COMPARISON OF THE DIFFERENT IMPUTATION PROCEDURES IN THE MONTHLY RETAIL SURVEY

Elizabeth T. Huang, U.S. Bureau of the Census

I. Introduction

The Monthly Retail Trade Survey is a panel survey that provides retail sales information from a probability sample. The list sample for any given data month consists of certainty units from a fixed panel and sampling units from a rotating panel. The certainty units report the sales monthly; while the sampling units from the rotating panel report both current and previous month sales every three months.

The missing item of the current month retail sales is imputed by multiplying the nonresponding unit's previous month sales (reported or imputed) by a measure of trend (the so called ratio of identicals) computed from those responding units whose size and kind of business characteristics are similar to the nonresponding units. The trend is calculated based on the weighted sum of the current month sales and the weighted sum of the previous month sales for each adjustment cell. The sample is partitioned into imputation cells defined by kind of business (3-or-4 digit Standard Industrial Classification (SIC) code is used), firm size (Group I and Group II; Group I has 3 different firm sizes (1-3, 4-10, 11+ establishments), Group II is the firm with 11+ establishments and certainty alpha) and size of sales (defined arbitrarily). Missing previous month sales for a sampling unit in a rotating panel is imputed in the similar fashion using historical data.

In this paper, we compare several ratio and regression adjustment procedures and a variety of imputation cell formations under a Monte Carlo study. We treated data reported in the Monthly Retail Trade Survey from 9 SIC's as our complete data set, and assumed that the data were missing at random. The missing items were imputed by different imputation procedures. The bias and mean square error (MSE) of the estimated totals for the given data set are derived in the following section. The conclusion in the study will be summarized in section III.

II. Monte Carlo Study

In Huang (1984), a Monte Carlo study was carried out to evaluate different imputation procedures based on a given set of complete data (reported list sample from SIC 562 in the December 1982 Retail Trade Survey). Five sets of incomplete data were generated from the complete data. For each of the five sets, data were randomly suppressed from each imputation cell of complete data according to its current imputation rate. The reader is cautioned that since only five incomplete data sets were used, the results of the comparisons may not give an accurate picture. In the following, the bias and MSE of the estimated total for a given complete data set were derived under the assumption that the missing data are a random sample of the complete data set.

A sample of size n is assumed to be drawn from a population of size N. Only one imputation cell is assumed. The sampling unit i has inclusion probability π_i. In a sample of size n, there are n_r units reported, and $n_{\bar{r}}$ units not reported. In the following we treat these reported n_r units as our complete data set. Assuming the nonresponse mechanism is ignorable, i.e., the data are missing at random, the incomplete data sets are generated in which $n'_{\bar{r}}$ units are suppressed randomly from the complete data set of n_r units, and different ratio type imputation procedures are used to impute $n'_{\bar{r}}$ missing units of y values (current month sales) using the auxiliary variable x (previous month sales), which is available for all n_r units.

Let Y be the estimated total using the complete data set of n_r units

$$Y = \sum_{i=1}^{n_r} y_i/\pi_i .$$

Let $\hat{Y}$ be the estimated total using incomplete data set of n_r units, of which n'_r units are reported, and $n'_{\bar{r}}$ units are imputed, i.e.,

$$\hat{Y} = \sum_{i=1}^{n'_r} y_i/\pi_i + \sum_{i=1}^{n'_{\bar{r}}} \hat{y}_i/\pi_i$$

where

$$\hat{y}_i = \hat{R}_{n'_r} x_i , \quad i=1,\ldots,n'_{\bar{r}} ,$$

$$n_r = n'_r + n'_{\bar{r}} ,$$

$\hat{R}_{n'_r}$ is one of the four ratio type estimators in (2.1)-(2.4) using n'_r units.

$$\hat{R}^{(1)} = (\Sigma_i\ y_i/\pi_i) \,/\, (\Sigma_i\ x_i/\pi_i) \qquad (2.1)$$

$$\hat{R}^{(2)} = \Sigma_i\ y_i \,/\, \Sigma_i\ x_i \qquad (2.2)$$

$$\hat{R}^{(3)} = (\Sigma_i \frac{1}{\pi_i}\frac{y_i}{x_i}) \,/\, (\Sigma_i \frac{1}{\pi_i}) \qquad (2.3)$$

$$\hat{R}^{(4)} = \frac{1}{n'_r} \sum_{i=1}^{n'_r} (y_i/x_i) \qquad (2.4)$$

$\hat{R}^{(1)}$ is the current imputation ratio.

