BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-84/13

AN IMPUTATION STUDY FOR THE MONTHLY RETAIL TRADE SURVEY

by

Elizabeth T. Huang
Statistical Research Division
U.S. Bureau of the Census
Room 3524, F.O.B. #3
Washington, D.C. 20233

(301)763-1820

Recommended by: Myron J. Katzoff

Report completed: July 30, 1984

Report issued: August 3, 1984

An Imputation Study for the Monthly Retail Trade Survey

I. Introduction

The Census Bureau conducts the monthly Retail Trade Survey of the business universe in order to provide timely estimates of the level and trend sales. The data for each establishment are subjected to a series of edit checks and to be imputed if they are missing. The problem of handling missing data for the Monthly Retail Trade Survey is examined in the paper. The Monthly Retail Trade Survey is composed of a list sample and an area sample, where the list sample contains 96% of the total sample size. The list sample consists of a fixed panels of certainty units (which report every month) and rotating panels of non-certainty sampling units (which report every three months). A stratified random sampling design was used (See (5)). The main variables collected in the rotating panel cases are the monthly retail sales for the current month and the previous month. For fixed panel cases, only current monthly sales are collected. These items are sometimes not reported or suppressed because of edit failure.

The current imputation procedure in the Monthly Retail Trade Survey takes advantage of the rotating nature of the sample panels and 'historical' data. The procedure operates by multiplying a nonresponding unit's 'historical' data by a measure of trend computed from those responding units whose size and kind of business characteristics are similar to the nonresponding unit's. This method assumes that trends in the nonresponse stratum are similar to those in the response stratum. The sample is partitioned into imputation cells defined by kind of business (KB), firm size (Group I and Group II) and size of sales. Kind of business is based upon one of the Standard Industrial Classification (SIC) codes defined in the Standard Industrial Classification Mannual. The SIC codes are two, three, or four digits depending on the detail of classification. In each imputation cell, the trend is calculated from the reported items. If the 'current' month sales are missing, it is imputed based on the 'previous' month sales of the same unit. Let y_i be the current month sales and x_i be the previous month sales of the i^{th} unit that reported in the current month. Let z_i be the previous month sales reported 3 months ago by the i^{th} unit of the same panel. For the list sample of noncertainty units, the trends or the so-called ratios of identicals for each imputation cell are calculated by

$$\hat{R}_p = \sum w_i x_i / \sum w_i z_i \qquad (1.1)$$

$$\hat{R}_c = \sum w_i y_i / \sum w_i x_i \qquad (1.2)$$

where w_i denotes the sampling weight of the i^{th} responding unit. The summations in $\hat{R}_p$ are taken over all units in the imputation cell whose data x_i and z_i were reported. The ratio, $\hat{R}_p$, estimates the previous month to previous three months ago sales trend for each imputation cell in the domain of respondents. Similarly, the summations in $\hat{R}_c$ are taken over all units in the imputation cell whose data, y_i and x_i were reported. The ratio, $\hat{R}_c$, estimates the current month to previous month sales trend for each imputation cell in the domain of respondents.

After forming the ratio of identicals for each imputation cell, the next step is to test whether the ratio $\hat{R}_p$ satisfies the conditions

$\hat{R}_p \in [m_1, M_1]$ and $N_1 > 15$, where N_1 denotes the number of units defining the ratio $\hat{R}_p$. The interval limits, m_1 and M_1, vary by KB and by month. If one or both of these conditions are not met in a given imputation cell, then the ratio $\hat{R}_p$ is recalculated over all reported x_i and z_i units within a collapsed cell which is defined by KB and firm size. In a similar manner, the ratio $\hat{R}_c$ is tested for each imputation cell for possible distortion and recalculated when necessary. If the ratio is accepted, the ratio will be used to impute the missing item. The ratio in (1.1) is used to impute the missing item in the case of previous month sales (x), and the ratio in (1.2) is used to impute the missing item in the case of current month sales (y).

Cassel, Särndal and Wretman (1979) outlined an approach that builds on an underlying linear regression model for estimation of the finite population mean when nonresponse has occurred. They developed two estimators; one estimator can be constructed to have built-in adjustment for varying response probabilities, and another estimator is simplified by leaving out such adjustment. The latter case takes the risk of design biased inferences when nonresponse occurs and the underlying model is false. They also extended the techniques to the case when only sample auxilary information is available instead of population auxilary information.

The current imputation procedure of Monthly Retail Trade Survey (as I view it) is a kind of latter case where the linear model going through the origin is assumed for each of the imputation cells of the sample . The missing item is imputed from the model using the sample auxiliary information.

For each KB, for each imputation cell ij (group size x sales size) i=1,...I, j=1,...J, the current month sales (y) are assumed to have a linear relationship with the previous month sales (x),

$$y = R_{ij}\ x + \varepsilon \qquad \varepsilon|x \sim N\ (0,\ x\ \sigma^2) \qquad (1.3)$$

where x is assumed to be known for every unit in the sample.

When nonresponse y occurs, the $\hat{R}_{ij}$ is calculated from the response data of imputation cell ij by using (1.2) which is a least squares estimate of R_{ij} under model (1.3) and incorporating the sampling weights. If model is true, both least squares estimate ($\Sigma\ y_i/\Sigma x_i$) and $\hat{R}_{ij}$ are unbiased estimate of R_{ij}. $\hat{R}_{ij}$ is one of the estimators under model (1.3) discussed in Cassel, Särndal and Wretman (1979). The imputed value for the missing item y is $\hat{R}_{ij}$ x. The current imputation procedure puts further restrictions on the estimated $\hat{R}_{ij}$. If $\hat{R}_{ij}$ is not in the prior limits [m_1, M_1] or the number of respondents in cell ij is less than 15, a collapsed cell is defined within group i. The following linear model is assumed in the collapsed cell i,

$$y = R_i\ x + \varepsilon \qquad \varepsilon|x \sim N(0,\ x\sigma^2) \qquad (1.4)$$

which assumes that the R differs by firm group size. By way of definition, group 2 consists of the certainty company reporting units with 11 or more retail establishments, while group 1 consists of the rest of the list sample establishments.

The same model assumption is used for the previous month sales (x) of the current month reporting unit, and the previous month sales (z) reported 3 months ago. All missing items of the previous month sales (x) are imputed before imputing the missing items of current month sales (y).

When nonresponse occurs, under the current stratified sample design and the current imputation procedure, the Horvitz-Thompson estimator of total sales y is a ratio type estimator. This is illustrated by the two imputation situations described below.

(a) <u>The trend (ratio of identicals) calculated from the respondents falls in the prior limits</u>. For each KB, let $\hat{R}_k$ be the ratio of identicals for each imputation cell k,

$$\hat{R}_k = \left(\sum_h \sum_{i=1}^{n_{hrk}} w_h\, y_{hik}\right) / \left(\sum_h \sum_{i=1}^{n_{hrk}} w_h\, x_{hik}\right) \tag{1.5}$$

where w_h is the sampling weight for the h^{th} stratum,

n_{hrk} is the number of the respondents in the h^{th} stratum, and imputation cell k,

y_{hik} is the current month sales of the i^{th} responding unit in the h^{th} stratum, and imputation cell k,

x_{hik} is the previous month sales of the i^{th} responding unit in the h^{th} stratum, and imputation cell k.

