasonably close to the distribution of the level of substance-related problems, the magnitude of the estimates could not be readily substantiated. With regard to alcohol-related problems, the estimate of approximately 130,000 alcohol abusers compared favorably with that derived by the Jelinek method (102,500) that appeared in the Oregon State Alcohol Plan for 1976-1977.

TABLE 3

Substance Abuse Indicators

Substance Class	Indicators
Alcohol	alcohol-related emergency roan admissions;[1] percent of traffic accidents with blood alcohol involved;[3] alcohol sales;[5] State hospital admissions diagnosed alcoholic;[1] cirrhosis deaths.[4]
Opiates	opiate-related emergency room admissions;[1] opiate-related arrests (State and Federal);[2] opiate-related deaths;[4] serum hepatitis.[4]
Depressants	depressant-related emergency room admissions;[1] depressant-related deaths.[4]
Stimulants	stim&nt-related emergency room admissions;[1] stimulant-related deaths.[4]
Other Drugs	other drug-related emergency room admissions;[1] other dangerous drug arrests (State and Federal);[2] other drug-related deaths.[4]

[1]Drug and Alcohol Program Office, Mental Health Division, Salem, Oregon

[2]Oregon State Criminal Justice Information System, Salem, Oregon; includes State and Federal agency cases

[3]Oregon Department of Motor Vehicles

[4]Oregon State Department of Health, Portland, Oregon

[5]Oregon State Liquor Control Commission

TABLE 4

Correlation Results for Potential Clientele Estimates and Substance Abuse Problem Indexes (N=36)

Substance Class	r	r^2	
Alcohol	.41	.17	$p < .05$
Opiates	.64	.41	$p < .05$
Depressants	.03	.001	NS
Stimulants	.37	.14	$p < .05$
other drugs	.61	.37	$p < .05$

An additional source of data provided estmates based on a different national survey of alcohol use patterns (Marden ND). While the demographic categories, as well as the problem definitions, were different from those employed in this study, the resulting estimates of total potential clientele for alcohol services based on the different survey was 125,000. Thus, the estimates of alcohol problems seemed to be in agreement with a number of sources. However, with regard to various categories of drug abuse, no similar figures were available for easy comparison 'The Oregon State Drug Plan for 1976-1977 estimated that, overall, close to 30,000 persons had a conspicuous involvement with drugs, i.e., "it had resulted in arrest, incarceration, admission to a hospital or treatment program, or death." This estimate was not directly comparable to the estimates provided by this study, since different use patterns and potential problems were involved. 'The addition of the categories of opiates, depressants, stimulants, and other drugs yields a total roughly twice that of the estimates for conspicuous drug users (69,816 versus 30,000). However, the study was also interested in "inconspicuous," as well as "conspicuous" or known substance abuse, so that it should not be surprising for the numbers of notential clientele to be higher than the number of actual substance abuse-related clientele.

APPLICATION

In part because of its intuitive appeal, and in part because of a well-designed dissemination strategy, the information was accepted and utilized by a wide variety of audiences. Staff of the State Legislature used the estimates to prepare testimony in appropriations hearings for alcohol and drug programs funded by the State. A number of local county programs utilized the information to target needs within their counties as well as to compare the characteristics of those they were serving with the demographic distribution of the estimates of potential clientele. By identifying specific population groups that were underrepresented in their service strategies, these programs were able to make decisions about new programs that were needed as well as the appropriateness of existing eligibility and admission criteria. The information was employed in both the drug abuse and alcoholism statewide plans for fiscal year 1976-1977.

Perhaps the most concerted and systematic use made of the information was by a pilot project undertaken across the State involving the development of county alcoholism plans. The project may serve here as a case study of what is possible in actually utilizing the information in planning services (see Hardison 1977, for more detail). The project was carried out by the Office of Programs for Alcohol and Drug Problems and involved using the estimates to establish a uniform process for defining service needs across all Alcohol Subcontract Service Providers funded through the Oregon Mental Health Division. The planning process was implemented across all counties in the course of their plan development by means of a series of steps.

First, ranges of the expected number of admissions to a program for a particular demoqraphic category were computed. For this purpose, a procedure was used that computed a 90% tolerance interval about the numbers of potential clientele for a given demographic category, adjusted by a utilization ratio formed by dividing total actual admissions by total potential clientele. The resulting computations are shown for an illustrative county in Table 5.

Next, the distribution of expected admissions was compared with the distribution of actual admissions to determine whether a particular group was wer or under represented in their utilization of services. High Priority Groups were identified by rank ordering under represented groups on the basis of percent need met, i.e., actual admissions divided by potential admissions as illustrated in Table 6.

A number of further steps were carried out to complete the planning process. These may be highlighted. Having identified the high priority population groups that existing services were considered to address inadequately, discussion groups comprised of program representatives were held to identify those reasons that might be at the base of the problem. Issues of geographic, cultural, and psychological accessibility were generally surfaced. Further steps involved identifying what modifications in existing service procedures or capacity might be implemented to deal with such problems. Finally, local planning groups were set to the task of formulating measurable objectives whereby needed changes in services would be carried out.

ISSUES

All in all, the experience at. the local and State levels demonstrated that the synthetic estimates could be of practical utility in structuring decision-making in policy and program development. During the course of working with planners and providers in Oregon, several key issues emerged which may be generalized as of common concern in choosing the synthetic estimates technique as a needs assessment tool.