The summations in (2.1)-(2.4) are over the n_r reported units in practice, over n'_r units in the Monte Carlo study. In addition for $\hat{R}^{(4)}$, the factor $1/n'_r$ is used in the Monte Carlo study, whereas in practice the factor $1/n_r$ is used. We also assume that the nonresponse rate is such that

$$\lim_{n'_{\bar{r}} \to \infty} f_{n'_{\bar{r}}} = \lim_{n'_{\bar{r}} \to \infty} n'_{\bar{r}}/n_r = f_{\bar{r}}$$

where $f_{\bar{r}}$ is fixed, and $0 < f_{\bar{r}} < 1$.

Lemma 1. Under above notations and assumptions, for

large n'_r,

$$E((\hat{Y} - Y) \mid n_r) \doteq -(n'_r/n_r) \sum_{i=1}^{n_r} (\hat{e}_i/\pi_i) \quad ,$$

$$= 0 \ , \text{ if } \quad \hat{R}_{n'_r} = \hat{R}^{(1)}$$

where

$$\hat{e}_i = y_i - \hat{R}_{n_r} x_i \ ,$$

$\hat{R}_{n_r}$ is any of the four ratio type estimators (2.1)-(2.4) using the complete data of size n_r, $E(\cdot \mid n_r)$ is the expectation over all possible samples of size n'_r drawn from n_r.

Proof:

Since $\hat{y}_i = \hat{R}_{n'_r} x_i$, it can be proved that

$\hat{R}_{n'_r} = \hat{R}_{n_r} + O_p(n'^{-1/2}_r)$ for $\hat{R}_{n'_r}$ and $\hat{R}_{n_r}$ being any form defined in (2.1) - (2.4) using n'_r and n_r units respectively.

We then have

$$E((\hat{Y} - Y) \mid n_r)$$

$$= E((\sum_{i=1}^{n'_r} y_i/\pi_i + \sum_{i=1}^{n'_r} \hat{y}_i/\pi_i - \sum_{i=1}^{n_r} y_i/\pi_i) \mid n_r)$$

$$= n'_r E((\frac{1}{n'_r} \sum_{i=1}^{n'_r} (\hat{y}_i - y_i) / \pi_i) \mid n_r)$$

$$\doteq - \frac{n'_r}{n_r} \sum_{i=1}^{n_r} \frac{1}{\pi_i} (y_i - \hat{R}_{n_r} x_i) \ .$$

When $\hat{R}_{n_r}$ is a form of $\hat{R}^{(1)}$,

$\sum_{i=1}^{n_r} \frac{1}{\pi_i} (y_i - \hat{R}_{n_r} x_i) = 0$, and hence

$E((\hat{Y} - Y) \mid n_r) = 0$.

Lemma 2. Under the notation and assumptions defined in this section, for large n'_r,

$$E((\hat{Y} - Y)^2 \mid n_r)$$

$$\doteq \frac{n'_r}{n_r} (\sum_{i=1}^{n_r} \frac{\hat{e}_i^2}{\pi_i^2} + \frac{(n'_r - 1)}{(n_r - 1)} \sum_{i \neq j}^{n_r} \frac{\hat{e}_i}{\pi_i} \frac{\hat{e}_j}{\pi_j})$$

where

$\hat{e}_i = y_i - \hat{R}_{n_r} x_i$

$\hat{R}_{n_r}$ is an estimator defined in (2.1) - (2.4) using the complete data set of size n_r.

Proof:

$$E((\hat{Y} - Y)^2 \mid n_r)$$

$$= E((\sum_{i=1}^{n'_r} (\hat{y}_i - y_i)/\pi_i)^2 \mid n_r)$$

$$= E(\sum_{i=1}^{n'_r} ((y_i - \hat{y}_i)/\pi_i)^2 \mid n_r)$$

$$+ E(\sum_{i \neq j}^{n'_r} ((y_i - \hat{y}_i)/\pi_i) ((y_j - \hat{y}_j)/\pi_j) \mid n_r)$$

$$\doteq \frac{n'_r}{n_r} \sum_{i=1}^{n_r} (\frac{y_i^2}{\pi_i^2} - 2 \frac{y_i x_i}{\pi_i^2} \hat{R}_{n_r} + \frac{x_i^2}{\pi_i^2} \hat{R}^2_{n_r})$$