Let n_{hk} be the number of units in stratum h and imputation cell k. For each KB, for each imputation cell k, the total of the current month sales (Y) in imputation cell k is estimated in two parts, the respondent part and the imputed part:

$$\hat{Y}_k = \sum_h w_h \left[\sum_{i=1}^{n_{hrk}} y_{hik} + \sum_{i=n_{hrk}+1}^{n_{hk}} \hat{R}_K\, x_{hik}\right]$$

$$= \sum_h w_h \left[\sum_{i=1}^{n_{hrk}} y_{hik} + \sum_{i=1}^{n_{hk}} \hat{R}_K\, x_{hik} - \sum_{i=1}^{n_{hrk}} \hat{R}_K\, x_{hik}\right]$$

$$= \sum_h w_h \sum_{i=1}^{n_{hk}} \hat{R}_k\, x_{hik} = \hat{R}_k\, \hat{X}_k \tag{1.6}$$

Summing over all imputation cells, the estimated total Y of the KB is

$$\hat{Y} = \sum \hat{R}_k\, \hat{X}_k \tag{1.7}$$

where $\hat{X}_k$ is the estimated total X for imputation cell k from the stratified sample, and all of the x_{hi} (reported or imputed) in the cell k is used in the calculation.

The estimate of the total Y is the sum of the combined ratio type estimators over all imputation cells.

Note that when $n_{hrk} = n_{hk}$, $\hat{Y}$ reduces to the Horvitz-Thompson estimator of the total Y.

(b) Case (a) fails, and a collapsed cell is used (cells collapsed over sales size within the group size). For each KB, the $\hat{R}_k$ defined in (a) is calculated over the respondents in the collapsed cell k, and we again have $\hat{Y} = \Sigma_k \hat{R}_k \hat{X}_k$ as in (1.7) except summation is over the collapsed cells instead of original imputation cells.

We next examine the validity of the current imputation model as well as other alternative models using Monthly Retail Trade Survey data.

II. Examining Current Monthly Retail Trade Survey Data - December 1982 Retail Trade Survey - SIC 562 (Women's Ready-to-Wear Stores)

Monthly retail sales reported data were examined to see whether the current model holds. As mentioned before, the main variables in the Monthly Retail Trade Survey are the monthly retail sales for the current month (y) and the previous month (x) for the rotating panel cases. For the fixed panel, current month sales are collected every month. Excluding the birth and death units, the previously described ratio type imputation is currently used to estimate the missing items of y assuming all x's are known. For each imputation cell, all the current month sales and previous month sales data are assumed to have the following relationship:

$$y = \beta x + \varepsilon \quad , \quad \varepsilon \sim N(0, x\sigma^2) \tag{2.1}$$

The missing item y_i is currently estimated by $\hat{\beta} x_i$,

where $\hat{\beta} = (\sum_i w_i y_i) / (\sum_i w_i x_i)$,

w_i is the sampling weight corresponding to unit i, and the summation is taken over all reported x_i and y_i's.

Four alternative linear regression models are examined <u>for each imputation cell</u>:

(a) $y = \alpha + \beta x + \varepsilon$, $\varepsilon \sim n(0, \sigma^2)$ (2.2)

(b) $y = \alpha + \beta x + \varepsilon$, $\varepsilon \sim n(0, x\sigma^2)$ (2.3)

(c) $y = \alpha + \beta x + \varepsilon$, $\varepsilon \sim n(0, x^2\sigma^2)$ (2.4)

(d) $\log y = \alpha + \beta \log x + \varepsilon$, $\varepsilon \sim n(0, \sigma^2)$ (2.5)

If the intercept α is not different from 0, the linear regression models in (2.2), (2.3), and (2.4) will reduce to the following models:

(a') $y = \beta x + \varepsilon$, $\varepsilon \sim n(0, \sigma^2)$ (2.6)

(b') $y = \beta x + \varepsilon$, $\varepsilon \sim n(0, x\sigma^2)$ (2.7)

(c') $y = \beta x + \varepsilon$, $\varepsilon \sim n(0, x^2\sigma^2)$ (2.8)

The least squares estimates for β in models (2.6), (2.7), and (2.8) for a simple random sample of size n are

$$(\sum_{i=1}^{n} x_i y_i)/(\sum_{i=1}^{n} x_i^2), \quad (\sum_{i=1}^{n} y_i)/(\sum_{i=1}^{n} x_i), \quad (1/n)\sum_{i=1}^{n} (y_i/x_i)$$

respectively.

The data used are from the December 1982 Monthly Retail Trade Survey for SIC 562. The total sample size is 2937. The tabulation of the data according to the reported and imputed (R/I) code and panel code is given in Table (2.1). The reported and imputed code is given for both current month and previous month. The definitions of the R/I code, panel and firm size code are given in Appendix A.

The data used for model fitting are restricted to establishments in the list sample with reported nonzero sales for both current and previous months. (The reported data here are defined to have R/I code 1 or 4 for both previous and current months). The sample size for the reported data is 1448.

The 4 current imputation cells and 2 collapsed imputation cells for SIC 562 are defined as follows:

Collapsed imputation cell	Imputation cell
1. Group 2 (firm size code = 6)	1. Sales* $\geq$ \$50,000, Group 2
	2. Sales* < \$50,000, Group 2
2. Group 1 (firm size code = 2,3,4)	3. Sales* $\geq$ \$50,000, Group 1
	4. Sales* < \$50,000, Group 1

*The sales size indicator depends on which panel the unit belongs. For fixed panel, the previous month sales are used. For rotating panels, the current month sales of 3 months ago are used.

The data of imputation cell 1 (group = 2, sales* $\geq$ \$50,000) are first used in fitting the different models. By looking at the plots of the residuals, it seems that model (2.2) with constant variance does not fit well. To find the approximate relationship of the variance of the current month sales y with the previous month sales x, units were first sorted by the previous month sales and then grouped with 20 units in each class. The variance or the standard error of y and the mean of x for each class were calculated. The results are given in Table 2.2. For imputation cell 1, the ratio of s_y/x is pretty much constant for each class, while the ratio of s_y^2/x is increasing rapidly, especially for the last three classes.

Alternatively, the relationship of the variance of y with the mean of x can be estimated by least squares method using the log transformation of the following

$$(s_y^2)_i = \lambda \ \overline{x}_i^{\rho} \ , \ i = 1,\ldots k.$$

The estimated λ, ρ for 4 imputation cells and 2 collapsed cells for the data of SIC 562 of December 1982 are tabulated in Table 2.2A.

It seems that the error variance $x^2 \sigma^2$ is more appropriate than $x\sigma^2$ for each imputation cell. (The error variance for each imputation cell of other KB's and of SIC 562 of February 1983 was also investigated. It was tabulated in Table 2.2A.) Since the plots of the data of the imputation cells of SIC 562 of December 1982 are proportional to x in some form (see Figure 2.1), model (2.2) of constant variance was not used to fit the remaining of the imputation cells data. A linear model with error variance $x^2\sigma^2$ (equation (2.4)) was then used to fit each of the 4 imputation cells data to see whether the intercept is significantly different from zero.

In fitting each imputation cell data, the outliers were also examined and deleted in the analysis of residuals based on the Cook and Studentized statistics.

The results of the model fitting for each imputation cell and collapsed cell of December 1982 for SIC 562 are tabulated in Table 2.3. If we treat the data as a simple random sample, we can conclude the following:

(1) The error variance of the linear regression model for each imputation cell is approximately $x^2\sigma^2$. (Same conclusion for February 1983's data.)

(2) By fitting the linear regression model with error variance $x^2\sigma^2$ to each imputation cell, it showed that at 0.01 level the intercepts of all 4 imputation cells are not significantly different from zero; i.e., the ratio model (equation (2.8)) is more appropriate than the regression model (equation (2.4)) for the data. However, the intercepts of cells 1 and 2 are significantly different from zero with probability 0.0311 and 0.0396, respectively. (For SIC 562 of February 1983's data, the intercepts of cells 1 and 2 are not significantly different from zero. However, the intercepts of cells 3 and 4 are significantly different from zero with probability less than 0.01 [see Table 2.4]).

