USE AND EVALUATION OF SYNTHETIC ESTIMATES

Maria Elena Gonzalez, U.S. Bureau of the Census

Definition of synthetic estimates

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. For the smaller areas, the estimates are no longer unbiased. However, it is possible to measure an average mean square error (MSE) for this set of estimates.

The simplest synthetic estimates are obtained by assuming that for the statistic of interest the mean value in the large area applies to each subarea directly; more refined estimates can be obtained by making this assumption for subgroups of the population. In the case when subgroups of the total are used, they should be nonoverlapping and exhaustive; the statistical estimates for the subgroups of the larger area are combined using independently known weights for the smaller area (e.g., as found at the time of the census) to obtain synthetic estimates for the smaller areas.

One is interested in estimating a characteristic, X. Identify j subgroups in the population, which are nonoverlapping and exhaustive. From the larger area we obtain estimates, $x_{.j}$, for $j=1, 2, \ldots, G$.

A synthetic estimate is desired for subarea i, which is within the larger area. From the latest census we have weights p_{ij}, such that

$$\sum_{j=1}^{G} p_{ij} = 1.$$

The synthetic estimate, x_i^*, for characteristic X and subarea i is defined as

$$x_i^* = \sum_{j=1}^{G} p_{ij} x_{.j} \qquad 1)$$

This estimate associates the characteristic $x_{.j}$ of the larger area with each of the subareas i.

Use of synthetic estimates

Synthetic estimates are used primarily to develop small-area estimates when sample sizes are too small to give reliable results directly. Some examples of recent uses follow.

1) The National Center for Health Statistics has developed synthetic State estimates of disability based on the Health Interview Survey data. National rates of disability for 78 subgroups defined in terms of age, sex, size of household, income, industry, etc., were obtained from the data collected in the Health Interview Survey. These disability rates were weighted by the corresponding population in individual States, from the 1960 Census of Population, to derive synthetic State estimates of disability. [1]

2) The Bureau of the Census has used synthetic estimates for the imputation of population for units reported as vacant in the 1970 Census of Population and Housing, but which were actually occupied. A subsample of the housing units reported as vacant in the 1970 Census of Population and Housing was selected and interviewers were sent to these units to determine how accurately that determination had been made. About 11 percent of the housing units reported as vacant were determined to have been occupied at the time of the census. Separate estimates of such error rates were prepared for twelve geographic areas within the United States. Within each area the rate was applied to each enumeration district in the census and the applicable percentage of vacant units was converted to occupied units. The estimates of the error rates for areas such as cities, counties or States were synthetic estimates. [2]

3) In order to study the properties of synthetic estimates, an experiment was conducted to develop unbiased and synthetic estimates of unemployment for SMSA's for monthly, quarterly and annual estimates based on the Current Population Survey (CPS) data. A comparison of the reliability of the two types of estimates revealed that for monthly data the synthetic estimates were preferable, while for annual data the unbiased estimates were preferable; for the quarterly data, the two were of about equal reliability. [2]

4) In the 1960 Census of Housing enumerators were instructed to rate the physical condition of each housing unit into one of three categories: "sound," "deteriorating," or "dilapidated." One important purpose of this was to provide data on substandard housing defined by Federal and local housing agencies as comprising units lacking complete plumbing facilities plus units which were dilapidated but had all plumbing facilities.

In the 1970 Census of Housing information was again obtained about plumbing, but synthetic methods were used to develop estimates of housing units which were dilapidated with all plumbing facilities (DWAPF). To obtain these estimates census data on housing units with all plumbing facilities 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). 1/ From CINCH, estimates of DWAPF housing units were obtained for specified subgroups for 15 selected large SMSA's and for four balance of regions of the U.S. Synthetic estimates for the smaller areas within these nineteen geographic areas were derived using the corresponding set of DWAPF proportions.

Evaluation of synthetic estimates

Synthetic estimates are biased; to evaluate their reliability one can use the MSE, which can be expressed as the sum of the variance and the

square of the bias:

$$MSE(x_i^*) = \sum_{j=1}^{G} p_{ij}^2 \sigma_{x.j}^2 + (X_i^* - X_i)^2 \qquad 2)$$

where

$\sigma_{x.j}^2$ is the sampling variance of estimate $x_{.j}$,

X_i is the "true value" of the statistic for subarea i, and

X_i^* is the expected value of the synthetic estimate for subarea i.

The estimate given in formula 2 assumes that:

a. the p_{ij}'s are fixed and measured without error; and

b. the cov $(x_{.j}, x_{.k}) = 0$, for $j \neq k$.