$$+ \frac{n'_r (n'_r - 1)}{n_r (n_r - 1)} \sum_{i \neq j}^{n_r} \{ \frac{y_i y_j}{\pi_i \pi_j} - (\frac{y_j x_i}{\pi_j \pi_i} + \frac{y_i x_j}{\pi_i \pi_j}) \hat{R}_{n_r}$$

$$+ \frac{x_i x_j}{\pi_i \pi_j} \hat{R}^2_{n_r} \}$$

$$= \frac{n'_r}{n_r} (\sum_{i=1}^{n_r} \frac{\hat{e}_i^2}{\pi_i^2} + \frac{(n'_r - 1)}{(n_r - 1)} \sum_{i \neq j}^{n_r} \frac{\hat{e}_i}{\pi_i} \frac{\hat{e}_j}{\pi_j})$$

We used the fact that

$\hat{y}_i = \hat{R}_{n'_r} x_i$, and $\hat{R}_{n'_r} = \hat{R}_{n_r} + O_p(n'^{-1/2}_r)$ for $\hat{R}_{n'_r}$ being any forms defined in (2.1)-(2.4). Note that if $\hat{R}_{n_r}$ is of form of $R^{(1)}$, then

$$\sum_{i=1}^{n_r} \frac{\hat{e}_i^2}{\pi_i^2} = - \sum_{i \neq j}^{n_r} \frac{\hat{e}_i}{\pi_i} \frac{\hat{e}_j}{\pi_j}, \text{ because } \sum_{i=1}^{n_r} \frac{\hat{e}_i}{\pi_i} = 0 \ .$$

Lemma 3. Under the notation and assumptions defined in this section, redefine $\hat{y}_i = Rx_i$ for $i=1,\ldots, n'_r$, where R is a preassigned value, then

$$(1) E((\hat{Y} - Y) \mid n_r) = - \frac{n'_r}{n_r} \sum_{i=1}^{n_r} \frac{1}{\pi_i} (y_i - Rx_i) \ ,$$

$$(2) E((\hat{Y} - Y) \mid n_r)) = 0, \text{ iff } R = (\sum_{i=1}^{n_r} \frac{y_i}{\pi_i}) / (\sum_{i=1}^{n_r} \frac{x_i}{\pi_i})$$

$$(3) E((\hat{Y} - Y)^2 \mid n_r))$$

$$= \frac{n'_{\bar{r}}}{n_r} \sum_{i=1}^{n_r} \frac{1}{\pi_i^2} (y_i - Rx_i)^2$$

$$+ \frac{n'_{\bar{r}}\,(n'_{\bar{r}} - 1)}{n_r\,(n_r - 1)} \sum_{i \neq j}^{n_r} \frac{1}{\pi_i \pi_j} (y_i - Rx_i)(y_j - Rx_j)$$

(4) The value of R that minimizes (3) is

$$\hat{R}_{opt} = \frac{\left(\sum_{i=1}^{n_r} \frac{x_i y_i}{\pi_i^2} + \frac{(n'_{\bar{r}} - 1)}{(n_r - 1)} \sum_{i \neq j}^{n_r} \frac{x_i y_j}{\pi_i \pi_j} \right)}{\left(\sum_{i=1}^{n_r} \frac{x_i}{\pi_i^2} + \frac{(n'_{\bar{r}} - 1)}{(n_r - 1)} \sum_{i \neq j}^{n_r} \frac{x_i x_j}{\pi_i \pi_j} \right)^2} . \quad (2.5)$$

Proof: Since R is a fixed value, by the definition of $\hat{y}_i = Rx_i$, and the missing at random assumption of $n'_{\bar{r}}$ missing units of y_i, following the similar proof in Lemma 1 and 2, we can easily derive the results.

Lemma 4. The bias and MSE of the estimated total $\hat{Y}$ given n_r, by using $\hat{R}_{opt}$ to impute missing y_i, is

$$E((\hat{Y} - Y)|n_r) = - \frac{n'_{\bar{r}}}{n_r} \sum_{i=1}^{n_r} \frac{1}{\pi_i} (y_i - \hat{R}_{opt} x_i) ,$$

$$E((\hat{Y} - Y)^2|n_r) = \frac{n'_{\bar{r}}}{n_r} \sum_{i=1}^{n_r} \frac{1}{\pi_i^2} (y_i - \hat{R}_{opt} x_i)^2$$

$$+ \frac{n'_{\bar{r}}\,(n'_{\bar{r}} - 1)}{n_r\,(n_r - 1)} \sum_{i \neq j}^{n_r} \frac{1}{\pi_i \pi_j} (y_i - \hat{R}_{opt} x_i)\, x$$

$$(y_j - \hat{R}_{opt} x_j).$$

Proof:

$\hat{R}_{opt}$ is a function of n_r units of a complete data set which is the population that the incomplete data samples are randomly generated in the Monte Carlo study. For a given complete data set of size n_r, $\hat{R}_{opt}$ is a fixed value. Following the similar proof as in Lemma 3, we have the results.