(3) The log scaled linear model (2.5) was also fitted to the data in each of the 4 imputation cells. The histograms of the standardized residuals and the scatter plots of the residuals showed no gross deviations from the assumptions of the model. The imputed value of y based on the log scaled linear model (2.5) is

$$\hat{y}_i = \exp\left(\hat{\alpha} + \hat{\beta} \log x_i + \hat{\sigma}^2/2\right)$$

where $\hat{\sigma}^2$ is the residuals mean square error from the regression. Since our data came from a stratified sample, the inclusion probability (or the sampling weight) for each sampling unit varies considerably for units in the different strata, especially between the certainty stratum and the noncertainty strata. The mean sampling weights for all reported data is 20.407, the range is from 1 to 512.080. The regression analysis using t-test in Table 2.3 is the standard test assuming that the data come from a simple random sample and all the model assumptions are met.

DuMouchel and Duncan (1983) proposed to use the difference between the weighted and unweighted estimates (where the weights are sampling weights) as an aid in choosing the appropriate model and hence the appropriate estimator.

Let $\hat{\beta}$ and $\hat{\beta}_W$ be the ordinary least squares estimator, and the weighted least squares estimator of the regression coefficients of y on x respectively, where the weight is the sampling weight.

$$\hat{\beta} = (X'X)^{-1} X'Y$$

$$\hat{\beta}_W = (X'WX)^{-1} X'WY$$

where

$W = \text{diag}(w_i)$, w_i is the sampling weight of the i-th unit.

Let $\hat{\Delta} = \hat{\beta}_W - \hat{\beta}$

Under the usual linear regression model,

$$Y = X\beta + e \quad , \quad e|X \sim N(0, I\sigma^2)$$

DuMouchel and Duncan (1983) showed that

$$\hat{\Delta} = \hat{\beta}_W - \hat{\beta} = DY = [(X'WX)^{-1}X'W - (X'X)^{-1}X']Y$$

and

$$V(\hat{\Delta}) = V(\hat{\beta}_W) - V(\hat{\beta})$$

Since $V(\hat{\beta}_w) = (X'WX)^{-1} (X'W^2X) (X'WX)^{-1} \sigma^2 \neq (X'WX)^{-1} \sigma^2$, the standard errors and t statistics output by most weighted regression computer program are invalid. Since DX=0, $\hat{\Delta}$ is orthogonal to the columns of X. The sum of the squared residuals from the unweighted regression can be partitioned into a part due to Δ and to error. This leads to the following ANOVA table.

Formulas for ANOVA Table Comparing
Weighted and Unweighted Regressions

Source	degrees of freedom	Sum of Squares	Mean Square
Regression*	p	$SS_R = \hat{\beta}'(X'X)\hat{\beta}$	$MS_R = SS_R/p$
Weight	p	$SS_W = \hat{\Delta}' V_\Delta^{-1} \hat{\Delta}$	$MS_W = SS_W/p$
Error	n-2p	SS_E = remainder	$\hat{\sigma}^2 = SS_E/(n-2p)$
Total	n	Y'Y	

*The source labeled regression here includes the constant term if it is present in the model. In the example later, the effect of the grand mean is omitted and the degrees of freedom for regression is p-1, while SS_R and Y'Y are reduced by the square of the grand mean.

To get the above ANOVA table, DuMouchel and Duncan (1983) suggested two methods:

Method A. Perform the regression of Y on X and Z, where Z = WX, and then refit the regression, dropping the Z variables. The two "due to regression" sums of squares will be $SS_R + SS_W$, and SS_R, respectively.

Method B. Perform the regressions of Y and Z on X. Then perform the regression of Y on the residuals of the Z on X regressions. The last "due to regression" sum of squares will be SS_W.

Method A has been used for the above monthly reported data for the ratio model with variance proportional to x^2.

The alternative way to write the ratio model (2.8) for all units in the 4 imputation cells (labled as Model A) is

$$R \equiv y/x = \sum_{j=0}^{3} \alpha_j z_j + e \text{ , } e \sim n(0, \sigma^2) \tag{2.9}$$

where

$z_0 \equiv 1$

$z_1 = 1$, if the unit is in the imputation cell 1

$= 0$, otherwise;

$z_2 = 1$, if the unit is in the imputation cell 2

$= 0$, otherwise;

$z_3 = 1$, if the unit is in the imputation cell 3

$= 0$, otherwise.

To test whether there is any difference of the weighted and unweighted regression coefficients in the model, $E(\hat{\beta}_w - \hat{\beta}) = 0$, we use method A of DuMouchel and Duncan (1983). The ordinary regressions were used to regress R on z, and R on z and zw (where zw is the variable z multiplied by the sampling weights w). The ANOVA table is the following:

Table 2.a

ANOVA Table Comparing Weighted and Unweighted Regressions (Model A)

Source	Degrees of Freedom	sum of squares	Mean Square	F	Sig F
Regression	3	8.58874	2.86291	17.42277	0.0000
Weights	4	2.11691	0.529228	3.22072	0.0157
Error	1434	235.635	0.164320		
Total	1441	246.341			

There is a significant difference between $E(\hat{\beta}_w)$ and $E(\hat{\beta})$ at 0.0157 significant level. Hence there is a difference in using sampling weights in estimating the parameters in Model A.

In DuMouchel and Duncan's paper, it suggested to check for interactions among the variables when the hypothes is of $E(\hat{\beta}_w - \hat{\beta}) = 0$ is rejected. For retail monthly survey data, not many other variables can be added for use in the model. Other variables are current and previous month sales for 3 months or 12 months ago data, sampling weights, and number of establishments. I did some preliminary modeling (without using sampling weights) by adding some of those variables or the interactions of them. It seems that none of those variables contribute much to the model.

The alternative way to write the ratio model (2.7) (the current imputation model) for all units in the 4 imputation cells (labeled as Model B) is

$$R = y/x = \sum_{j=0}^{3} \alpha_j z_j + e, \quad e \sim n(0, \sigma^2/x) \qquad (2.10)$$

where z's are defined in (2.9).
The estimate of the mean rate (or trend) for each imputation cell under Models A and B with or without sampling weights (designated as w/wt, wo/wt, respectively) is given below.

Table 2.b

Cell Mean Rate (Trend) for Each Imputation Cell Under Different Models

Imputation Cell	Model A (ratio model with $V(\varepsilon) = x^2\sigma^2$)		Model B (ratio model with $V(\varepsilon) = x\sigma^2$	
	wo/wt	w/wt	wo/wt	w/wt
1 (GP2, Sales $\geq$ \$50,000)	1.61598	1.61117	1.53371	1.50170
2 (GP2, Sales < \$50,000)	1.68204	1.60483	1.66189	1.63338
3 (GP1, Sales $\geq$ \$50,000)	1.48462	1.47382	1.48338	1.41526*
4 (GP1, Sales < \$50,000)	1.49881	1.37942*	1.43179*	1.41607*

n = 1442, 6 outliers were deleted

In the current imputation procedure, the estimated mean rate from each imputation cell is checked to see whether it is in the prior limits of 1.443879 and 1.718664. If any of the estimates do not fall in the range, it will be recalculated using the appropriate collapsed cell.

It can be seen in Table 2.b that the estimates of all 4 cells from Model A wo/wt are in the desired range, while some others indicated by '*' are outside of the range. If the present range is a good prior, we'll conclude that Model A (wo/wt) is a good model for the data.

In the next section, a simulation study is conducted. The objective of the study is to evaluate the different imputation procedures.

III. Simulation Study

One way to compare different imputation procedures is to do a simulation study. The simulation study described below uses only full reported survey data as a complete data set, and simulates the missing values from the complete data set. Different imputation procedures are then applied on the simulated data set, the imputed values are then compared with the original values. Ford, Kleweno and Tortora (1980) did a simulation study using agriculture survey data, and Kalton (1981) did a simulation study using ISDP data.