Perhaps the major obstacle confronted in attempting to utilize the synthetic estimates technique in policy and programs concerned establishing some degree of understanding of what the

TABLE 5

Identifying Underserved Client Groups

Race/ Ethnicity	Estimated Number at Risk	MHIS* Admissions	Percent of Need Met	Range of Expected Admissions	Representation
Total	42,155	4,813			
Native American	1,278	601	17.03	124-168	Over
Black	2,203	218	9.90	222-282	Under
Spanish Language	1,452	115	7.92	142-190	Under
White	37,222	3,669	9.86	4068-4435	Under
Total Identified by Race	→	4,603			
Age					
12-17	984	9	.91	93-132	Under
18-21	1,590	112	7.04	157-207	Under
22-25	2,747	229	8.34	280-348	Under
26-34	5,822	770	13.23	611-719	Over
35-49	11,265	1,761	15.63	1205-1369	Over
50-64	17,863	1,425	7.98	1930-2151	Under
65+	1,884	102	5.41	188-243	Under
Total Identified by Age	→	4,408			
Sex					
Female	21,355	363	1.70	2315-2564	Under
Male	20,800	4,054	19.49	2254-2498	Over
Total Identified by Sex	→	4,417			

* Mental Health Information System

TABLE 6

High Priority Populations

Rank-Order	Population Descriptor	Percent of Need Met
1st	Age: 12-17	.91
2nd	Female	1.70
3rd	Age: 65+	5.41
4th	Age: 18-21	7.04

estimates really meant. On the one hand, the information was in a form that was intuitively appealing and those who most often had little or no data with which to compare were sorely tempted to use the estimates. On the other hand, the computational procedures used in the technique were somewhat obscure to the range of audiences to which the information was directed. Those responsible for policymaking were rightfully suspicious of information developed in a manner they did not understand. Concerns over the timeliness of the information compounded these issues. This may be endemic, to the extent the technique relies on secondary data which, allowing for dissemination lag, may be several years old. In situations where the estimates are attempting to describe a condition which is unstable or in flux, the resulting information may be rejected out of hand by those working in the field.

A number of issues hindered the precision of the technique. At sone level of demographic or geographical detail, the size of the survey sample limits the ability of the survey to maintain representativeness and accuracy in the information disaggregated to a local area. Additionally, the application of the survey rates to the population base of a community is limited by the size of the area. Issues concerning the nature of the survey used (e.g., sample characteristics, representativeness, sampling error), the sampling error of census information,and geographic uniqueness of the area all serve to limit the degree to which estimates for a small area may be accurate. These issues became more telling for areas which have unique characteristics or conditions. Such factors as special population pockets, geographical diversity and unique cultural features all serve to reduce the relevance of estimates based on more general expectations. For example, several counties in Oregon had Indian reservations, while others were influenced by transient farm labor. The estimates derived for these counties certainly could not adequately reflect the circumstances confronted by local programs serving such areas.

CONCLUSION

To what extent can synthetic estimates be relied upon in policy decisions, particularly if such decisions are to materially affect funding allocations, program operation, and the utilization of services? From one perspective, this question can be examined by assessing the technical validity or reliability of the estimates themselves. Here, limitations in the body of information used for prediction or estimation, errors in computational procedures, as well as the nature of the area to which the technique is applied all serve to define reasonable boundaries for the use of the estimates in decisiomaking. However, one has to distinguish between what is statistically acceptable and what is useful in practice. While achieving technical standards of validity usually heightens practical utility, infomation can be of practical use that my not meet standards of statistical rigor. In part, it is a matter of degree; more often it is a question of what alternative sources of information are available and whether they are more or less technically acceptable. While the synthetic estimates technique has much to recommend it as part of the methodological armamentarium of quantitative needs assessment, a broader view recognizes that the utility of synthetic estimates rests on their ability to inject an element of objectivity in policy decisions. The experience of Oregon suggests that when the information was used as a basis for discussion or combined with other information or perspectives, planning decisions were relatively more systematic and comprehensive. Thus, when the estimates were not taken as being exact and precise statements of commnity need but rather used to structure a closer examination which included the qualitative and subjective viewpoints of those working in the field, their role was more useful in motivating program changes and improvement.

REFERENCES

Abelson, H., and Atkinson, R. Public Experience with Psychoactive Substances (National Survey--NIDA; Main Findings 1974). Princeton: Response Analysis, 1975.

Elinson, J., and Nurco,D.,eds. Operational Definitions in Socio-Behavioral Drug Use Research. NIDA Research Monograph 2. DHEW Pub. No. (ADM) 76-292. Washington: Superintenden of Documents, U.S. Government Printing Office, 1975.

Froland, C. Substance Abuse in Oregon. Salem, Oregon: Oregon Mental Health Division, Department of Human Resources, 1976.

Hardison, J. Criteria for a Minimum Definition of Need. Salem, Oregon: Managrment Support Services, Oregon Mental Health Division, 1977.

Lindbloom, C. The Policymaking Process. New York: John Wiley, 1973.

Marden, P. A procedure for estimating the potential number of alcoholism service program clientele. Washington, D.C.: NIAAA, unpublished, no date.

Oregon State Alcohol Plan for 1976-1977. Salem, Oregon: Mental Health Program Office, Oregon Mental Health Division, 1977.

Discussion

Reuben Cohen

Most of us present at this workshop deal with large data bases and design research or provide data at the national level. Charles Froland's paper has added a significant dimension to our discussion. He has told us how real life decisions are made at the county and community level. For any given jurisdiction in which program funds are allocated, the number of dollars involved may be relatively small. But those local allocations aggregrate to millions of dollars and affect large numbers of human lives.

A recurring question has been posed at this workshop: What are the alternatives available to us? One message in Froland's paper is that there are few, if any, alternatives to making appropriate use of national survey data for needs assessment at the community level. Surveys adequate to the task of providing direct estimates of needs at the community level might cost as much as or more than the amount available for program use. Poorly conceived or loosely executed data collection procedures might be worse than none at all.

I am reminded that I was involved in planning and interpreting results of a national survey preceding the Presidential election of 1968. Since Joe Waksberg told his election story yesterday, I will tell mine today. Pre-election surveys may be unique in that, in addition to national samples, there are generally more State surveys than there are States, and the actual election results are available almost immediately to help evaluate the results of State as well as national polls. Many of you will recall that the pre-election poll results (and indeed the election itself) were very close to 50-50 between Nixon and Humphrey in 1968. Estimates for specific States became very important because electoral votes would actually elect the new President.

Just prior to the election, one of my tasks was to estimate the electoral vote distribution based on survey results and any other information available to me. I made very little use of the State survey results and would have done better had I not used them at all. I discarded all but a few of the State surveys because (1) I was concerned about bias of the auspices (some of the surveys were done in behalf of the political candidates); (2) I was doubtful about the methodology (either the sampling or interviewing was suspect); or (3) the sample size was too small to be useful.