An estimator of $\hat{R}_{opt}$ by using n'_r reported units in the Monte Carlo study is

$$\tilde{R}_{opt} = \frac{\frac{1}{n'_r} \left(\sum_{i=1}^{n'_r} \frac{x_i y_i}{\pi_i^2} + \frac{n'_{\bar{r}} - 1}{n'_r - 1} \sum_{i \neq j}^{n'_r} \frac{x_i y_j}{\pi_i \pi_j} \right)}{\frac{1}{n'_r} \left(\sum_{i=1}^{n'_r} \frac{x_i}{\pi_i^2} + \frac{n'_{\bar{r}} - 1}{n'_r - 1} \sum_{i \neq j}^{n'_r} \frac{x_i x_j}{\pi_i \pi_j} \right)^2}$$

$$\text{for } n'_r > 2. \quad (2.6)$$

Lemma 5. The bias and MSE of the estimated total $\hat{Y}$ given n_r, by using $\tilde{R}_{opt}$ to impute missing y_i is

$$E((\hat{Y} - Y)|n_r) \doteq - \frac{n'_{\bar{r}}}{n_r} \sum_{i=1}^{n_r} \frac{1}{\pi_i} (y_i - \hat{R}_{opt} x_i)$$

$$E((\hat{Y} - Y)^2|n_r) \doteq \frac{n'_{\bar{r}}}{n_r} \sum_{i=1}^{n_r} \frac{1}{\pi_i^2} (y_i - \hat{R}_{opt} x_i)^2$$

$$+ \frac{n'_{\bar{r}}\,(n'_{\bar{r}} - 1)}{n_r\,(n_r - 1)} \sum_{i \neq j}^{n_r} \frac{1}{\pi_i \pi_j} (y_i - \hat{R}_{opt} x_i)\, x$$

$$(y_j - \hat{R}_{opt} x_j) .$$

Proof:

Following the similar proofs in lemmas 1 and 2, and the fact that for large n'_r, $\tilde{R}_{opt} = \hat{R}_{opt} + O_p(n_r'^{-1/2})$, the results follow.

To use $\tilde{R}_{opt}$ we need to know the number of nonresponse items $n'_{\bar{r}}$, and the number of response items n'_r in the sample. If the factor $(n'_{\bar{r}} - 1)(n'_r - 1)^{-1}$ is not used in $\tilde{R}_{opt}$, then

$$\tilde{R}_{opt} = \frac{\sum_{i=1}^{n'_r} \frac{x_i y_i}{\pi_i^2} + \sum_{i \neq j}^{n'_r} \frac{x_i y_j}{\pi_i \pi_j}}{\sum_{i=1}^{n'_r} \frac{x_i}{\pi_i^2} + \sum_{i \neq j}^{n'_r} \frac{x_i x_j}{\pi_i \pi_j}} = \frac{\sum_{i=1}^{n'_r} \frac{y_i}{\pi_i}}{\sum_{i=1}^{n'_r} \frac{x_i}{\pi_i}} .$$

$\tilde{R}_{opt}$ is reduced to the current imputation ratio $R^{(1)}$.

Another estimator of $\hat{R}_{opt}$ is

$$R^{(5)} = \left(\sum_{i=1}^{n'_r} \frac{x_i y_i}{\pi_i^2} \right) \Big/ \left(\sum_{i=1}^{n'_r} \frac{x_i^2}{\pi_i^2} \right) . \quad (2.7)$$

It can be shown that when $R^{(5)} x_i$ is used to impute missing y_i, for large n'_r, the bias and MSE of $\hat{Y}$ for a given complete data set n_r are given in lemma 1 and 2, where

$$\hat{R}_{n_r} = \left(\sum_{i=1}^{n_r} \frac{x_i y_i}{\pi_i^2} \right) \Big/ \left(\sum_{i=1}^{n_r} \frac{x_i^2}{\pi_i^2} \right) .$$