The simulation study was conducted using the Monthly Retail Trade Survey data. The complete data set is the reported list sample of December 1982 retail sales of SIC 562, where the reported and imputed codes for both current and previous months are 1. There are 1445 units. A random mechanism is used to designate the missing values from the complete data set. (i.e., It is assumed that the data are missing at random.)

For each imputation cell, the establishment's current month sales are designated missing randomly according to the current nonresponse rate for the cell. (See Tables 3.3 and 3.4.) Five sets of missing data are generated.

From the previous study of the complete data set, it seems that for each imputation cell, the ratio model is a reasonable model and the model error variance is proportional to the square of the previous month sales. The current imputation procedure assumed a ratio model with model error variance proportional to the previous month sales. Models A and B defined in (2.9), (2.10) are used for the imputation comparisons.

Models A and B are defined for 4 imputation cells. It also is defined for the collapsed cells. Recall that the current imputation procedure will collapse within the group if the mean rate of any cell does not fall within the prior limits. Hence 6 imputation procedures are applied to the 5 simulated data.

1. Model A (4 cells) without using sampling weights in the estimation procedure [Model A (4 cells) wo/wt].

2. Model A (4 cells) incorporating the sampling weights in the estimation procedure [Model A (4 cells) w/wt].

3. Model B (4 cells) without using sampling weights in the estimation procedure [Model B (4 cells) wo/wt].

4. Model B (4 cells) incorporating the sampling weights in the estimation procedure [Model B (4 cells) w/wt].

5. Model B (2 cells) incorporating the sampling weights in the estimation procedure [Model B (2 cells) w/wt].

6. Current procedure: Use Model B (4 cells) w/wt to estimate the mean rate, if any of these rates falls off the range, the rates will be recalculated using the collapsed cells.

The criteria used to evaluate different imputation procedures are the following:

A. The mean deviation defined as $\sum (\hat{y}_i - y_i)/m$, where $\hat{y}_i$ is the imputed value and y_i is the actual value for unit i, (i = 1,2,...m), m is the number of missing values.

B. The mean absolute deviation defined as $\sum |\hat{y}_i - y_i|/m$.

C. The root mean square deviation defined as $\{\sum (\hat{y}_i - y_i)^2/m\}^{1/2}$.

D. The bias of the estimated totals due to imputation,

$\sum_{i=1}^{m} w_i(\hat{y}_i - y_i)$, where w_i is the sampling weight for the i^{th} unit.

E. The relative bias of the estimated totals due to imputation

$$\left(\sum_{i=1}^{m} w_i(\hat{y}_i - y_i) / \sum_{i=1}^{n} w_i y_i \right) \times 100, \text{ where } \sum_{i=1}^{n} w_i y_i \text{ is the estimated total}$$

from the complete data set.

The "errors" due to imputations of the above five types are calculated for each simulated data set, and the average of the five data sets is tabulated in Table 3.5

The mean deviation measures the bias in the imputed values. On the average, Model A over imputes the actual value, Model B slightly under estimates the actual value. The average over imputed value using Model A (4 cells) wo/wt is $16,529; when incorporating sampling weights in the estimation, it is $15,954. The average under imputed value using Model B without and with sampling weights are $1,356 and $4,501, respectively. The average under imputed value under current procedure is $2,390.

The mean absolute deviations and the root mean squares deviations measure the "closeness" of the imputed value ($\hat{y}_i$) with the true value y_i. On the average, Model A (4 cells) wo/wt has $2,255 larger mean absolute deviation than Model B (4 cells) wo/wt, and $2,610 larger mean absolute deviation than the current procedure. The current procedure has the smallest mean absolute deviation $56,975. Model B (4 cells) w/wt gives the smallest mean square deviation $186,324.

Since our sample is a stratified random sample, sampling units from different strata have different inclusion probabilities. To estimate total sales, the bias due to imputation is of most interest. The average bias of these five data sets is $2281 x 10^3 when using the current imputation procedure. The smallest bias is $1,969 x 10^3 by using Model B (4 cells) with sampling weights. Model A (4 cells) without sampling weights has the largest bias $31,659 x 10^3. Note that under the same model and the same number of cells, the bias of the estimated total is smaller by using sampling weights than not using sampling weights. This occurred for both Models A and B. In comparing Model B (w/wt) with 2 imputation cells and 4 imputation cells, the bias of the estimated total of 4 imputation cells is 40% less than 2 imputation cells.

The relative bias is 0.1394% for the current procedure, and 0.1203% for Model B (4 cells) with sampling weights, and 1.9344% for Model A (4 cells) without sampling weights.

For the current imputation procedure, all ratios of identicals of the five data sets for cells 3 and 4 exceeded the prior limits. Beside data set 1, the recalculated ratios of identicals from the other four data set for the collapsed cell of Group 1 still exceeded the prior limits. (See Table 3.6.)

Table 3.6 shows that the estimated ratios in the imputation cells within Group 1 are more likely outside of the prior limits. The validity of the prior limits apparently need to be studied further using long-term time series data.

IV. Summary and Recommendation

We have reviewed the imputation procedure of the Monthly Retail Trade Survey. The data of December 1982 retail sales were examined. We summarize the results as follows:

1. The current imputation procedure is a fairly simple procedure which assumes a ratio model (1.3) for the reported data in each imputation

cell. By examining the reported retail sales data of December 1982, the error variance of the model for each imputation cell for most selected SIC's is proportional to x^2 instead of x (where x is the previous month sales). It seems that the current definitions of the imputation cells need to be modified so that the data will conform with the assumed model. The sales sizes of some imputation cells need to be adjusted for some SIC's, e.g., SIC's 541, 572, 592, and 5813. For SIC 551, an alternative definition of imputation cell may be needed. There are only 21 reported list sample firms in Group 2, the 753 firms are in Group 1. (See Table 2.2A.) Finer firm size codes (2, 3, 4, 6) or regions may be used as an alternative to Groups 1 and 2. (See Table 2.2B.)

2. The simulation study using SIC 562 of December 1982's data (it is assumed that the data are missing at random) shows that the current imputation procedure gives lesser bias than the other imputation procedures studied in estimation of the population total. It also shows that 4 imputation cells give less bias than 2 collapsed cells; using sampling weights in the estimation gives less bias than not using sampling weights. It is suggested that finer imputation cells may be needed for each KB.

3. Field follow-up on the nonresponse data is necessary so we can better understand the nonresponse characteristics. This would show, for example, whether the distribution of nonresponse is the same as the distribution of response. Or whether the nonresponse rate systematically increases or decreases with sales size.

4. Revising or incorporating the prior limits used in the imputation

procedure. For the data we examined (December 1982 - SIC 562), two ratios of identicals are outside of the prior limits (see Table 2.b last column). The ratio will then be recalculated within a bigger cell and it will be used in the imputation whether the new ratio is in the prior limits or not. If the prior limits are good, it should be used in the imputation procedure when the ratio of identicals is outside of the limits. For example, using the closest bound of the limits to replace the ratio that is out of range. If the prior is out of date, it seems that it should be revised more often by using the existing ratios that have been calculated through the years.

5. The current imputation procedure is a mean imputation one (see Sedransk and Titterington (1980)), i.e., to impute for missing sales using a mean of the predictive distribution conditional on the known predictors. The mean imputation usually gives less variance of the total than the random imputation (where some error has been added to each predicted value). Since the objective of the Monthly Retail Trade Survey is to publish the total of the monthly sales, the mean imputation is used in the current imputation procedure. If furnishing the public use tape is also needed monthly, then in order to preserve the distribution of the monthly sales data the random imputation should be used, i.e., some 'error' should be added to the predicted value. These errors can take the form of random normal deviates defined in the model or randomly selected residuals from the model.