The alternative I had was to use regional data from the national survey and make State estimates based on relationships among the States within a region observed in earlier elections. Except in the South, those relationships had been reasonably consistent through the Presidential elections of the 1950's and 1960's. Some States consistently voted more Democratic than the region as a whole, others were more Republican. The point of these remarks is that a rough and ready "synthetic" procedure provided better estimates than State surveys of questionable quality.

A significant point in Froland's paper is the importance of the strategy used to disseminate statistical results and the need to distinguish between what is statistically acceptable and what is useful in practice. As he points out, estimates do not have to meet rigorous statistical standards in order to be useful. There is an urgent need to continue to suggest ways in which national survey data can be useful to community program administrators.

General Discussion

* Things have been put on a different level from what was talked about this morning. One point suggested by Charles Froland's paper is that the real issue is: How well did the individual characteristics in Reuben Cohen's survey correlate with the alternative data that he had available? Has someone ever done these kinds of correlations for local areas for which survey data were actually collected? That would really have been useful information for the process.

A general point is that if there is an assessment of error, then the data is a lot more useful than if there is no assessment of error.

There is another point being made: The context in Froland's paper is different from the context of Cohen's paper. Froland did not have a policy situation with a great deal of money; this is different from the context where millions or billions of dollars are involved. When that is true, then a much higher standard of accuracy should be called for.

I think it's amazing that the demand for accuracy is probably ten times the demand that we're talking about. It's remarkable to see the statisticians' desire to get error down below levels that don't really make any difference for the purpose.

* There is no indicator of demand here--there is an indicator of a crude level of use. There is an indicator of the proportion that met some arbitrary criterion which cannot clearly be defined as need or as demand in any sense. If you want a better match between the estimated number at risk and the MHIS admissions, I think this is one of the messages: All you have to do is pick a different arbitrary level to define risk and you get a better correspondence. But these will not necessarily be the same people, which is another factor to consider.

* Look at the changes in levels of activity. If you compare it with Table 6, where you're really coming down to a few percentage points difference in what might be regarded as a policymaker's potential clientele, I'm a little overwhelmed by the mixture of levels of accuracy.

* One of the things that was considered was the issue of error. We knew it was there--how great it was was something that we didn't know. That was one of the primary reasons for structuring a process.

* It's appropriate for a lot of situations. People who from day to day have to deal with the problems can look at these as results of one method. There could be discussion of: What do you think about it; does that help you? It's a step up from what normally goes on in planning discussions.

* Sample size has not been mentioned but there certainly are some startling findings. The fact that the female orevalence rate is higher than the male prevalence rate runs against a lot of experience, but it's hard for me to believe, because quite a different dependent variable was used. It certainly seems that the consumers of these things are probably more than happy to see a high prevalence rate; but what about the relative distribution problem?

* What the data represents is a combination of a proportion within the area times the rate. The rate was a sex-specific rate.

* In the alcohol field the issue of definition of what you want to measure has a degree of arbitrariness associated with it. There are material differences you can get simply by setting a standard of a few drinks more or less as to what is a drinking problem. What you want to measure is a much more difficult issue than the question of how to count it once you have defined it. The statistical aspects of this workshop have been very interesting. But there are real problems of definition that are bigger than the problems of statistical differences and errors. Beyond that, it is necessary to be aware that there is a ten to one ratio between prevalence and utilization and some utilization data aren't very accurate and deal with counts of admissions rather than individuals. You really have to wonder about this juxtaposition of differing levels of accuracy and interest.

*The ten to one ratio is actually a small one. A lot of literature found it much higher, depending upon the kind of problem, or the area, or the availability of services. So we weren't at all surprised to find that kind of difference.

* Are you implying that there are ten times as many people that need treatment as are getting treatment? Or are we talking here of prevention modality? It seems to me that we've discussed primarily secondary and tertiary treatment modality. I see most of these data as indicating some population at risk.

Concerning the alcohol field--there are a lot of gradations of use which don't suggest that someone has a full blown alcohol problem. But if they keep it up, most medical evidence would indicate that in a period of time they might have the problem.

* The data as I see it would be more useful really for people designing prevention programs than treatment programs.

* It used to be that if there was some physiological damage then you could be sure that you had an alcoholic on your hands. More and more the judgment in treatment services tends to take a much broader view of who is at risk. A second point about the difference in magnitude: From a practical standpoint, it really doesn't make any difference to the people in the field whether it is ten times or twenty times,

* Did you discover any groups that weren't served at all by any programs? The thing that is apparent is the volume of users that live in the suburbs: Mostly white women, middle class, and there are very few programs for people like that. I think one of the kinds of things that a needs analysis should do is to make a population estimate of groups that no program exists for. Did that occur?

* There has been some mention of data problems. Have some of the indicators used for testing the quality of your process been tried in regression estimates as a composite estimator with the synthetic estimates? I wonder what advice we would give you if you were asking--given the fact that you used something to test the method, but the results are also available for inclusion in the estimate.

* I want to clear up one point. I wasn't being critical of mixtures of levels of accuracy. I was thinking of the different levels of accuracy that were being discussed in different papers. I think this kind of work suggests the value of the observation that we sure know a lot more just by knowing admissions. We are a lot closer to some sort of reality for practical purposes in terms of predicting anticipated admissions for next year by looking at last year or this year. There seems to be a circle that's been traveled. To start, apparently planners and program people don't like their own current statistics and are looking for something that tells them more about prevalence and need. This is a logical step that is mediated through their presumed knowledge of treatment of needs. Then it circles all the way around and comes back to how many people they are seeing in the program anyway. It seems somewhat of a circular process. The only people, in a way, who are benefiting from it are the estimators and the people who get some benefit from having a prevalence that way outstrips their ability to serve that prevalence. It does seem that past admissions are a lot more trustworthy figure.

* As has been noted there are some data that had been used for synthetic estimates and there are some data on indicators. What advice would you give?

* As a general rule, any time you have two estimators for a group of small areas that you think are equally good (both are reasonably good, or both poor) you should consider combining somehow; and, you're not going to do any worse than the poorer of the two.

* Another theme that has been raised is the disparity, or seeming disparity, between what is acceptable when millions or billions of dollars are at stake versus what the person has to do when there is a small staff and little time. Is it really that different a problem?