If the inclusion probability π is not used in (2.7), we have

$$R^{(6)} = \left(\sum_{i=1}^{n'_r} x_i y_i \right) \Big/ \left(\sum_{i=1}^{n'_r} x_i^2 \right) \quad (2.8)$$

which is the least squares estimate of R of the ratio model with constant error variance (i.e., $y = Rx + e$, e is independently identically distributed with mean zero and variance σ^2). It can be shown that the bias and

MSE of $\hat{Y}$ for a given complete data set are given in lemma 1 and 2 with

$$\hat{R}_{n_r} = \left(\sum_{i=1}^{n_r} x_i y_i\right) / \left(\sum_{i=1}^{n_r} x_i^2\right).$$

If the ordinary regression estimator is used to impute missing item y_i, $i = 1,\ldots, n'_{\bar{r}}$,

$$\hat{y}_i = \hat{\alpha}_{n'_r} + \hat{\beta}_{n'_r} x_i , \qquad (2.9)$$

where

$$\hat{\alpha}_{n'_r} = \bar{y}_{n'_r} - \hat{\beta}_{n'_r} \bar{x}_{n'_r} , \qquad (2.10)$$

$$\bar{y}_{n'_r} = \frac{1}{n'_r} \sum_{i=1}^{n'_r} y_i ,$$

$$\bar{x}_{n'_r} = \frac{1}{n'_r} \sum_{i=1}^{n'_r} x_i ,$$

$$\hat{\beta}_{n'_r} = \sum_{i=1}^{n'_r} (x_i - \bar{x}_{n'_r})(y_i - \bar{y}_{n'_r}) / \sum_{i=1}^{n'_r} (x_i - \bar{x}_{n'_r})^2 , \qquad (2.11)$$

then the bias and MSE of $\hat{Y}$ for a given complete data set n_r are given in lemma 1 and 2, with

$$\hat{e}_i = y_i - \hat{\alpha}_{n_r} - \hat{\beta}_{n_r} x_i , \text{ and} \qquad (2.12)$$

$$\hat{\alpha}_{n_r} = \bar{y}_{n_r} - \hat{\beta}_{n_r} \bar{x}_{n_r} ,$$

$$\bar{y}_{n_r} = \frac{1}{n_r} \sum_{i=1}^{n_r} y_i ,$$

$$\bar{x}_{n_r} = \frac{1}{n_r} \sum_{i=1}^{n_r} x_i ,$$

$$\hat{\beta}_{n_r} = \sum_{i=1}^{n_r} (x_i - \bar{x}_{n_r})(y_i - \bar{y}_{n_r}) / \sum_{i=1}^{n_r} (x_i - \bar{x}_{n_r})^2 .$$

The above results can be extended to more than one imputation cell if we generate the incomplete data set from the complete data set independently for each imputation cell. Let $\hat{Y}_k$, Y_k be the estimated totals of the incomplete data set and the complete data set respectively from imputation cell k, k=1,...,K. Then

$$\hat{Y} = \sum_{k=1}^{K} \hat{Y}_k , \quad Y = \sum_{k=1}^{K} Y_k .$$

Let n_{r_k} be the sample size of the reported data of imputation cell k, and we randomly suppress $n'_{\bar{r}_k}$ units from this complete data set. Let n_r be the sample size of the complete data set from all K imputation cells.

The bias of the estimated total $\hat{Y}$ given the complete data set is

$$E((\hat{Y} - Y) \mid n_r)$$

$$= \sum_{k=1}^{K} E((\hat{Y}_k - Y_k) \mid n_{r_k})$$

$$\doteq - \sum_{k=1}^{K} \frac{n'_{\bar{r}_k}}{n_{r_k}} \sum_{i=1}^{n_{r_k}} \frac{1}{\pi_{ki}} \hat{e}_{ki},$$

where

$$\hat{e}_{ki} = y_{ki} - \hat{R}_{n_{r_k}} x_{ki} , \quad k=1,\ldots,K, \quad i=1,\ldots,n_{r_k} .$$