1. Reported and Imputed Code (R/I): Reported and imputed flags for current and previous month's data for each form and key.

   1 = reported - data accepted
   2 = imputed - no report
   3 = imputed - tolerance failure
   4 = reported - tolerance failure (data accepted)
   5 = prorated establishment data obtained from sampling unit total

2. Panel Code: Designates the months for which data are collected and tabulated.

   0 = canvassed monthly
   1 = Jan., April, July, October
   2 = Feb., May, August, Nov.
   3 = March, June, Sept., Dec.
   5 = canvassed monthly

3. Firm Size Code

   0 = Area sample nonemployer
   1 = Area sample employer
   2 = Group I (1-3 establishments)
   3 = Group I (4-10 establishments)
   4 = Group I (11 + establishments)
   6 = Group II (11 + establishments and certainty alpha)

Table (2.1) Crosstabulation by Panel and R/I Code
Monthly Retail Trade Survey of December 1982 - SIC 562

Reported & Imputed Code (previous month, current month)	Panel 0	2	3	5	Total
11	802	21	635	32	1490
12	68	0	1	3	72
21	32	0	1	1	34
22	661	0	87	5	753
23	1	0	0	0	1
33	0	0	1	0	1
41	2	0	3	0	5
44	0	0	4	0	4
55	577	0	0	0	577
TOTAL	2143	21	732	41	2937

Table 2.2 The Variance of Current Month Sales (y) and Mean Sales for Previous Month (x) of Each Sales Class For Imputation Cell 1

SIC 562 - December 1982

	N_i	$\bar{X}_i$	S_{y_i}	$S_{y_i}/\bar{X}_i$	$S^2_{y_i}/\bar{X}_i$
		$	$		
1	20	53589.70	23299.20	.43477	10129.80
2	20	61375.15	22784.82	.37124	8458.60
3	20	66902.35	31037.83	.46393	14399.34
4	20	75298.35	21491.90	.28542	6134.29
5	20	85678.90	31313.63	.36548	11444.40
6	20	98271.55	31651.67	.32208	10194.49
7	20	108040.35	28618.81	.26489	7580.84
8	20	119017.00	50504.87	.42435	21431.75
9	20	130416.30	45786.41	.35108	16074.64
10	20	148169.05	48830.86	.32956	16092.78
11	20	163565.05	50617.00	.30946	15663.99
12	20	187467.35	53693.16	.28641	15378.44
13	20	217480.45	61671.26	.28357	17488.21
14	20	253893.30	90767.27	.35750	32449.44
15	20	304998.35	97160.85	.31856	30951.74
16	20	364519.05	107030.16	.29362	31426.22
17	20	453640.35	160211.72	.35317	56581.82
18	20	628361.15	241197.89	.38385	92584.37
19	20	870042.34	309432.53	.35565	110050.32
20	20	1551478.09	718237.21	.46294	332498.86
21	11	3465163.91	1513233.64	.43670	660827.63

Table 2.2 A The Estimated λ , ρ for each Imputation Cell

	December 1982					February 1983		
	GP	Sales	n	λ	ρ	n	λ	ρ
SIC 562	(Women's Ready-to-Wear Stores)							
Imputation Cells								
1	2	> $ 50,000	411	0.0513	2.0709	402	0.141970	1.87255
2	2	< $ 50,000	249	8.4806	1.6594	265	0.002676	2.22700
3	1	> $ 50,000	354	0.0468	2.1115	276	0.003741	2.21092
4	1	< $ 50,000	431	1.8031	1.7804	612	0.140595	1.60352
Collapsed Cells								
1	2	(0, ∞)	660	1.12407	1.7873	667	0.038760	1.99410
2	1	(0, ∞)	785	0.52022	1.8271	888	0.152564	1.93086
Total			1,445					
SIC 521	(Building Materials Stores)							
Imputation Cells								
1	2	> $183,333	137	39,878.79	1.11455			
2	2	< $183,333	84	1.692573	1.81312			
3	1	> $183,333	192	0.023958	2.08377			
4	1	< $183,333	222	34,776,763	0.52032			
Collapsed Cells								
1	2	(0, ∞)	221	229.20	1.45410			
2	1	(0, ∞)	414	33.90	1.58289			
Total			635					
SIC 531	(Department Stores)							
Imputation Cells								
1	2	> $501,667	5,319	0.452028	1.86277			
2	2	< $501,667	1,757	531,359,832	0.48661			
3	1	> $501,667	148	0.020041	2.10604			
4	1	< $501,667	333	1,330.66	1.35004			
Collapsed Cells								
1	2	(0, ∞)	7,076	12.09	1.66243			
2	1	(0, ∞)	481	0.47482	1.90221			
Total			7,557					

Table 2.2 A The Estimated λ , ρ for each Imputation Cell
(continued)

	December 1982				
	GP	Sales	n	λ	ρ
SIC 541	(Grocery Stores)				
Imputation Cells					
1	2	> $146,667	1,462	0.001128	2.16946
2	2	< $146,667	149	45142.53	0.99589
3	1	> $146,667	506	1.098186	1.71815
4	1	< $146,667	311	0.007389	2.16929
Collapsed Cells					
1	2	(0, ∞)	1,611	5.89	1.62492
2	1	(0, ∞)	817	22.81	1.51683
Total			2,428		
SIC 551	(Motor Vehicle Dealers (Franchised))				
Imputation Cells					
1	2	> $375,000	17	NA	NA
2	2	< $375,000	4	NA	NA
3	1	> $375,000	596	0.008398	2.12081
4	1	< $375,000	157	131.90	1.54960
Collapsed Cells					
1	2	(0, ∞)	21	NA	NA
2	1	(0, ∞)	753	4.460942	1.74458
Total			774		
SIC 572	(Household Appliance Stores, Radio and Television Stores)				
Imputation Cells					
1	2	> $ 58,333	130	0.342361	1.93154
2	2	< $ 58,333	13	NA	NA
3	1	> $ 58,333	225	0.005206	2.25830
4	1	< $ 58,333	132	2.576959	1.88564

Table 2.2 A The Estimated λ , ρ for each Imputation Cell
(continued)

	December 1982				
	GP	Sales	n	λ	ρ
SIC 572 (con't)					
Collapsed Cells					
1	2	(0, ∞)	143	1682.44	1.35209
2	1	(0, ∞)	357	7.592493	1.75393
Total			500		
SIC 592	(Liquor Stores)				
Imputation Cells					
1	2	> $ 32,500	127	60.81	1.64058
2	2	< $ 32,500	31	NA	NA
3	1	> $ 32,500	253	0.002175	2.29045
4	1	< $ 32,500	131	0.814729	1.83754
Collapsed Cells					
1	2	(0, ∞)	158	1647.62	1.41048
2	1	(0, ∞)	384	0.041157	2.07049
Total			542		
SIC 5812	(Eating Places)				
Imputation Cells					
1	2	> $ 34,167	474	0.069592	1.97725
2	2	< $ 34,167	200	1813.1483	1.14356
3	1	> $ 34,167	539	36.789041	1.47826
4	1	< $ 34,167	318	0.572605	1.86140
Collapsed Cells					
1	2	(0, ∞)	674	0.142359	1.92301
2	1	(0, ∞)	857	315.8603	1.32191
Total			1531		

Table 2.2A The Estimated λ , ρ for each Imputation Cell
(continued)

	December 1982				
	GP	Sales	n	λ	ρ
SIC 5813	(Drinking Places)				
Imputation Cells					
1	2	> $ 7,500	48	1,428,935	0.74658
2	2	< $ 7,500	1	NA	NA
3	1	> $ 7,500	302	0.003442	2.20535
4	1	< $ 7,500	69	0.000005	2.91040
Collapsed Cells					
1	2	(0, ∞)	49	9,575,324	0.60170
2	1	(0, ∞)	371	0.001431	2.27896