* I want to comment on the point you made, which I think is a very fundamental one, not in terms of the question as posed but rather on: What are the requirements for precision when you are dealing with a billion dollar program as compared with a tens of thousand dollar program? The real question is, when you're dealing with these billion dollar programs whether synthetic estimates are the right thing to do or whether you should be pressing for the kind of money that would give you better kinds of estimates. When you are dealing with small programs, such as Froland's, just from the point of view of any kind of cost benefit ratio, it doesn't make sense to put more money into getting the statistics than you put into the program. If the estimates are synthetic and are crude, they would still be the best kind of allocation data for the purpose and the nominal cost involved.

Several things suggest themselves as a result of the sessions to this point. They relate to the basic question asked: Under what conditions should we use synthetic estimates? Part of the answer is, use synthetic estimates when it doesn't pay to put a lot of money into trying to get individual survey data for individual places. Thus, there are times when previous studies may indicate that getting data from a survey would result in quality so poor that almost certainly synthetic estimates would give you better information for local areas than from a survey or a census. The fact is that under some conditions, because of measurement error, we are likely to do better with synthetic estimates than with directly collected data. Unfortunately there isn't any nice set of rules that can be put down that would identify the specific circumstances. You have to think about the problem. and if it is likely there would be substantial measurement error', at least in some cases, synthetic estimates would be a useful solution.

Under some circumstances this would hold true even for larger places up to and including the United States as a whole. If the figure on dilapidated units in the 1960 census is compared to what was gotten in a housing survey done simultaneously with the census, but by better trained and better quality interviewers, one figure is found to be fifty percent higher than the other. If an overall result for the United States is subject to problems of quality of this magnitude, imagine what it must be at the local area level where a small number of interviewers are involved.

* There is a question of use which needs to be examined. Are data needed on level for a large area or are data needed on the relative order of differences among small areas such as counties or tracts, as illustrated in Froland's paper? Attention needs to be paid to who are the users and what are their data needs, both on geographic level and quality. For if one ignores the users, after data have been published for local areas, if there is distrust, then users will ignore the data that have been compiled and use either synthetic estimators or direct collection of data that they believe are relevant and have the needed accuracy.

* For some purposes large relative errors for areas with small populations don't make very much difference, whereas small relative errors for large population areas make a lot of difference. There hasn't been very much discussion about how synthetic estimates and direct estimates

can be used jointly: where the direct estimates are used only for the very large States or for very large local units and synthetic estimates for the others. It is net only the size of an area in population that should be considered but also the importance of an area for analysis. For example, if I wanted to construct a conservation target for home heating fuel it is going to make a big difference whether I'm talking about Minnesota or about a State with the same population in the deep South. I want more accurate data--not in terms of relative error--but in terms of absolute error--for the Northern State than for the Southern State in this instance.

* In a sense the thrust of the last comment needs to be kept in mind. That is, there are occasions when there is knowledge of an atypical situation and the method of synthetic estimates does not (as it stands right now) take it into account. In fact, if we were designing a survey we would take it into account by treating it as a separate self-representing area or we would do something special in estimation. It appears right now that for synthetic estimates we use two sets of data and a single algorithm and get the result. We ought to try to keep such possibilities of atypical areas in mind and suggest to the producers and users of the synthetic estimation approach how to deal with these kinds of identifiable problems. We've heard one method that has been proposed: The use of symptomatic indicators. But, how do we provide that in certain circumstances the symptomatic indicators be used for problem areas when the synthetic estimators should not be used; whereas, for the other areas the synthetic estimator should be used? Perhaps we have to get away from what might be called a push button approach and create a joint composite estimator approach that comes close to the complex kinds of sample designs we construct.

* A bit less general way of doing what you have just described was in Wes Schaible's paper.

* I think, in one place in Bob Fay's paper, there is a distinction between two populations: those above and those below the median. This seems to me is the kernel of an idea with respect to use of measures of position for a symptomatic indicator for defining subsectors of the population. For one subsector there would be proper use of a single kind of estimator, say, a synthetic estimator; for other subsectors one would use a more complex system (including a composite estimator with varying weights for each of the specific subsectors).

* I'm not sure problems are quite so complicated. Of course, there are likely to be exceptions. The example of a conservation target for home heating fuel may be amenable to a less complex approach. It seems to me that if I were looking at heating fuel I would use weather information for classifying States into tiers. If it turns out that Minnesota is unusual among its tier of Northern States, then, of course, there would be trouble with the Minnesota estimate.

* Another point which is related is that you need to consider the properties of the variable that you are estimating. There has been reference to the fact that synthetic estimates for diseases seem to be relatively accurate compared with estimates for other variables. That seemed to make sense. But there are other variables where you would

expect the synthetic estimates to be bad. I think unemployment is a very good example of that because of the nature of the economy. If you have a synthetic estimate that is based on industry like, let's say, the steel industry, for example? that doesn't mean that every steel plant lays off ten percent of its workers. There is likely to be variation. What actually happens is that Youngstown Steel closes down in Youngstown, so you have a place with a thirty percent unemployment rate. They just happened to close down before Inland Steel did in Gary so the unemployment rate in Gary would be lower. So there is reason to think that the synthetic estimates for local area unemployment would be bad because this is how unemployment arises.

* Is the suggestion perhaps that it would be useful to build in a current indicator of local area variation if such data exist? Are we coming around full circle to the question: What are the resources, what do we know about the between and within variances, and how current are the indicator data?

* Perhaps it is a bit different. In the absence of other data there are substantial reasons why you would not use synthetic estimates for unemployment, but there are substantial reasons why you would use synthetic estimates of death rates due to certain diseases. All you have to do is look at unemployment rates as far back as they go. If there was high unemployment, it was uneven and it lends credence to the point. So, it is more an argument of when you don't use synthetic estimates in the absence of other information.