$\hat{R}_{n_{r_k}}$ is a ratio estimator of (2.1) to (2.4) and (2.6)-(2.8) using the complete data set of size n_{r_k} from each imputation cell k. For the regression estimator defined in (2.9), $\hat{e}_i$ is defined in (2.12) for each imputation cell. Similarly, the mean square errors of the estimated total $\hat{Y}$ given the complete data set can be written as

$$E((\hat{Y} - Y)^2 \mid n_r)$$

$$= \sum_{k=1}^{K} E((\hat{Y}_k - Y_k)^2 \mid n_{r_k})$$

$$\doteq \sum_{k=1}^{K} \frac{n'_{\bar{r}_k}}{n_{r_k}} \left(\sum_{i=1}^{n_{r_k}} \frac{\hat{e}_{ki}^2}{\pi_{ki}^2} + \frac{(n'_{\bar{r}_k} - 1)}{(n_{r_k} - 1)} \sum_{i \neq j}^{n_{r_k}} \frac{\hat{e}_{ki}}{\pi_{ki}} \frac{\hat{e}_{kj}}{\pi_{kj}} \right).$$

Under the assumption that the data are missing at random, we have already shown that the bias and MSE of the estimated total given the complete data set using various ratio and regression imputation procedures are functions of residuals of the complete data, and the nonresponse rate of each imputation cell. To compare different ratio and regression imputation procedures defined in (2.1) - (2.4) and (2.6) - (2.9) empirically, we can thus compute these biases and MSE's using Monthly Retail Trade Survey reported data and current nonresponse rates without randomly generating all possible incomplete samples.

The Monthly Retail Trade Survey reported data of December 1982 for nine SIC's were used to compare the bias and MSE of the estimated totals of the different ratio and regression type imputation procedures. The trends are calculated from the reported data of each imputation cell by these different estimators. The trends calculated by the optimum ratio procedure and the current imputation procedure are fairly close for most SIC's. (See Huang (1986)). The bias and MSE of the estimated totals by using these different imputation procedures are tabulated in Tables 1.1 and 1.2. Algebraically, we have already shown that given the complete data set, the current imputation procedure is unbiased with respect to the estimated reported total for each

imputation cell, and so are the empirical results. The relative biasses (bias/estimated reported total) of the other ratio imputation procedures are relatively small, less than 3% for most data.

The optimum ratio imputation procedure, $\tilde{R}_{opt}$, gave the minimum mean square error among all the ratio type imputation procedures. However, the gain in efficiency of $\tilde{R}_{opt}$ in comparing with the current imputation procedure is at most 0.002. The current imputation procedure is fairly competitive with the optimum ratio imputation procedure and is easier to compute.

Note that all the inferences of the Monte Carlo study are restricted to the data we used. The derivations of the bias and MSE are based on the assumption that the data are missing at random. The data used for the Monte Carlo study were examined to investigate the validity of this assumption. The imputation rates by sales classes of each imputation cell were calculated. There is no apparent relationship between item nonresponse rates and sales classes. The imputation rates by regions of each imputation cell were also calculated. The imputation rates are different for different regions but there is no specific pattern.

Based on the current imputation procedure, we also used mean square error (MSE) criterion to evaluate different imputation cell definitions, e.g., to answer the question of what quantiles (median, 1/4 or 1/8 or 1/16 quantiles) should be used for the cutoff of sales size classes if sales sizes are used within each firm size (group I and II) for imputation cell definition as opposed to the current fixed cutoff. The reported data for 9 SIC's from December 1982 were used. The empirical results showed that for SIC 562 the smaller the imputation cell is, the better the MSE. However, the most drastic reduction in MSE is the cell definition using 1/4 quantiles as sales cutoffs. There was an approximate 44% reduction in MSE as compared to the MSE under the current imputation cell definition. Using 1/8 quantiles as sales cutoffs a further 6% reduction over 1/4 quantiles was observed; and using 1/16 quantiles a further 3% reduction over 1/8 quantiles was observed. Overall, the empirical results varied by SIC's. In 6 of 9 SIC's, the reductions in MSE ranged from 12% (-3%) to 59% (44%) by using 1/8 (1/4) quantiles instead of the current fixed cutoff. Most of these reductions in MSE came from group II. For SIC's 541, 551, and 5813, there was little, if any, gain in using any of the quantiles considered. (See Huang (1986)).

III. Summary

We have evaluated the bias and MSE of the estimated totals using different ratio and regression type imputation procedures (including the currently used imputation procedure) under a Monte Carlo study for a given data set.

Under the assumption that the data are missing at random, the bias and MSE of the estimated total using different ratio type imputation procedures with respect to the estimated reported total were derived for the given reported data. An optimum ratio imputation estimator was also derived along with several variants. The bias and MSE were calculated for each of nine SIC's using December 1982 retail sales data. For the given data set, the empirical results showed that the estimated total using the current imputation procedure is unbiased and has the second smallest MSE among all ratio type imputation procedures in the study.