Table 2.2B Frequency Table by Region
(Reported list sample with all nonzero sales stores)

December 1982

SIC Code	Northeast	Northcentral	South	North	Total
521	107	153	195	180	635
531	1464	2163	2544	1386	7557
541	498	638	786	506	2428
551	115	167	307	185	774
562	321	309	479	336	1445
572	103	135	156	106	500
592	103	109	193	137	542
5812	1531	278	381	560	312
5813	412	102	130	94	94

February 1983

SIC Code	Northeast	Northcentral	South	North	Total
562	331	408	525	291	1555

Table 2.3 Models Comparisons Using December 1982 Detailed Survey Data - SIC 562 (Outliers were removed)

	Regression Model ($y = \alpha + \beta x + \varepsilon$, $\varepsilon \sim n(0,v)$) (1) $v = x^2\sigma^2$			(2) $v = x\sigma^2$			Ratio Model ($y = \beta x + \varepsilon$, $\varepsilon \sim n(0,v)$) (3) $v = x^2\sigma^2$		(4) $v = x\sigma^2$		Log Model ($\log y = \alpha + \beta \log x + \varepsilon$, $\varepsilon \sim n(0,v)$) $v = \sigma^2$		
	n	$\hat{\alpha}$	$\hat{\beta}$	n	$\hat{\alpha}$	$\hat{\beta}$	n	$\hat{\beta}$	n	$\hat{\beta}$	n	α	β
Group 1	785	133.90791	1.49165**	782	922.84357	1.44890**	785	1.497					
t p		0.576 0.5647	78.324 0.0000		1.434 0.1521	87.664 0.0000							
Group 2	658	2002.8040**	1.59809**	651	5074.80864**	1.53727**	658	1.642					
t p		3.486 0.0005	83.060 0.0000		4.232 0.0000	134.838 0.0000							
Imputation Cell 1 Gp2, Sales ≥$50,000	411	6705.62*	1.56858**	411	16450.91**	1.49786**	411	1.617	411	1.54094	411	0.75225**	0.97589**
t p		2.163 0.0311	56.483 0.0000		4.156 0.0000	86.805 0.0000						5.691 0.000	90.444 0.000
Imputation Cell 2 Gp2, Sales <$50,000	247	2108.74*	1.58554**	247	1963.53	1.59218**	247	1.682	247	1.66189	247	0.90824*	0.95761**
t p		2.068 0.0396	29.115 0.0000		1.419 0.1570	28.556 0.0000						2.317 0.0213	24.790 0.0000
Imputation Cell 3 Gp1, Sales ≥$50,000	353	2049.13	1.45436**	353	195.05	1.48227**	353	1.485	353	1.48377	353	0.75347**	0.9652*
t p		1.277 0.2026	44.846 0.0000		0.070 0.9443	49.248 0.0000						3.571 0.0004	52.390 0.0000
Imputation Cell 4 Gp1, Sales <$50,000	432	-147.81	1.5092**	432	1664.25**	1.3756**	432	1.498	432	1.4307	429	0.58015**	0.97874**
t p		-0.991 0.3224	62.484 0.0000		2.875 0.0042	48.08 0.0000						3.365 0.0000	56.865 0.0008

Prior limits for December and SIC 562 are (1.443879 ≤ β ≤ 1.718664)

** significant level 0.01
* significant level 0.05

Table 2.4 Results of Regression Model
Using February 1983 Retailed Survey Data--SIC 562

$(y = \alpha + \beta x + \varepsilon \; , \; \varepsilon \sim n\;(o, x^2\;\sigma^2))$

Imputation Cells	n	α	β
Cell 1 (GP=2, sales ≥ $50,000)	402	893.05	1.00387**
t		0.656	70.992
p		0.5119	0.0000
Cell 2 (GP=2, sales< $50,000)	265	-1161.91	1.07922**
t		-1.167	22.421
p		0.2443	0.0000
Cell 3 (GP=1, sales ≥ $50,000)	276	16,216.53**	0.86146**
t		6.661	25.884
p		0.0000	0.0000
Cell 4 (GP=1, sales< $50,000)	612	882.702**	0.98123**
t		6.278	70.225
p		0.0000	0.0000

Prior limits for February and SIC 562 are [0.837801, 0.967831]
** significant level 0.01
* significant level 0.05

FIGURE 2.1 The Data Plot of the Current Month Sales vs. Previous Month Sales of SIC 562 of December 1982 for Four Imputation Cells-Reported List Sample Data

