CHOOSING WEIGHTS FOR COMPOSITE ESTIMATORS FOR SMALL AREA STATISTICS

Wesley L. Schaible, National Center for Health Statistics

I. Introduction

Data from sample surveys are sometimes used to produce estimates for geographically defined domains or "small areas" whose boundaries were not used as strata in the original sample design. In such cases the sample in a particular small area may be unrepresentative, small, or even non-existent. Several estimators have been suggested for use in these situations. In the publication Synthetic State Estimates of Disability (1968) the authors state that the sample size (and design) of the Health Interview Survey was inadequate to make State estimates by conventional procedures and after considering several estimators suggest the use of a synthetic estimator. Since this publication, other estimators, including modifications of the synthetic estimator, have been investigated by Levy (1971), Royall (1978), and Gonzalez and Hoza (1978). Much of the research in this area has been devoted to evaluations of 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 Current Population Survey 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.

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 Metropolitan 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 better than choosing one over the other.

Estimators that are weighted sums of two component estimators are not new and have been used to produce estimates by government agencies such as the Bureau of Labor Statistics, Bureau of the Census, and National Center for Health Statistics. In addition, the concept has been discussed in a variety of situations. A composite estimator consisting of a synthetic estimator and an adaptation of a regression estimator was considered by the National Center for Health Statistics (NCHS, 1968). 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." Marks (1977) investigated the use of a composite technique in the construction of the Consumer Price Index of the Bureau of Labor Statistics. Schaible, Brock, and Schnack (1977b) and Brock and Peyton (1978) specified a particular composite estimator and compared its performance with that of direct and synthetic component estimators. Although the James-Stein estimator, James and Stein (1961), and generalizations by Efron and Morris (1973, 1975) were not developed as the sum of two estimators, the weighting schemes given by their approach can generally be viewed as weighting schemes for composite estimators. Fay (1978) considered the use of a James-Stein estimator for the production of estimates of per capita income.

In this paper the minimum mean square error weighting scheme and an approximation are discussed as well as conditions under which the composite estimator gives large reductions in mean square error. In addition, the effect of errors in estimates of the minimum mean square error weight is investigated.

II. Composite Estimators

To define the composite estimator more precisely let Y'_d and 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

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

The mean square error (MSE) of this estimator is

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

By minimizing this quantity with respect to C_d, it is easily shown that the minimum mean square error weight is

$$C_d^* = \frac{\text{MSE } Y''_d - E(Y'_d-\bar{Y}_d)(Y''_d-\bar{Y}_d)}{\text{MSE } Y'_d + \text{MSE } Y''_d - 2E(Y'_d-\bar{Y}_d)(Y''_d-\bar{Y}_d)} \tag{2}$$

If the component estimators are independent and if either is unbiased with estimable variance then C_d^* can be estimated in a straightforward manner. However, the independence assumption may not be valid and an unbiased estimator may not be available. An alternative, and perhaps less restrictive, condition under which C_d^* becomes more manageable is when $E(Y'_d-\bar{Y}_d)(Y''_d-\bar{Y}_d)$ is small rela-

tive to MSE Y''_d. In this case C^*_d may be approximated by

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

where R_d = MSE Y'_d/MSE Y''_d. The weighting 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 range of the exact weight C^*_d is not restricted. A desirable feature of the weight C^{**}_d is that individual estimates of the component mean square errors are not needed to estimate this weight, only an estimate of their relative size is required. Schaible (1978) found that the use of the approximate weight rather than the exact weight produced negligible increases in average squared errors for selected models and variables.