Since the decrease of the MSE by using the optimum imputation procedure is trivial, and extra computation and information are needed to implement this optimum imputation procedure, we do not recommend any changes of the current ratio type imputation procedure in the Monthly Retail Trade Survey.

For the given data set, there is no apparent relationship of nonresponse rate with sales within each imputation cell for all nine SIC's.

In the current imputation cells, for some SIC's, the number of establishments in Group II dominates the number in Group I; for other SIC's, the number of establishments in Group I dominates the number in Group II. The empirical results suggested that for some of the nine SIC's included in the study, we can do better by using alternative imputation cells, i.e., use sales quantiles as cutoffs within groups as opposed to the current fixed sales cutoffs. The decrease in MSE in 6 of 9 SIC's ranges from 12% to 59% by using 1/8 quantiles. We recommend that changes in the current imputation cells definition be considered, especially where empirical studies show that a significant reduction in the MSE can be achieved by increasing the number of imputation cells. We also suggest that further similar empirical studies be carried out on recent monthly data to provide a basis for changes in cell definitions for other SIC's. This will tell us whether there is a gain in using alternative imputation cells and what quantiles to use for a given SIC in a given month.

VI. Acknowledgement

The author gratefully acknowledges the helpful comments and criticisms of Cary Isaki, Nash Monsour and Carl Konschnik.

V. References

1. Bailey, L., Chapman, D.W. and Kasprzyk, D. (1985). "Nonresponse Adjustment Procedures at the Bureau of the Census: A Review. Paper presented at First Annual Census Bureau Research Conference, Reston, Virginia.
2. Cassel, C.M., Sarndal, C.E., and Wretman, J.H. (1979). "Some Uses of Statistical Models in Connection with the Nonresponse Problem," Incomplete Data in Sample Surveys. Volume 3. W.G. Madow, I. Olkin and D.B. Rubin, eds., Academic Press: New York pg. 143-160.
3. Cochran, W.G. (1977). Sampling Techniques. John Wiley and Sons, New York.
4. Fuller, W.A. (1976). Introduction to Statistical Time Series. John Wiley and Sons, New York.
5. Huang, E.T. (1984). "An Imputation Study for the Monthly Retail Trade Survey," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 610-615.
6. Huang, E.T. (1986). "Report on the Imputation Research for the Monthly Retail Trade Survey" Statistical Research Division report series, report number: CENSUS/SRD/RR-86-09/. Bureau of the Census.
7. Little, R.J.A (1986). "Missing Data in Census Bureau Surveys". Paper presented at Second Annual Census Bureau Research Conference, Reston, Virginia.

TABLE 1.1 The Bias (Relative Bias (%)) of the Estimated Total
By Using Different Imputation Procedures
December 1982

Unit: U.S. Dollars

SIC	n	Estimated Reported Total	$R^{(1)}$	$R^{(2)}$	$R^{(3)}$	$R^{(4)}$	$\tilde{R}_{opt}$	$R^{(5)}$	$R^{(6)}$	Regression Estimator
562 (Women's Ready-to-Wear Stores)	1445	1,636,658,834	0	4,745,104	18,303,757	29,244,857	-495,262	-18,584,772	-49,527,438	73,566,518
(%)			(0)	(0.290)	(1.118)	(1.787)	(-0.030)	(-1.136)	(-3.026)	(4.495)
521 (Building Materials Stores)	635	1,933,849,833	0	7,282,673	14,289,364	24,869,789	-133,689	-5,946,523	-850,347	8,085,767
(%)			(0)	(0.377)	(0.739)	(1.286)	(-0.007)	(-0.307)	(-0.044)	(0.418)
531 (Department Stores)	7557	14,758,285,090	0	202,901	71,895,263	72,247,748	-340,476	-77,969,143	-78,411,093	425,515
(%)			(0)	(0.001)	(0.487)	(0.490)	(-0.002)	(-0.528)	(-0.531)	(0.003)
541 (Grocery Stores)	2428	12,374,995,572	0	51,782,545	-13,933,939	24,927,603	-409,814	-1,945,308	59,579,151	39,159,736
(%)			(0)	(0.418)	(-0.113)	(0.201)	(-0.003)	(-0.016)	(0.481)	(0.316)
551 (Motor Vehicle Dealers)	753	14,565,413,603	0	14,544,341	59,652,581	61,644,105	-3,169,391	-53,164,307	-5,047,197	57,052,376
(%)			(0)	(0.100)	(0.410)	(0.423)	(-0.022)	(-0.365)	(-0.035)	(0.392)
572 (Household Appliances, Radio,TV Stores)	500	571,806,693	0	6,876,815	2,920,688	5,941,472	-266,227	-4,071,594	13,020,508	-2,565,749
(%)			(0)	(1.203)	(0.511)	(1.039)	(-0.047)	(-0.712)	(2.277)	(-0.449)
5812 (Eating Places)	1531	6,055,819,018	0	-2,022,721	31,094,652	47,959,481	-720,701	-18,773,433	-41,890,604	24,760,883
(%)			(0)	(-0.033)	(0.513)	(0.792)	(-0.012)	(-0.310)	(-0.692)	(0.409)
5813 (Drinking Places)	420	642,146,909	0	-163,904	-647,688	-151,990	59,308	1,057,819	634,619	-2,087,831
(%)			(0)	(-0.026)	(-0.101)	(-0.024)	(0.009)	(0.165)	(0.099)	(-0.325)
592 (Liquor Stores)	542	1,740,095,873	0	-3,303,540	3,987,841	1,672,854	-345,361	-2,725,450	-18,100,886	2,683,752
(%)			(0)	(-0.190)	(0.229)	(0.096)	(-0.020)	(-0.157)	(-1.040)	(0.154)