* Another variable in this situation is the group that is producing the estimates. Should the Census Bureau be producing synthetic estimates? It is a different thing than if the local area produces the local estimates. You expect the Census Bureau to do a thorough analysis of the methods and to try to understand the errors and develop a model that you feel reasonably sure fits the situation. It is no different from conditions under which you do a survey, or the conditions under which you produce an estimate from the survey, or the conditions under which you will not. If a national statistical agency is putting out the data, it has a different connotation than if a local area is producing that local area estimate for its own use. That is part of the problem that we have here. We may lose sight of an important part of the problem, and it may be that the national statistical agencies will simply refrain from producing certain local area statistics because they feel that the errors are too large or that the errors are not measurable. It isn't the size of the error that bothers you so much. It is whether or not you have an idea of how large that error is. If you feel that you don't have a reasonable fix on the size of the error, you may decide that as a Federal agency, you will not produce the data. That does not prevent the local area from going ahead and using the data if it wants to, for its own purposes. There is an official character to the data that is produced by a Federal agency, and there is an expectation of accuracy, deliberateness, and thoughtfulness. I don't see why that aspect should be any different as it applies to synthetic or any other kinds of estimate since it applies to the statistics produced directly from surveys.

* Federal agencies have a responsibility when they work with local people who want to produce an estimate for some particular characteristic, to determine whether a situation exists where there are anomalies for small areas. It is this sort of thing that synthetic estimation has trouble handling. And it is up to the Federal agency at that point, whether or not estimates of the error involved in the particular statistics can be determined, to advise the local people that based on previous experience synthetic estimates won't work.

* It appears from the work that Bob Fay and Gene Ericksen have told us about, that it really takes a good deal of work to understand what is going on with synthetic and regression estimators. It really requires that we dig into it quite hard to know what is going on. It may be that if people ask (and if in fact that is what it takes and there don't appear to be any shortcuts that anyone has thought of to setting up criteria), you may have to say sometimes: "I really can't tell you. I don't have the experience or the knowledge, and I can't advise you to do this."

* I would like to be a devil's advocate for a minute. It seems that what we are trying to do is to provide only Grade A statistics and if it is not Grade A, then data are not to be provided. Perhaps groups that like to have a Grade A symbol attached to their work need to examine whether some lower grade should be made available with an indication of the level of quality which is associated with the data. If it can't be done in a quantitative sense, it could be attempted in a qualitative sense. In Britain, for example, in certain programs, they do use this system of Grade A, Grade B, and Grade C as a way of distinguishing, in a qualitative sense, among a variety of statistical outputs. It's handled in a way so that users are put on notice that there are problems in the lower grade categories that demand attention.

* I don't think that that was what I was saying. What I was saying was that you don't put out everything just because there might be a need for it. In survey work there is a screening. You consider what you can do and what you can't do. I don't see why the same kind of consciousness of what is the best type shouldn't apply to the statistics that don't come out of surveys as applies in statistics that do come out of surveys. It may be that there will be political reasons why you have to put out some poorer statistics anyway. But, we should make the distinction between the political reason for doing it and us as statisticians proposing to do it.

* I think that we all agree that putting out good data is a good idea. The next question is, is putting out bad data worse than no data? I think Froland said at one point--this is better than giving the money to people who cry the loudest. I'd like to put in a good word for people who cry the loudest. Crying the loudest is often a very helpful thing for the system. It teaches people about argument; you've got to get in there and say what's going on at the level of providing service. This notion of providing a more rationalized system for the distribution of resources is not necessarily as desirable in all respects. I suppose one could go on and speak a whole essay about that. But, just one word for the people who cry loud.

* Have we found it or have we lost it? We started our discussion about synthetic estimates on the note that some would be enthusiastic and some skeptical. In the course of the presentations and the discussion a number of different methods have been discussed. Some have been carefully documented and have led to a feeling that synthetic estimates do provide useful means for creating estimates. On the other hand, there has been discussion leading to the feeling that perhaps we are not yet to the point where these methods can be and should be used in every instance.

* I just want to say regarding that: I think that people feared we might have a how-to-do-it manual coming out, and instead I think we're going to have a very fine consumer's report on synthetic estimates, which will serve the field very well.

(Contributing to the general discussion during this period were: Ira Cisin, Reuben Cohen, Eugene Ericksen, Dwight French, Charles Froland, Maria Gonzalez, Louise Richards, Ron Roizen, Wes Schaible, Walt Simmons, Monroe Sirken, Joseph Steinberg, Joseph Waksberg, and Robert Wilson.)

Expansion of Remarks

Wait R. Simmons

Recapping statements of my own and of others, we should follow these guidelines in order to increase the utility of a synthetic estimate:

Let $$\bar{Z}'_c = \sum_a W_{ca} \bar{X}'_a,$$

where $\bar{X}'_a$ is the direct estimate for the a^{th} category of persons, secured from a probability survey, W_{ca} is the proportion of persons in community c that fall into category a, and $\bar{Z}'_c$ is the synthetic estimate for community c.

The efficiency of the Z - estimate depends upon four factors:

A. The variability of $\bar{X}'_a$ measures among a-classes. Design should make this variability as great as feasible.

B. The variability of the X - measure among persons within an a-class. We seek a-classes for which this variability is relatively small.

C. The sampling variances of the estimates $\bar{X}'_a$, which in turn are a function of B above, and sample size. This means that sample sizes of the a-classes must be adequate to yield tolerable sampling error.

D. The variability of the W_{ca} values among the c-communities of interest for a given a-class. The guidelines require a search for a-classes for which this variability is as great as available data permit.

It seemed to me that the majority view of conferees was that the best choice is a composite estimate that is a weighted average of a direct estimate and either a simple synthetic estimate or some form of regression estimate.

For many purposes, data for a homogeneous class of small areas -- where class is defined in socioeconomic terms; for example, central cities in the North Central U.S. with 200,000 to 1,000,000 population, median household income under $10,000, and more than 20 percent black -- are acceptable in lieu of data for a specific small area and may have greater validity. Average relative measurement error may be quite large for individual small areas, but may be substantially less for the direct estimate for the homogeneous class of areas, and thus lead to superior final conclusions.

Afterword

Joseph Steinberg

The participants in this workshop gave the existing techniques of Synthetic Estimates for Small Areas a mixed review. Synthetic estimates are useful in some situations where small area data are not available. There are other situations where synthetic estimates are not useful and in some cases may be worse than no data at all.

