Survey Methodology, June 1986
Vol. 12, No. 1, pp. 73–83
Statistics Canada

Statistical Editing and Imputation for Periodic Business Surveys

M.A. HIDIROGLOU and J.-M. BERTHELOT[1]

ABSTRACT

For periodic business surveys which are conducted on a monthly, quarterly or annual basis, the data for responding units must be edited and the data for non-responding units must be imputed. This paper reports on methods which can be used for editing and imputing data. The editing is comprised of consistency and statistical edits. The imputation is done for both total non-response and partial non-response.

KEY WORDS: Periodic survey; Statistical editing; Total/partial non-response; Imputation.

1. INTRODUCTION

Data are routinely collected by large organizations such as Statistics Canada based on properly designed sample surveys. If such data are collected on a periodic basis from the same sampling unit, there are several possibilities which will occur with respect to the data consistency (quality) over a given time period. The sampling unit may report the data faithfully with no dramatic departure in continuity ("smoothness") as time progresses. The data may be reported faithfully, with questionable jumps between two time periods. The sampling unit may not report all the requested data items: this is known as partial non-response. The sampling unit may report data sporadically with breaks of total non-response for some periods. These can occur simultaneously in a periodic survey which collects required data from a large number of sampling units.

The problems which will be addressed in this article are the editing and imputation of data for sampling units that are contacted on a periodic basis by a surveying organization. The methods discussed are general for data of a multivariate nature composed of both quantitative and qualitative variables. The editing will include consistency and statistical edits.

For quantitative data, consistency edits ensure that linear combination of the data fields within a given time period satisfy given requirements. For qualitative data, consistency edits ensure that variables correspond to well defined values.

Statistical edits are used to isolate sampling units which may report some of their quantitative data fields in an inconsistent manner either from time period to time period or within a specific time period. Units with unusually high or low values will be termed "outliers". The identification of "outliers" is extremely important in an ongoing survey for two reasons. First, they influence statistics of the data set which may be for instance totals. This point has been studied by Hidiroglou and Srinath (1981). Second, the imputation of quantitative data for non-response units for periodic business surveys is usually based on trends or means: the removal of outlier units from the computation of these trends or means, will produce statistics that are not contaminated with there observations. For units which have partial non-response, data must be imputed for the missing fields.

For large data sets, where timely release of the summary information is crucial, the editing and the imputation of data should be automatic and computer handled given some well specified rules. This is in agreement with Gentleman and Wilk (1975), and Fellegi and Holt (1976).

[1] M.A. Hidiroglou and J.-M. Berthelot, Business Survey Methods Division, 11th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

2. EDITING PERIODIC DATA

2.0 Consistency Edits

For a given unit i and time period t, let $\underset{\sim}{x}_i(t)$ represent the vector of data which is to be collected. The vector $\underset{\sim}{x}_i(t)$ may be decomposed into a series of elementary vectors for which independent editing and imputation are required.

That is, $$\underset{\sim}{x}_i(t) = (\underset{\sim}{x}_i^{(1)}(t), \ldots, \underset{\sim}{x}_i^{(P)}(t))$$

where $$\underset{\sim}{x}_i^{(p)}(t) = (x_{il}^{(p)}(t), \ldots, x_{ik_p}^{(p)}(t))$$

for $$i=1, \ldots, n;\ p=1, \ldots, P;\ t=1, \ldots, T$$

and k_p is the number of variables in the p:*th* elementary vector.

For each elementary vector $\underset{\sim}{x}_i^{(p)}(t)$, the consistency edits may be represented as

$$\underset{\sim}{A}^{(p)}(\underset{\sim}{x}_i^{(p)}(t))' \leq (\underset{\sim}{c}^{(p)})'$$

where $\underset{\sim}{A}^{(p)}$ is a ℓ_p by k_p matrix representing the rules that the elements of the elementary vector $\underset{\sim}{x}_i^{(p)}(t)$ must obey, and $\underset{\sim}{c}^{(p)}$ is a 1 by ℓ_p vector which represents the constraints. This formulation allows one to define consistency edits for both qualitative and quantitative variables. For qualitative variables, the consistency edits could be used to check if the variables correspond to well-defined values. For quantitative variables, the consistency edits can check if certain variables are not larger (or smaller) than other variables or that a linear combination is equal to (or greater than or less than) a given variable.