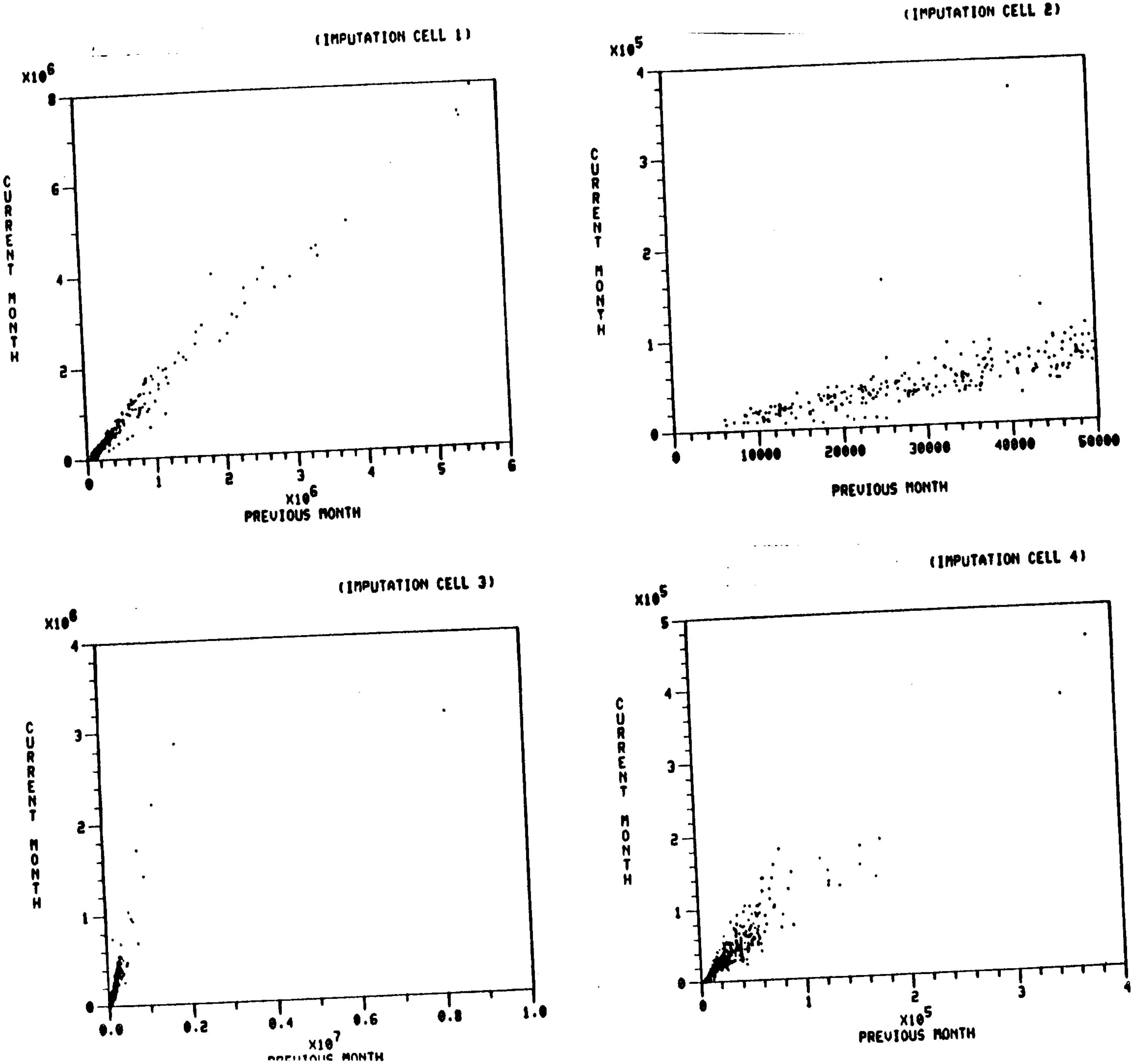

FIGURE 2.2 The Subdomain of the Plots of Figure 1

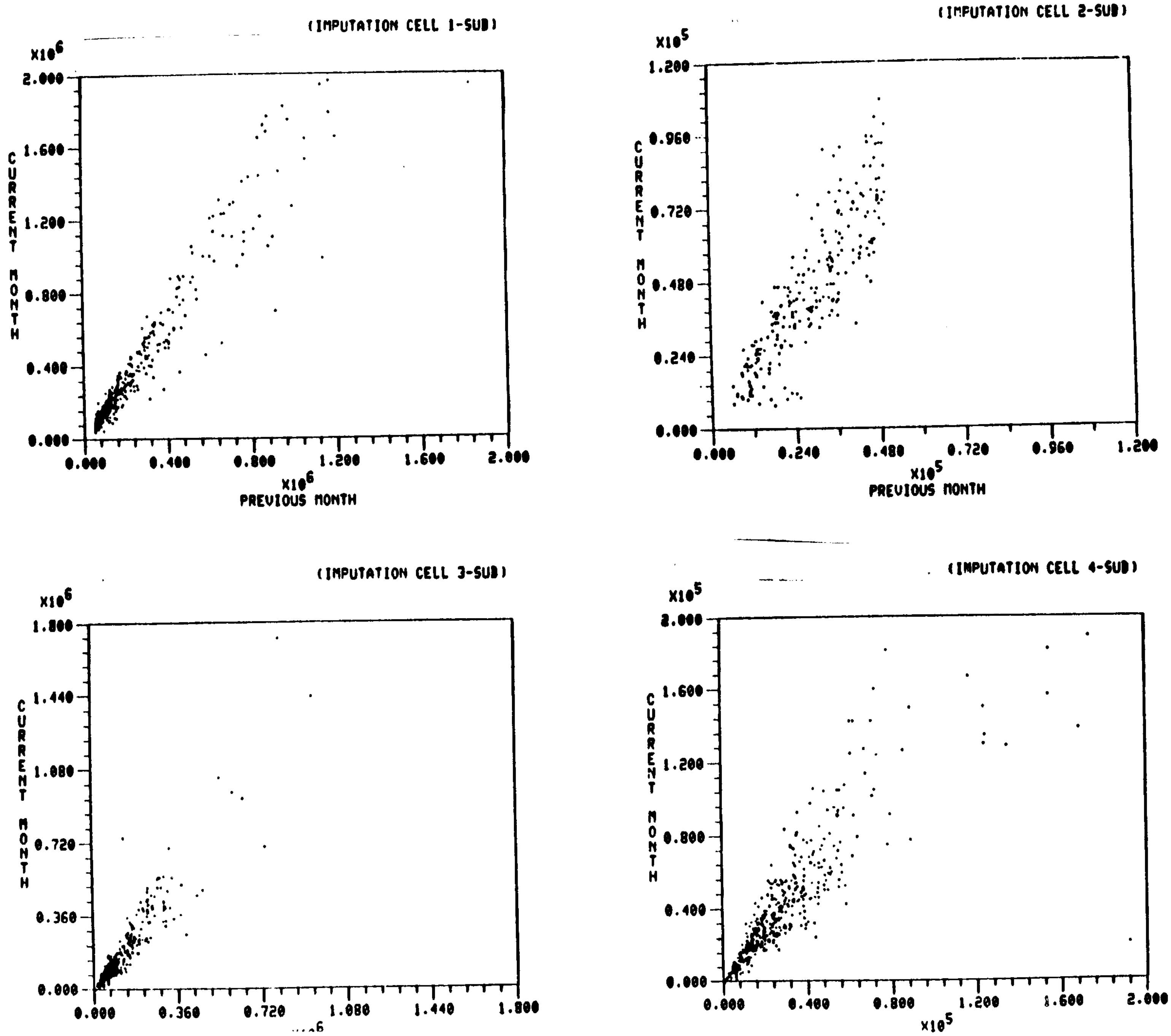

FIGURE 2.3 The Histogram of the Current Month Sales of SIC 562 of December 1982 for Four Imputation Cells-Reported List Sample Data