If $E(Y'_d-\bar{Y}_d)(Y''_d-\bar{Y}_d)$/MSE Y''_d is small then the mean square error of the composite estimator can be written in multiples of the mean square error of the second component estimator as

$$\text{MSE } \hat{\bar{Y}}_d/\text{MSE } Y''_d = (R_d+1)C_d^2 - 2C_d + 1 \quad . \qquad (4)$$

Figure 1. Mean Square Error of the Composite Estimator Relative to that of the Second Component Estimator as a Function of the Composite Weight, C_d: (MSE Y'_d = MSE Y''_d)

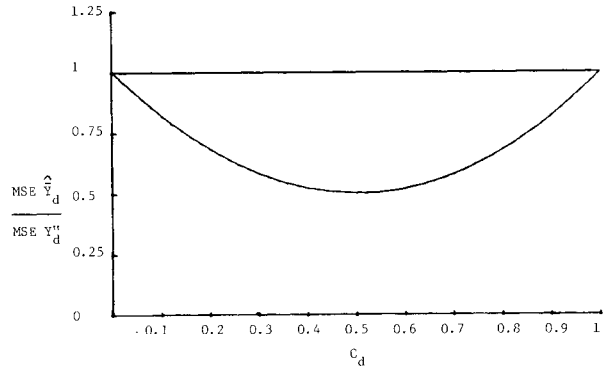


Figure 2. Mean Square Error of the Composite Estimator Relative to that of the Second Component Estimator as a Function of the Composite Weight, C_d: (MSE Y'_d = 2MSE Y''_d)

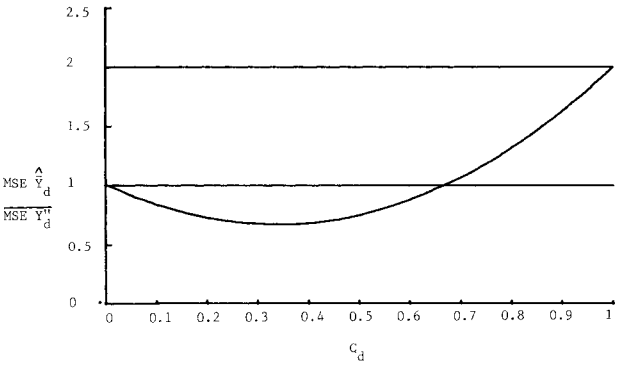
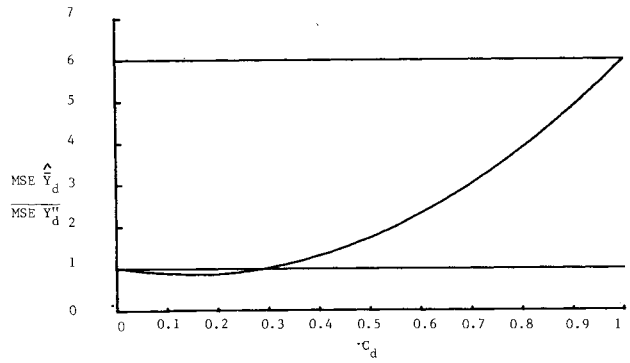


Figure 3. Mean Square Error of the Composite Estimator Relative to that of the Second Component Estimator as a Function of the Composite Weight, C_d: (MSE Y'_d = 6MSE Y''_d)

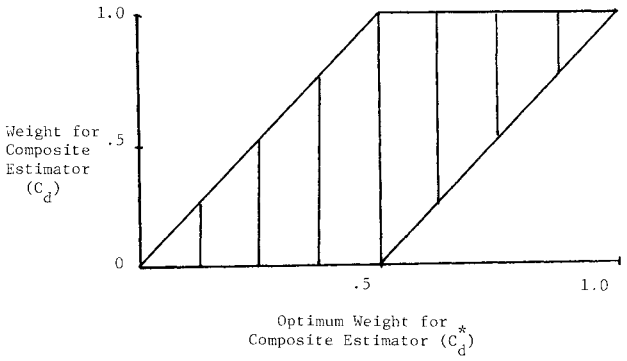


Figures 1, 2 and 3 illustrate this relationship for values of R_d equal to 1, 2 and 6 respectively. The mean square errors of the component estimators are indicated by horizontal lines. In practice, the optimum weight is not known and estimates of the quantity are, of course, subject to error. It is clear from these figures that sizable errors can be made in estimating the optimum weight without producing large increases in the mean square error of the composite estimator. This is consistent with results reported by Royall (1978) which show that in the case of unbiased component estimators the variance curve of the composite estimator is flat in the vicinity of the optimum weight. Another characteristic of composite estimators is evident from these figures. That is, 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. The figures also illustrate the fact that reduction in mean square error and the range of weights for which the composite estimator has smaller mean square error than either component estimator both vary with the magnitude of R_d. Both situations are most advantageous when R_d is close to one.