Throughout the course of the workshop, there have been comments and advice concerning criteria about when to use and when not to use synthetic estimates. Walt Simmons in his Expansion of Remarks suggests guidelines for increasing the utility of a synthetic estimate. It was felt that where there were going to be important decisions involving substantial sums of money there should be significant efforts to obtain funding of direct survey estimates with usable precision. For other situations, especially where funds were limited for program needs and where cost benefit analyses dictated it, synthetic estimates may serve in the absence of anything else. In such situations they are likely to be better as a basis for decisions than opinions or pressure (although some may prefer pressure as a decision-making tool).

Surveys or census results may not provide the answer to small area data needs if there are relatively large measurement errors in direct data collection. If the data are needed retrospectively, there will be no opportunity to do surveys and all that is feasible is one or another indirect estimation, if anything is to be provided.

Anomalies need to be recognized. Symptomatic data may be helpful in recognizing such situations. Sometimes the symptomatic data, used in a regression function, may provide one useful component of a composite local area estimator. The other component could be a direct estimate or a synthetic estimate. James-Stein estimators should be considered. Symptomatic data may be helpful in the efficient design of a basic sample survey geared to the needs of synthetic estimation. Multilevel survey design strategies need to be considered. The efficiencies of designs using random digit dialing techniques for one aspect should be explored.

A variety of estimators have been discussed during the workshop. Each use was related to particular circumstances. The nature of the variable being estimated may suggest the desirability (or its lack) of use of a simple synthetic or regression estimator or a composite estimator. Synthetic estimates may not be a good way of ordering areas if they are based on demographic characteristics since such characteristics may not vary much among local areas; care was advised for such intended use.

If there is some means of determining quality of estimate, publication of synthetic estimates could be considered. Availability only of average approximate measures of quality should be considered reasonable for synthetic estimates as are average approximate standard errors when publishing probability sample survey data.

After evaluation of likely quality, it seems clear that professional statistical judgment needs to be exercised before synthetic estimation use is recommended.

There is a need for continuing research on estimators and evaluation methods. It is unlikely that many small area data needs--including some where substantial resource allocation is involved--are going to be met by direct surveys. Continuing efforts to improve small area estimation techniques are needed to serve the many and varied policy and administrative needs of our society for objective planning, allocation, and decision.

Appendix

A: Attendees at Workshop

B: Workshop Program

Appendix A: Attendees at Workshop

Herbert I. Abelson, Ph.D.
Response Analysis Corporation
Box 158
Princeton, New Jersey 08540

Barbara A. Bailar, Ph.D.
Research Center for Measurement Methods
Bureau of the Census
Washington, DC 20233

Ira Cisin, Ph.D.
Social Research Group
The George Washington University
2401 Virginia Avenue, NW
Washington, DC 20037

Judy Coakley
Division of Extramural Research
National Institute on Alcoholism and Alcohol Abuse
Rockville, Maryland 20857

Reuben Cohen
Response Analysis Corporation
Box 158
Princeton, New Jersey 08540

Steven B. Cohen, Ph.D.
National Center for Health Services Research
Hyattsville, Maryland 20782

Eugene P. Ericksen, Ph.D.
Institute for Survey Research
Temple University
1601 North Broad Street
Philadelphia, Pennsylvania 19122

Donna O. Farley
Suburban Cook County - DuPage County Health Systems Agency
1010 Lake Street
Oak Park, Illinois 60301

Robert E. Fay, Ph.D.
Statistical Research Division
Bureau of the Census
Washington, DC 20233

Dwight K. French
National Center for Health Statistics
Hyattsville, Maryland 20782

Charles Froland, Ph.D.
Regional Research Institute
Portland State University
Portland, Oregon 97207

Charles J. Furst, Ph.D.
Neuropsychiatric Institute - Department of Psychiatry
University of California Center for the Health Sciences
Los Angeles, California 90024

Maria Elena Gonzalez
Office of Federal Statistical Policy and Standards
Department of Commerce
Washington, DC 20230

Warren G. Holland
National Clearinghouse for Alcohol Information -
Alcohol Epidemiologic Data System
PO Box 2345
Rockville, Maryland 20852

Gary G. Koch, Ph.D.
Department of Biostatistics
School of Public Health
University of North Carolina
Chapel Hill, North Carolina 27514

Paul S. Levy, Sc.D.
Department of Biostatistics
School of Health Sciences
University of Massachusetts
Amherst, Massachusetts 01003

(affiliation at time of workshop:

School of Public Health
University of Illinois at the Medical Center
Chicago, Illinois 60680)

Lillian H. Madow
Bureau of Labor Statistics
441 G Street, NW
Washington, DC 20212

Harold Nisselson
Statistical Standards and Methodology
Bureau of the Census
Washington, DC 20233

Frederick J. Oeltjen
Division of Community Assistance
National Institute on Drug Abuse
Rockville, Maryland 20857

David M. Promisel, Ph.D.
Policy Studies and Special Reports Branch
National Institute on Alcoholism and Alcohol Abuse
Rockville, Maryland 20857

Louise G. Richards, Ph.D.
Psychosocial Branch, Division of Research
National Institute- on Drug Abuse
Rockville, Maryland 20857

Joan Rittenhouse, Ph.D.
Office of Medical and Professional Affairs
National Institute on Drug Abuse
Rockville, Maryland 20857

Ron Roizen
Social Research Group
School of public Health
University of California
1918 Bonita Avenue
Berkeley, California 94704

Richard M. Royall, Ph.D.
Department of Biostatistics
School of Hygiene and Public Health
The Johns Hopkins University
615 North Wolfe Street
Baltimore, Maryland 21205

Wesley L. Schaible, Ph.D.
Office of Statistical Research
National Center for Health Statistics
Hyattsville, Maryland 20782

Walt R. Simmons
1525 Belle Haven Road
Alexandria, Virginia 22307

Monroe G. Sirken, Ph.D.
Office for Mathematical Statistics
National Center for Health Statistics
Hyattsville, Maryland 20782

Joseph Steinberg
Survey Design, Inc.
1320 Fenwick Lane
Silver Spring, Maryland 20910

Joseph Waksberg
Westat, Inc.
11600 Nebel Street
Rockville, Maryland 20852

Robert Wilson, Ph.D.
College of Urban Affairs and Public Policy
University of Delavare
Newark, Delaware 19711