In general, the values of X_i are not known and consequently the MSE of an individual synthetic estimate cannot be calculated for a particular area "i." However, if we establish M subareas within the survey population, the average MSE of the synthetic estimate over the M subareas (which may be unequal in size) can be estimated from the sample. Let

$$E\left[\frac{1}{M} \sum_{i=1}^{M} (x_i^* - X_i)^2\right] = \alpha \qquad 3)$$

The average MSE can be estimated by using the following approximation:

$$\hat{\alpha} \doteq \frac{1}{M} \sum_{i=1}^{M} (x_i^* - \sum_{j=1}^{G} p_{ij} x_{ij})^2 - \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{G} p_{ij}^2 (1-2f_{ij}) \sigma_{x_{ij}}^2 \quad [3] \qquad 4)$$

where

x_{ij} is the estimated statistic from the sample for subclass j and area i,

f_{ij} is the sample estimate of the proportion of the total for the j-th subclass that is in the i-th subarea, that is

$f_{ij} = n_{ij} / \sum_{i=1}^{M} n_{ij}$, and

$\sigma_{x_{ij}}^2$ is the sampling variance of estimate x_{ij}.

The interpretation of the square root of the mean square error is not altogether clear. Probability statements which can be made using the standard error for an unbiased estimate do not necessarily hold when the estimates are biased and the root mean square error is used as a measure of reliability. To try to understand the situation, an empirical study of the root mean square errors was made, using 1960 housing data. Synthetic estimates for a census were compared with actual measurements of the item for the same census to obtain an estimate of the bias. For various groupings of areas we then computed an average root mean square error (RMSE); this estimate together with the distribution of the biases was then used to compare an empirical distribution of the biases of synthetic estimates with the normal distribution.

As part of the publication of the 1970 Census of Housing, data on housing units dilapidated with all plumbing facilities collected in the 1960 Census of Housing were available for comparison with a set of synthetic housing estimates of DWAPF derived from the same data. The average mean square error for a set of M areas is given by

$$\text{Average MSE} = \frac{1}{M} \sum_{i=1}^{M} (x_i - x_i^*)^2 \qquad 5)$$

where

x_i is the census estimate for area i.

The use of formula 5 to estimate the average MSE assumes that the second term of formula 4 is negligible. This assumption is reasonable for large areas. The square root of the average MSE gives the estimate for the RMSE.

For all States we have two estimates of dilapidated housing units with all plumbing facilities in 1960; the 25-percent census estimate and a synthetic estimate based on a particular set of subgroups. The difference between these two estimates will be used as an estimate of the bias of the synthetic estimation procedure. Table 1 shows estimates of the proportion of the set of synthetic estimates for States with a relative bias within specified values. The relative bias for an area is defined as the difference between the synthetic estimate and the census estimate divided by the synthetic estimate.2/

Table 1. Distribution of Relative Biases of Synthetic Estimates for States

State estimate	Number of areas	Proportion with relative biases				
		0-9%	10-19%	20-29%	30-49%	50%+
1,000-2,499	7	0.14	0.29	0.29	0.14	0.14
2,500-4,999	6	0.50	0.17	0.17	0.17	0.0
5,000-9,999	13	0.23	0.46	0.15	0.08	0.08
10,000-19,999	16	0.38	0.38	0.19	0.06	0.0
20,000 or more	8	0.38	0.38	0.25	0.0	0.0
Total	50	0.32	0.36	0.20	0.08	0.04

This table shows that the proportion of estimates with large relative biases diminishes as the size of the synthetic estimate increases. For example, for synthetic estimates between 1,000 and 2,499 DWAPF housing units, 57 percent have relative biases of at least 20 percent; however, for synthetic estimates of over 10,000 units only about 25 percent have relative biases greater than 20 percent. When we consider the State synthetic estimates for all States, we note that 32 percent have relative biases of 20 percent or more, 12 percent have relative biases of 30 percent or more and 4 percent have relative biases of 50 percent or more. The average number of DWAPF housing units for States is about 13,000; the estimated average root mean square error is about 2,500; the ratio of the $\overline{RMSE}$ divided by the average size of State synthetic estimates of DWAPF housing units is 0.19. A high variability of the synthetic estimate is shown by the fact that the $\overline{RMSE}$ divided by the mean is about 20 percent. This shows that the synthetic estimates obtained do not account for a large part of the variability among areas. The synthetic estimates of housing units DWAPF are computed using a particular set of subgroups, defined in terms of tenure, race of head of household and other characteristics related to the quality of housing unit. The use of other subgroups would produce a different set of synthetic estimates.

From the point of view of ascertaining whether the average root mean square error can be used to make probability statements the results are more encouraging. Table 2 gives some comparisons of the distribution of the difference between State synthetic estimates of DWAPF housing units for 1960 and the estimates reported in the 1960 census. The first two columns of the table show the expected percentage of the normal distribution at different multiples of the standard error (σ). For example, 95 percent of the normal distribution is expected within two standard errors of the mean. The empirical distributions of the biases of the synthetic State estimates of DWAPF housing units are given in columns 3, 4 and 5. For example, 48 percent of the biases for estimates of total for States are less than one-half the estimated $\overline{RMSE}$.