Royall (1978) 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 4. 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 little information to add, and it is likely that if the relative sizes

of the component estimator mean square errors are known, the estimator with small mean square error would be used rather than a composite 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 figures 1 and 4, the composite will outperform either component estimator regardless of the weight chosen.

Figure 4. The Range of Weights (C_d) for which the Composite Estimator has Smaller MSE than either Component Estimator

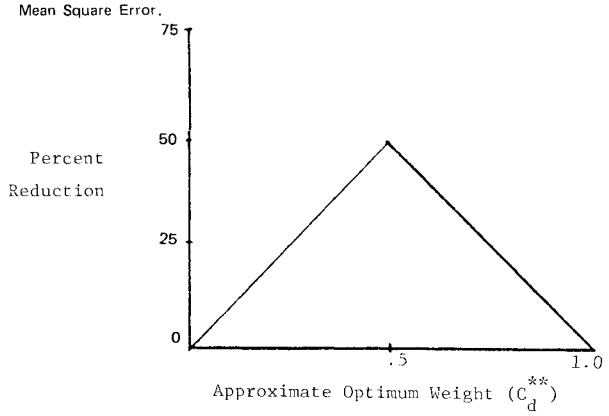


If the expected cross-product term in equation (2) is small relative to the mean square error of the second component estimator then the percent reduction in the mean square error of the composite estimator as compared to the smaller of the mean square errors of the two component estimators is

$$\text{Percent reduction} \approx \begin{cases} 100\ C_d^{**}, & 0 < C_d^{**} \leq .5 \\ 100\ (1-C_d^{**}), & .5 \leq C_d^{**} < 1 \end{cases}$$

This relationship is illustrated in figure 5. 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 percent reduction is 50 percent when the optimum weight is one half and approaches 100 percent as the optimum weight approaches zero or one. If the mean square error of the composite estimator is compared to the average of the mean square errors of the two component estimators then the percent reduction is 50 percent regardless of the value of the optimum weight.

Figure 5. 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.



III. Empirical Results

To further investigate the choice of weights for the composite estimator and to compare this estimator 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 (HIS). For each of five variables within each State estimates were made for a range of weights for the two weighting schemes described below. Estimates were compared with corresponding census values and squared errors averaged over States were calculated. The following five variables obtained in a similar manner in the 1970 census and in the HIS were selected: percent of the population married, separated, less than one year of age, having completed high school, and having completed college. Census values were obtained from the Bureau of Census Public Use Sample Tapes and treated as population values ($\bar{Y}_d$). The sample mean or simple direct estimator (Y_d') and the synthetic estimator (Y_d'') were chosen as the two component estimators. These estimators are defined as follows.

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

$$Y_d' = \sum_{\alpha=1}^{K} \sum_{i=1}^{n_{d\alpha}} Y_{d\alpha i}/n_d \ .$$

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 full sample is

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

and the synthetic estimator for small area d is then

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

This estimator, with the addition of sampling weights and a ratio-adjustment, was used to produce the synthetic estimates for this paper. The ratio-adjustment forces the weighted sum of the individual State synthetic estimates in a geographic region to be consistent with the usual HIS probability estimate for that region. The α-cells 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 members or more
5. Industry of head of family: Standard Industrial Classifications: (1) Forestry and fisheries, agriculture, construction,

mining and manufacturing; (2) all other industries.