Philip Wirtz
Social Research Group
The George Washington University
2401 Virginia Avenue, NW
Washington, DC 20037

Thomas H. Woteki, Ph.D.
Office of Data Development
Energy Information Administration
Department of Energy
Washington, DC 20461

Appendix B: Workshop Program

Thursday, April 13, 1978

9:00 CONVENING OF WORKSHOP	Louise G. Richards National Institute on Drug Abuse
	Monroe G. Sirken National Center for Health Statistics
Remarks by Chair	Joseph Steinberg Survey Design, Inc.
9:15 PAPER Small Area Estimation -- Synthetic and Other Procedures, 1968-1978	Paul S. Levy School of Public Health University of Illinois at the Medical Center
Discussants	Walt R. Simmons Consultant, National Academy of Sciences
	Gary G. Koch School of Public Health The University of North Carolina at Chapel Hill
10:45 PAPER Drug Abuse Applications: Some Regression Explorations with National Survey Data	Reuben Cohen Response Analysis Corporation
Discussants	Monroe G. Sirken
	Ira Cisin Social Research Group The George Washington University
2:00 PAPER A Composite Estimator for Small Area Statistics	Wesley L. Schaible National Center for Health Statistics
Discussant	Barbara A. Bailar Bureau of the Census

3:00 PAPER Prediction Models in Small Area Estimation
Richard M. Royall
School of Hygiene and Public Health
The Johns Hopkins University

Discussant — Harold Nisselson
Bureau of the Census

4:00 PAPER A Modified Approach to Small Area Estimation
Steven B. Cohen

Discussant — Joseph Waksberg
Westat, Inc.

Friday, April 14, 1978

9:00 PAPER Applications of Synthetic Estimates to Alcoholism and Problem Drinking
David M. Promisel
National Institute on Alcohol Abuse and Alcoholism

Discussant — Donna O. Farley
Suburban Cook County-DuPage County Health Systems Agency

10:15 PAPERS Case Studies on the Use and Accuracy of Synthetic Estimates: Unemployment and Housing Applications
Maria E. Gonzalez
Office of Federal Statistical Policy and Standards

Some Recent Census Bureau Applications of Regression Techniques to Estimation
Robert E. Fay
Bureau of the Census

Discussant — Eugene P. Ericksen
Institute of Survey Research
Temple University

2:00 PAPER Synthetic Estimates as an Approach to Needs Assessment: Issues and Experience
Charles Froland
Berkeley Planning Associates

Discussant — Reuben Cohen

3:00 Summary Remarks — Joseph Steinberg

National Institute on Drug Abuse

Research monograph series

While limited supplies last, single copies of the monographs may he obtained free of charge from the National Clearinghouse for Drug Abuse Information (NCDAI). Please contact NCDAI also for information about availability of coming issues and other publications of the National Institute on Drug Abuse relevant to drug abuse research.

Additional copies may be purchased from the U.S. Government Printing Office (GPO) and/or the National Technical Information Service (NTIS) as indicated. NTIS prices are for paper copy. Microfiche copies, at $3, are also available from NTIS. Prices from either source are subject to change.

Addresses are:

NCDAI
National Clearinghouse for Drug Abuse Information
Boom 10-A-56
5600 Fishers Lane
Rockville, Maryland 20857

GPO
Superintendent of Documents
U.S. Government Printing Office
Washington, D.C. 20402

NTIS
National Technical Information Service
U.S. Department of Commerce
Springfield, Virginia 22161

1 FINDINGS OF DRUG ABUSE RESEARCH. *An annotated bibliography of NIMH- and NIDA-supported extramural grant research, 1964-74. Volume 1, 384 pp., Volume 2, 377 pp.*

Vol.1: GPO out of stock	NTIS PB #272 867/AS $14
Vo1.2: GPO Stock #017-024-0466-9 $5.05	NTIS PB #272 868/AS $15

2 OPERATIONAL DEFINITIONS IN SOCIO-BEHAVIORAL DRUG USE RESEARCH 1975. Jack Elinson, Ph.D., and David Nurco, Ph.D., editors. *Task Force articles proposing consensual definitions of concepts and terms used in psychosocial research to achieve operational comparability. 167* pp.
GPO out of stock NTIS PB #246 338/AS $8

3 AMINERGIC HYPOTHESES OF BEHAVIOR: REALITY OR CLICHE? Bruce J. Bernard, Ph.D., editor. *Articles examining the relation of the brain monoamines to a range of animal and human behaviors.*
GPO Stock #017-024-00486-3 $2.25 NTIS PB #246 687/AS $8

4 NARCOTIC ANTAGONISTS: THE SEARCH FOR LONG-ACTING PREPARATIONS. Robert Willette, Ph.D., editor. *Articles reporting current alternative inserted sustained-release or long-acting drug devices.*
GPO Stock #017-024-00488-0 $1.10 NTIS PB #247 096/AS $4.50

5 YOUNG MEN AND DRUGS: A NATIONWIDE SURVEY. John A. O'Donnell, Ph.D., et al. *Report of a national survey* of drug *use by men 20-30 years old* in *1974-5. 144 pp.*
GPO Stock #017-024-00511-8 $2.25 NTIS PB #247 446/AS $8

6 EFFECTS OF LABELING THE "DRUG ABUSER": AN INQUIRY. JAY R. Williams, Ph.D. *Analysis and revew of the literature examining effects of drug use apprehension or arrest on the adolescent.*
GPO Stock #017-024-00512-6 $1.05 NTIS PB #249 092/AS $4.50

7 CANNABINOID ASSAYS IN HUMANS. Robert Willette, Ph.D., editor. *Articles describing current developments in methods for measuring cannabinoid levels in the human body by immunoassay, liquid and dual column chromatography and mass spectroscopy techniques. 120 pp.*
GPO Stock #017-024-00510-0 $1.95 NTIS PB #251 905/AS $7.25

8 R :3x/WEEK LAAM - ALTERNATIVE TO METHADONE. Jack Blaine, M.D., and Pierre Renault, M.D., editors. *Comprehensive summary of development of LAAM (Levo-alpha-acetyl methadol), a new drug for treatment of narcotic addiction. 127* pp.
Not available from GPO NTIS PB #253 763/AS $7.25