TABLE 1.2 The MSE of the Estimated Total By Using Different Imputation Procedures
(And the Ratio to its Current Imputation Procedure)
December 1982

Unit: $\$10^6$

SIC	n	$R^{(1)}$	$R^{(2)}$	$R^{(3)}$	$R^{(4)}$	$\tilde{R}_{opt}$	$R^{(5)}$	$R^{(6)}$	Regression Estimator
562 (Women's Ready-to-Wear Stores)	1445	122,250,188	149,423,484	485,083,727	542,620,951	122,151,733	254,714,015	1,384,741,428	2,236,292,027
		(1)	(1.2223)	(3.9680)	(4.4386)	(0.9992)	(2.0835)	(11.327)	(18.293)
521 (Building Materials Stores)	635	373,293,260	508,178,456	634,679,773	810,125,338	373,238,170	405,332,341	425,909,403	482,528,485
		(1)	(1.3613)	(1.7002)	(2.1702)	(0.9999)	(1.0858)	(1.1410)	(1.2926)
531 (Department Stores)	7557	148,129,951	148,187,203	5,362,890,768	5,363,246,499	148,069,922	6,176,598,292	6,178,473,892	121,809,138
		(1)	(1.0004)	(36.204)	(36.206)	(0.9996)	(41.697)	(41.710)	(0.8223)
541 (Grocery Stores)	2428	975,155,627	2,732,057,367	1,383,526,763	2,492,746,633	975,070,021	1,306,876,310	2,902,580,498	2,341,096,627
		(1)	(2.8017)	(1.4188)	(2.5563)	(0.9999)	(1.3402)	(2.9765)	(2.4007)
551 (Motor Vehicle Dealers)	753	2,617,510,636	2,906,824,784	4,919,457,310	5,040,731,723	2,610,996,151	4,108,583,221	2,689,858,236	5,023,064,502
		(1)	(1.1105)	(1.8794)	(1.9258)	(0.9975)	(1.5697)	(1.0276)	(1.9190)
572 (Household Appliances, Radio/TV Stores)	500	25,604,966	60,624,441	41,244,917	52,676,300	25,562,855	35,419,411	134,778,334	83,056,483
		(1)	(2.3677)	(1.6108)	(2.0573)	(0.9984)	(1.3833)	(5.2638)	(3.2438)
5812 (Eating Places)	1531	410,794,971	477,812,049	1,134,373,917	1,339,936,313	410,512,304	551,848,720	1,168,011,026	669,950,846
		(1)	(1.1631)	(2.7614)	(3.2618)	(0.9993)	(1.3434)	(2.8433)	(1.6309)
5813 (Drinking Places)	420	4,150,594	4,205,883	4,691,465	4,309,152	4,145,352	5,229,200	4,577,926	8,792,564
		(1)	(1.0133)	(1.1303)	(1.0382)	(0.9987)	(1.2599)	(1.1030)	(2.1184)
592 (Liquor Stores)	542	110,351,071	150,489,266	183,571,784	118,555,857	110,152,107	119,873,690	428,523,608	144,471,410
		(1)	(1.3637)	(1.6635)	(1.0744)	(0.9982)	(1.0863)	(3.8833)	(1.3092)