Weighting schemes for the composite estimator were defined under two models or sets of assumptions about the behavior of the mean square errors of the component estimators. The first model allows the mean square errors of the two component estimators to vary between but not within small areas, i.e. MSE Y'_d = MSE Y''_d. Under this model the approximate minimum mean square error weight is $C^{**}_d = (1+R)^{-1}$, where R is constant for all small areas. The second model assumes that the mean square error of the simple direct estimator varies inversely with the small area sample size whereas that of the synthetic estimator remains constant, i.e. MSE $Y'_d = b'/n_d$, MSE $Y''_d = b''$. Under this model the approximate minimum mean square error weight is $C^{**}_d = (1+R'/n_d)^{-1}$ where $R' = b'/b''$. It is of interest to note that under this model R' is the small area sample size, n_d, at which the mean square errors of the two component estimators are equal.

For the variables investigated, figures 6 through 10 show the average squared error of the composite estimator over the range of R and R' from infinity to zero. Each R and R' specifies a value of the approximate optimum weight which ranges from zero to one. The average squared errors of the two component estimators are indicated by horizontal lines. As 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. In these figures the range of weights for which the composite estimator has smaller average squared error than both component estimators is dependent on the relative magnitude of the average squared errors of the two component estimators. The ranges found in these figures are consistent with the ranges of mean square error indicated in figure 4. Also, as would be expected from figure 5 the maximum percent reduction in average squared error is greatest in those variables where the average squared errors of the component estimators are of a similar magnitude. This is the case in four of the five variables investigated. The exception is the variable "percent of the population less than one year of age" where the average squared error of the direct estimator is eight times that of the synthetic.

The composite estimator performs well under both weighting schemes. However, the $(1+R'/n_d)^{-1}$ weighting scheme seems to perform slightly better when all five variables are considered. The composite estimator specified by the $(1+R)^{-1}$ weighting scheme when R=1 gives near minimum average squared errors in three of the five variables considered and performs reasonably well for the remaining two variables. This composite estimator is the simple average of the component estimators.

In all five figures the average squared error curve is flat in the vicinity of the optimum weight. This result and the general shape of the curves in these figures are consistent with equation (4). The usefulness of composite estimators is greatly enhanced by this general insensitivity to poor estimates of the optimum weight.

Figure 6. Average Squared Errors of Two Composite Estimators, Percent of the Population Married, Health Interview Survey, Forty-Nine States, 1969-71

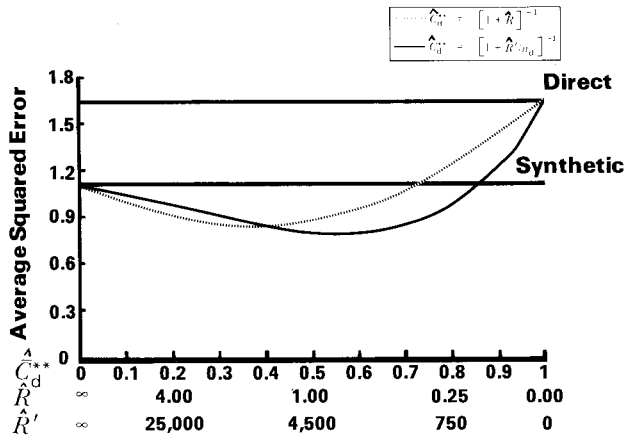


Figure 7. Average Squared Errors of Two Composite Estimators, Percent of the Population Separated, Health Interview Survey, Forty-Nine States, 1969-71

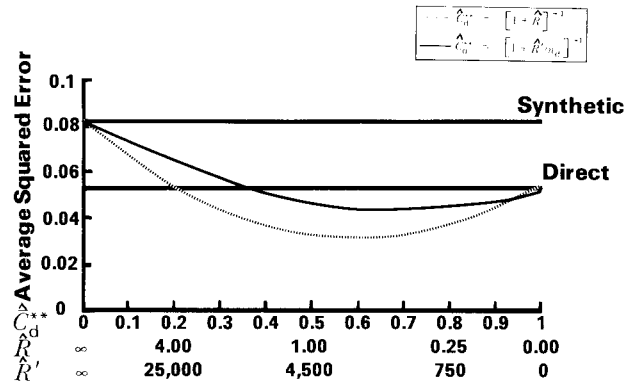


Figure 8. Average Squared Errors of Two Composite Estimators, Percent of the Population Less Than One Year of Age, Health Interview Survey, Forty-Nine States, 1969-71.