9 NARCOTIC ANTAGONISTS: NALTREXONE. Demetrios Julius, M.D., and Pierre Renault, M.D., editors. *Progress report of development, preclinical and clinical studies of naltrexone, a new drug for treatment of narcotic addiction. 182 pp.*
GPO Stock #017-024-00521-5 $2.55 NTIS PB #255 833/AS $9

10 EPIDMIOLOGY OF DRUG ABUSE: CURRENT ISSUES. Louise G. Richards, Ph.D., and Louise B. Blevens, editors. *Conference Proceedings. Examination of methodological problems in surveys and data collection. 259*
GPO Stock #017-024-00571-1 $2.60 NTIS PB #266 691/AS $10.75

11 DRUGS AND DRIVING. Robert Willette, Ph.D., editor. *State-of-the-art revew of current research on the effects of different drugs on perfoformance impairment, particularly on driving. 137* pp.
GPO Stock #017-024-00576-2 $1.70 NTIS PB # 269 602/AS $8

12 PSYCHODYNAMICS OF DRUG DEPENDENCE. Jack D. Blaine, M.D., and Demetrios A. Julius, M.D., editors. *A pioneering collection of papers to discover the part played by individual psychodynamics in drug dependence. 187 pp.*
GPO Stock #017-024-00642-4 $2.75 NTIS PB #276 084/AS $9

13 COCAINE: 1977. Robert C. Petersen, Ph.D., and Richard C. Stillman, M.D., editors. *A series of reports developing a picture of the extent and limits of current knowledge of the drug, its use and misuse. 223 pp.*
GPO Stock #017-024-00592-4 $3 NTIS PB #269 175/AS $9.25

14 MARIHUANA RESEARCH FINDINGS: 1976. Robert C. Petersen, Ph.D., editor. *Technical papers on epidemiology, chemistry and metabolism, toxicological and pharmacological effects, learned and unlearned behavior, genetic and immune system effects, and therapeutic aspects of marihuana use.*
GPO Stock #017-024-00622-0 $3 NTIS PB #271 279/AS $10.75

15 REVIEW OF INHALANTS: EUPHORIA TO DYSFUNCTION. Charles Wm. Sharp, Ph.D., and Mary Lee Brehm, Ph.D., editors. *A broad review of inhalant abuse, including sociocultural, behavioral, clinical, pharmacological, and toxicological aspects. Extensive bibliography. 347 pp.*
GPO Stock #017-024-00650-5 $4.25 NTIS PB #275 798/AS $12.50

16 THE EPIDEMIOLOGY OF HEROIN AND OTHER NARCOTICS. Joan Dunne Rittenhouse, Ph.D., editor. *Task Force report on measurement* of *heroin-narcotic use, gaps in knowledge and how to address them, improved research technologies, and research implications. 249 pp.*
GPO Stock #017-024-00690-4 $3.50 NTIS PB #276 357/AS $9.50

17 RESEARCH ON SMOKING BEHAVIOR. Murray E. Jarvik, M.D., Ph.D., et al., editors. *State-of-the-art* of *research* on *smoking behavior, including epidemiology, etiology, socioeconomic and physical consequences* of *use, and approaches to behavioral change. From a NIDA-supported UCLA* conference. *383 pp.*
GPO Stock #017-024-00694-7 $4.50 NTIS PB #276 353/AS $13

18 BEHAVIORAL TOLERANCE: RESEARCH AND TREATMENT IMPLICATIONS. Norman A. Krasnegor, Ph.D., editor. *Conference papers discuss theoretical and empirical studies of nonpharmacologic factors in development of tolerance to a variety of drugs in animal and human subjects. 151 pp.*
GPO Stock #017-024-00699-8 $2.75 NTIS PB #276 337/AS $8

19 THE INTERNATIONAL CHALLENGE OF DRUG ABUSE. Robert C. Petersen, Ph.D., editor. *A monograph based on papers presented at the World Psychiatric Association 1977 meeting in Honolulu. Emphasis is on emerging patterns of drug use, international aspects of research, and therapeutic issues of particular interest worldwide.*
GPO Stock #017-024-00822-2 $4.50 NTIS PB # to be assigned

20 SELF-ADMINISTRATION OF ABUSED SUBSTANCES: METHODS FOR STUDY. Norman A. Krasnegor, Ph.D., editor. *Papers from a technical review on methods used to study self-administration of abused substances. Discussions include Overview, methodological analysis, and future planning* of *research on a variety of substanses: drugs, ethanol, food, and tobacco. 246 pp.*
Not available from GPO NTIS PB #288 471/AS $10.75

21 PHENCYCLIDINE (PCP) ABUSE: AN APPRAISAL. Robert C. Petersen, Ph.D., and Richard C. Stillman, M.D., editors. *Monograph derived from a technical review to assess the present state* of *knowledge about phencyclidine and to focus on additional areas* of *research. Papers are aimed at a professional and scientific readership concerned about how to cope with the problem of PCP abuse. 313 pp.*
GPO Stock #017-024-00785-4 $4.25 NTIS PB #288 472/AS $11.75

22 QUASAR: QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIPS OF ANALGESICS, NARCOTIC ANTAGONISTS, AND HALLUCINOGENS. Gene Barnett, Ph.D., Milan Trisc, Ph.D., and Robert E. Willette, Ph.D., editors. *Reports an interdisciplinary conference on the molecular nature of drug-receptor interactions. A broad range of quantitative techniques were applied to questions of molecular structure, correlation of mo1ecular properties with biological activity, and molecular interactions with the receptor(s). 487 pp.*
GPO Stock #017-024-00786-2 $5.25 NTIS PB #292 265/AS $15

23 CIGARETTE SMOKING AS A DEPENDENCE PROCESS. Norman A. Krasnegor, Ph.D., editor. *A review of the biological, behavioral, and psychosocial factors involved in the onset, maintenance, and cessation ot tobacco/nicotine use is presented, together with an agenda for further research on the cigarette smoking habit process.*

In Press

☆ U.S. GOVERNMENT PRINTING OFFICE : 1979 O—293-965