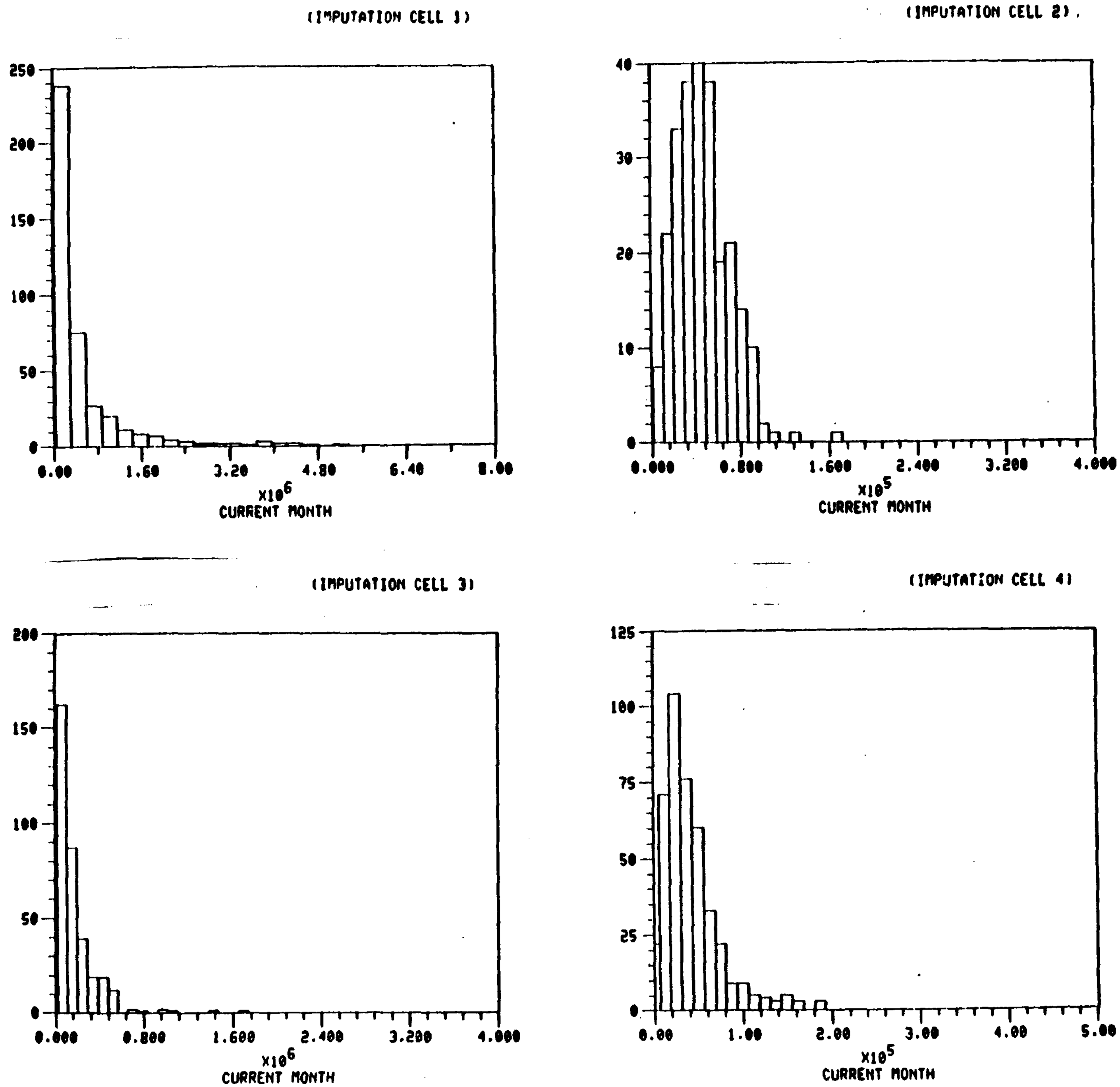

Table 3.1 Frequency Table: Nonresponse Pattern by Panel
December 1982 - SIC 562*

Panel	Nonresponse Pattern (RICM, RIPM)**									
	11	12	21	22	23	33	41	44	55	Total
0	800	66	12	539	1	0	2	0	576	1996
2	20	0	0	0	0	0	0	0	0	20
3	593	1	1	86	0	1	3	3	0	688
5	32	3	0	4	0	0	0	0	0	39
TOTAL	1445	70	13	629	1	1	5	3	576	2743

* List Sample data with positive sales for both current and previous months
** RICM: Report/impute code for current month
RIPM: Report/impute code for previous month

Table 3.2 Frequency Table: Nonresponse Pattern for Current Month
by Panel, December 1982 - SIC 562*

Panel	RICM					
	1	2	3	4	5	Total
0	814	605	1	0	576	1996
2	20	0	0	0	0	20
3	596	87	1	3	0	688
5	32	7	0	0	0	39
TOTAL	1463	699	2	3	576	2743

* List sample data with positive sales for both current and previous months

Table 3.3 Item Nonresponse Rate for Current Month Sales by Group and Sales Sizes December 1982 - SIC 562**

Imputation Cell	Group	Sales*	Total Units	%	Reported Units	Imputed Units	Imputed Rate %
1	2	≥ $50,000	836	38.58	415	421	50.36
2	2	< $50,000	417	19.24	259	158	37.89
3	1	≥ $50,000	416	19.20	357	59	14.18
4	1	< $50,000	498	22.98	435	63	12.65
			2167		1466	701	32.35

** List sample data with positive sales for both current and previous months; also units with R/I code 5 were excluded.

* For panel 0, previous month sales were used; for other panels, current month sales 3 months ago's data were used.

Table 3.4 The Pseudo Nonresponse Rate from the Simulated Data

Imputation Cell	Group	Sales*	Total Units	%	Units Deleted	%
1	2	≥ $50,000	410	28.83	206	50.24
2	2	< $50,000	250	17.30	95	38.00
3	1	≥ $50,000	354	24.50	50	14.12
4	1	< $50,000	431	29.83	55	12.76
			1445		406	28.10

Table 3.5 Summary of the Results of the Model Comparisons from the Average of the Five Data Sets

	Model A (4 cells) wo/wt	Model A (4 cells) w/wt	Model B (4 cells) wo/wt	Model B (4 cells) w/wt	Model B (2 cells) w/wt	Current Procedure
	$	$	$	$	$	$
Mean deviation	16.529×10^3	15.954×10^3	-1.356×10^3	-4.501×10^3	0.995×10^3	-2.390×10^3
Mean absolute deviation	59.585×10^3	59.892×10^3	57.330×10^3	57.347×10^3	56.994×10^3	56.975×10^3
Root mean square deviation	205.612×10^3	208.105×10^3	190.503×10^3	186.324×10^3	188.514×10^3	187.197×10^3
Bias of the estimated total	$31,659 \times 10^3$	$22,002 \times 10^3$	$6,316 \times 10^3$	$1,969 \times 10^3$	$3,247 \times 10^3$	$2,281 \times 10^3$
Relative bias of the estimated total (%)*	1.9344	1.3443	0.3860	0.1203	0.1984	0.1394

* The estimated total from the complete data is $\$1,636,659 \times 10^3$.

Table 3.6 The Mean Rate (Ratio of Identicals) for Each Imputation Cell Under Current Imputation Procedure for Five Simulated Data Sets

Imputation Cell	Group	Sales $ 10^3	Data 1		Data 2		Data 3		Data 4		Data 5		n_i
1	2	$\geq$ 50	1.48790	1.5457	1.46531		1.58687		1.49056		1.55891		205
2	2	< 50	1.76791*		1.70612		1.62313		1.61662		1.59659		154
3	1	$\geq$ 50	1.37457*	1.40009	1.42998*	1.42614*	1.40372*	1.41563*	1.40648*	1.42494*	1.38069*	1.40829*	304
4	1	< 50	1.41911*		1.42303*		1.42495*		1.43944*		1.43150*		376
TOTAL													1039

*The limits for mean rate (ratio of identical) for SIC 562 are [1.443879, 1.718664].

REFERENCES

1. Bethlehem, J. G. and Keller, W. J. (1983). 'Linear Models for Weighting Sample Survey Data,' contributed paper of 44th ISI meeting, pgs. 396-399.

2. Cassel, C.M., Särndal, C.E., and Wretman, J. H. (1979). 'Some Uses of Statistical Models in Connection with the Nonresponse Problem.' Symposium on Incomplete Data: Preliminary Proceedings, pgs. 188-215.

3. DuMouchel, W. H. and Duncan, G. J. (1983). 'Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples.' JASA 78, pgs. 535-543.

4. Ford, B. L., Kleweno, D. G., and Tortora, R. D. (1980). 'The effects of procedures which impute for missing items: a simulation study using an agricultural survey.' Proceedings of the Survey Research Methods Section of ASA, pgs. 251-256.

5. Isaki, C. T., Wolter, K. M., Sturdevant, T. R. and Monsour, N. J. (1976). 'Sample Redesign of the Census Bureau's Monthly Business Survey,' American Statistical Meeting, Proceedings of Business and Economic Statistics Section, pgs. 90-98.

6. Kalton, G. (1981). 'Compensating for Missing Survey Data' Institute for Social Research, The University of Michigan.

7. Little, R.J.A. and Rubin, D.B. (1982). 'Missing Data in Large Data Sets.' Paper for small conference on the Improvement of the Quality of Data Collected by Data Collection System, Tennessee.

8. Little, R. J. A. (1982). 'Models for Nonresponse in Sample Surveys.' Journal of the American Statistical Association, 77, pgs. 237-250.

9. Little, R. J. A. and Smith, P. J. (1983). 'Multivariate Edit and Imputation for Economic Data.' American Statistical Meeting, proceedings of Survey Research Methods Section, pgs. 518-522.

10. Little, R. J. A. and Samuhel, M. E. (1983). 'Alternative Models for CPS Income Imputation.' American Statistical Meeting, Proceedings of Survey Research Methods Section, pgs. 415-420.

11. Särndal, C. E. and Hui, T. K. (1981). 'Estimation for Nonresponse Situations: To What Extent Must We Rely On Models?' Current Topics in Survey Sampling. Academic Press Inc., pgs. 227-246.

12. Sedransk, J. and Titterington, D.M. (1980). 'Non-response in Sample Surveys.' Technical report, Bureau of the Census.

13. Wolter, K. M., Isaki, C. T., Sturdevant, T. R., and Monsour, N. J. (1976). 'Sample Selection and Estimation Aspects of the Census Bureau's Monthly Business Surveys.' American Statistical Meeting, Proceedings of Business and Economic Statistics Section, pgs. 99-109.