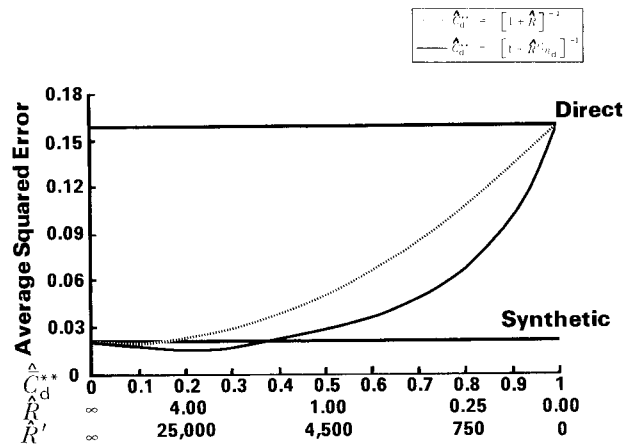


Figure 9. Average Squared Errors of Two Composite Estimators, Percent of the Population Having Completed High School, Health Interview Survey, Forty-Nine States, 1969-71

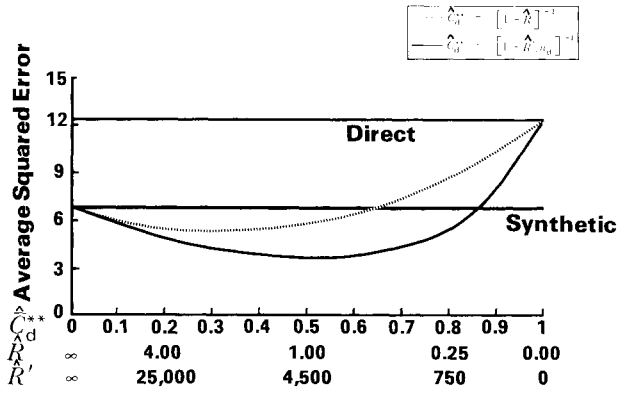
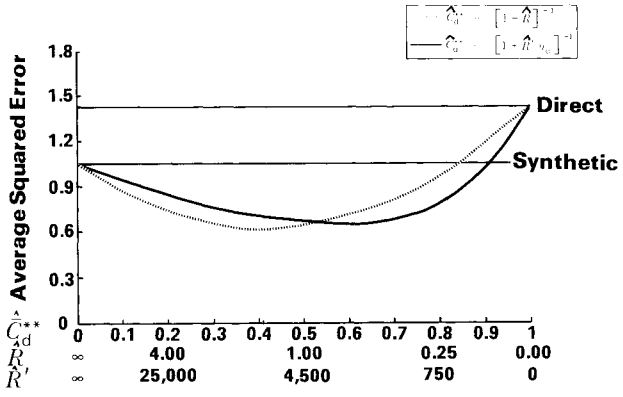


Figure 10. Average Squared Errors of Two Composite Estimators, Percent of the Population Having Completed College, Health Interview Survey, Forty-Nine States, 1969-71



IV. Summary

The composite estimator, a weighted sum of two component estimators of the same parameter, has a mean square error that is smaller than the larger of the mean square errors of the two component estimators. The mean square error of this estimator will be smaller than that of either component estimator when any one of an appropriate range of weights is used. When the mean square errors of the component estimators are equal, this range is from zero to one and the use of a composite estimator can achieve a reduction in mean square error of 50 percent. The mean square error curve of the composite estimator is relatively flat in the vicinity of the optimum weight and the composite estimator is surprisingly insensitive to poor estimates of this weight. Empirical results using average squared errors computed from the Health Interview Survey and Bureau of the Census data are consistent with these statements.