Table 2. Comparison of Empirical Distribution of the Biases of State Synthetic Estimates of Dilapidated Housing Units with All Plumbing Facilities with the Theoretical Normal Distribution

Multiple of standard error (σ)	Normal probability	Distribution of bias of state estimates (n = 50)		
		Total Aver = 12,970 $\overline{RMSE}$ = 2,490	Inside SMSA's Aver = 8,170 $\overline{RMSE}$ = 1,540	Outside SMSA's Aver = 4,800 $\overline{RMSE}$ = 1,340
0.50	38%	48%	50%	62%
0.75	55	62	62	68
1.00	68	74	68	74
1.25	79	86	82	84
1.50	87	88	90	88
1.75	92	88	90	88
2.00	95	92	94	94
2.25	97.6	94	94	98
2.50	98.8	96	96	98
3.00	99.7	100	100	98

The empirical distributions of the biases of synthetic State estimates are closer to the mean (on the average) for values within one standard error of the mean, than expected for the normal distribution. For example, for half a standard error the normal distribution expects to cover about 38 percent of the distribution; for State totals, the empirical distribution actually includes 48 percent of the distribution; for estimated units within SMSA's, the empirical distribution includes 50 percent of the distribution and for units outside SMSA's the empirical distribution includes 62 percent of the distribution. However, for values which are more than two standard errors from the mean, the empirical results are reversed: the frequency of synthetic estimates with biases more than two standard errors from the mean is greater than expected for normal distributions; for State synthetic estimates about 8 percent had biases which differed by more than two standard errors from the mean. That is, on the average there are more outliers for synthetic estimates than would be expected for a normal distribution.

Table 3 shows empirical distributions of the biases of the estimates of DWAPF housing units for non-Negro renters in counties within SMSA's. The distribution was computed separately depending on the magnitude of the estimate of DWAPF housing: less than 100, 100 to 499, and 500 or more units. It is possible to carry out the analysis separately for the three groups through the use of formula 5 for the average root mean square error. The results reveal a similar

pattern to the results for State estimates given in Table 2. For values within one standard error the empirical distribution gives conservative estimates of the probability of occurrence, except for the distribution of DWAPF housing units with 500 or more units. However, for values at three standard errors we find more outliers than expected for the normal distribution. For example, for synthetic DWAPF estimates less than 100, the normal distribution expects only 0.3 percent of the cases to be further than three standard errors, but we find that 2.5 percent of the values have biases larger than three times the average root mean square error.

Table 3. Comparison of Empirical Distribution of the Biases of Synthetic Estimates for Non-Negro Renters in Counties within SMSA's of Dilapidated Housing Units with All Plumbing Facilities with the Theoretical Normal Distribution

Multiple of standard error (σ)	Normal probability	Distribution of bias for non-Negro renters in counties within SMSA's		
		DWAPF <100 $\overline{RMSE}$ = 39 n = 160	DWAPF (100-499) $\overline{RMSE}$ = 125 n = 219	DWAPF (500+) $\overline{RMSE}$ = 300 n = 84
0.50	38%	50%	49%	37%
0.75	55	66	66	52
1.00	68	84	79	64
1.25	79	89	85	75
1.50	87	93	89	85
1.75	92	94.3	92.2	89.2
2.00	95	95.6	94.5	95.2
2.25	97.6	96.2	96.8	98.8
2.50	98.8	96.2	97.7	98.8
3.00	99.7	97.5	98.6	98.8

Conclusions

Census data allow us to compute synthetic estimates and to compare them directly to the census estimates. Therefore, the biases of synthetic estimates can be obtained and their distribution analyzed directly.

The results presented comparing 1960 estimates of dilapidated housing units with all plumbing facilities with synthetically derived estimates show that the synthetic estimates are highly variable, but that the distribution of their biases is not too far from normal.

The analysis presented is based on a particular set of synthetic estimates; alternative sets using other variables should be investigated in order to be able to select the subgroups which account for a large proportion of the variability of the local area estimates, with an aim toward improving local area estimates. The results presented here, based on a particular set of synthetic estimates, may not necessarily generalize to possible alternative sets of synthetic estimates.

FOOTNOTES

1/ The change in procedure in estimating DWAPF housing units was necessary because a majority of housing units in the 1970 Census of Housing were enumerated by a mail-out, mail-back procedure; in addition, studies of these data for 1960 indicated that statistics based on enumerator ratings are highly unreliable.

2/ The synthetic estimate was used as denominator, instead of the reported estimate, because in 1970 the synthetic estimate was the only one available.

REFERENCES

[1] "Synthetic Estimates of Disability," published in 1968 by the National Center for Health Statistics, PHS publication No. 1759.

[2] A more detailed discussion of this project is available in "Estimation of the Error of Synthetic Estimates," by Maria Elena Gonzalez and Joseph Waksberg, presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria, August 18-25, 1973.

[3] Appendix 1 of "Estimation of the Error of Synthetic Estimates," by Maria Elena Gonzalez and Joseph Waksberg gives the derivation of the approximation given in formula 4.