Although composite estimators are being used to produce estimates, there are two areas of research which deserve additional attention. The first is to further investigate weighting schemes and the estimation of the optimum weight for the composite estimator. Several approaches are being considered and, in fact, have been used, but further study is needed. A 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 estimators and is, in general, a difficult one. One way to provide information on the performance of biased small area estimators is to estimate measures of error averaged over small areas. Although this information can be useful, it would be more desirable to have a measure of the estimator's performance for a particular small area.

References

Brock, D.B. and Peyton, B.W.(1978) "Small Area Estimation: An Application of Three Methods to the U.S. National Health Interview Survey." Presented at the 36th Annual Meeting of the United States-Mexico Border Health Association. Reynosa, Tamaulipas, Mexico.

Efron, B. and Morris, C. (1973) "Stein's Estimation Rule and Its Competitors - An Empirical Bayes Approach." Journal of the American Statistical Association, 68(341):117-30.

Efron, B. and Morris, C. (1975) "Data Analysis Using Stein's Estimator and Its Generalizations." Journal of the American Statistical Association, 70(350):311-19.

Ericksen, E.P. (1973) "Recent Developments in Estimation for Local Areas." Proceedings of the American Statistical Association, Social Statistics Section, pp. 37-41.

Fay, R.E. III (1978) "Some Recent Census Bureau Applications of Regression Techniques to Estimation." Presented at the NIDA-NCHS Workshop on Synthetic Estimates, Princeton, N.J.

Gonzalez, M.E. (1973) "Use and Evaluation of Synthetic Estimates." Proceedings of the American Statistical Association, Social Statistics Section, pp. 33-36.

Gonzalez, M.E. and Waksberg, J.E. (1973) "Estimation of the Error of Synthetic Estimates." Presented at the International Association of Survey Statisticians, Vienna, Austria.

Gonzalez, M.E. and Hoza, C. (1975) "Small Area Estimation of Unemployment." Proceedings of the American Statistical Association, Social Statistics Section, pp. 437-43.

Gonzalez, M.E. and Hoza, C. (1978) "Small Area Estimation with Application to Unemployment and Housing Estimates." Journal of the American Statistical Association, 73(361):7-15.

James, W. and Stein, C. (1961) "Estimation with Quadratic Loss." Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1, Berkeley: University of California Press, pp. 361-79.

Levy, P.S. (1971) "The Use of Mortality Data in Evaluating Synthetic Estimates." Proceedings of the American Statistical Association, Social Statistics Section, pp. 328-31.

Levy, P.S. and French, D.K. (1977) "Synthetic Estimation of State Health Characteristics Based on the Health Interview Survey." Vital and Health Statistics, Series 2-75 (No. 78-1349), Public Health Service, National Center for Health Statistics.

Marks, H., (1977) "Estimation of Cost Weights for the Consumer Price Index." Presented at the April 19 meeting of the Washington Statistical Society, Washington, D.C.

Namekata, 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, 90:532-38.

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

Royall, R.M. (1973) "Discussion of two papers on Recent Developments in Estimation of Local Areas." Proceedings of the American Statistical Association, Social Statistics Section, pp. 43-44.

Royall, R.M. (1978) "Prediction Models in Small Area Estimation." Presented at the NIDA-NCHS Workshop on Synthetic Estimates, Princeton,N.J.

Schaible, W.L., Brock, D.B., and Schnack, G.A. (1977a) "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.

Schaible, W.L., Brock, D.B., and Schnack, G.A. (1977b) "An Empirical Comparison of the Simple Inflation, Synthetic and Composite Estimators for Small Area Statistics." Proceedings of the American Statistical Association, Social Statistics Section, pp. 1017-1021.

Schaible, W.L., (1978) "A Composite Estimator for Small Area Statistics," Presented at the NIDA-NCHS Workshop on Synthetic Estimates, Princeton N.J.

Acknowledgements

The author would like to thank Barry Peyton and Roy Heatwole of the Office of Statistical Research for computing estimates and providing graphical presentations for this paper. Also, the author thanks Bernadette Wendricks for typing the manuscript.