2.1 Statistical Edits

Given that data are reported periodically, the problem is to isolate outlying observations within the time series. In the present context, an outlying observation i, will be defined as one whose trend for the current period to a previous period, for given variables of the element vector $\underset{\sim}{x}_i(t)$, differs significantly from the corresponding overall trend of other observations belonging to the same subset of the population. Statistical edits can also be applied within a time period, by comparing the ratios of two correlated variables amongst themselves, within a given subset of the population. In this article, the statistical edit will only be discussed in terms of the trend between time periods. Similar, somewhat imprecise but working definitions of outliers have also been given by other authors, for example:

GRUBBS (1969) says that "An outlying observation, or outlier, is one that appears to deviate markedly from the other members of the sample in which it occurs."

GUMBEL (1960) says: "The outliers are values which seem either too large or too small as compared to the rest of the observations."

KENDALL and BUCKLAND (1957, p. 209), write: "In a sample of n observations it is possible for a limited number to be so far separated in value from the remainder that they give rise to the question whether they are from a different population, or that the sampling technique is at fault. Such values are called outliers. Tests are available to ascertain whether they can be accepted as homogeneous with the rest of the sample."

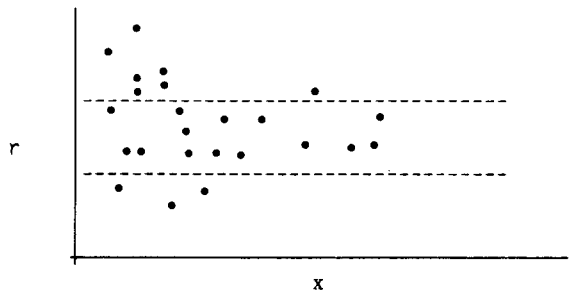
2.1.1 Review of Some Methods Currently Used

Methods for detecting outliers have been proposed by Dixon (1953), Grubbs (1969), Tietgen and Moore (1972), and Prescott (1978) to mention a few. Most of the test procedures for outlier detection proposed by these authors consider the problem as one of hypothesis testing. In the simplest cases, the null hypothesis is that the sample comes from a normal distribution with unspecified mean and variance, while the alternative hypothesis is that one or more of the observations come from a different distribution. Percentage points of a test statistic may be determined under the null hypothesis and compared with computed values of the test statistic in particular applications. Applying these methods to periodic data from large surveys presents problems for the following reasons. First, the assumption of normality of trends from one period to another may not hold. Second, these traditional methods require the existence of tables for determining critical values which define rejection regions. The method which we will propose in Section 2.1.2 does not have the above mentioned disadvantages. It can be easily implemented on the computer, does not require the assumption of normality, and does not make use of tables.

In our specific context, and given elements of the vectors $\underset{\sim}{x}_i(t)$ and $\underset{\sim}{x}_i(t+1)$, denote as $x_i(t)$ and $x_i(t+1)$ the responses for two consecutive periods for a given unit, where $i=1, \ldots, n$. Denote as r_i the ratio of current period data to previous period data. One method which is known as the range edit, is to simply define fixed upper and lower bounds based on experience for comparison purposes. Ratios found outside these bounds are declared as outliers. A major drawback with this method is that the definition of outlier is too subjective and does not make use of the distribution of the ratios.

A method that attempts to make use of the distribution of the ratios is the Chebychev inequality edit. This edit is constructed by computing the lower bound as $\bar{r} - ks_r$ and the upper bound as $\bar{r} + ks_r$ where $\bar{r} = \Sigma_{i=1}^{n} r_i/n$ and $s_r^2 = \Sigma_{i=1}^{n} (r_i - \bar{r})^2/(n-1)$. This edit has two main drawbacks. First, the choice of k is subjective and can result in having an edit that cannot detect any outliers. This last point has been demonstrated by Wilkinson (1982). Second, "large" outliers may hide "smaller" outliers. This effect is known as the masking effect.

An improvement to this method has been the use of quartiles and interquartile distances rather than the use of mean and standard error to come up with the upper and lower bounds. In this case, the edit is constructed by computing the lower bound as $r_M - k\,D_{r_{Q1}}$ and the up per bound as $r_M + k\,D_{r_{Q3}}$ where r_M is the median of the ratios, $D_{r_{Q1}}$ is the distance between the first quartile and the median, and $D_{r_{Q3}}$ is the distance between the third quartile and the median. Since the quartiles are not affected by the tails of the distribution, it greatly alleviates the masking effect problem. However, this method has two drawbacks. First, in some very specific circumstances, it is possible that the outliers on the left tail of the distribution are undetectable. Second this method does not take into account the fact that in most of the periodic business surveys, the variability of ratios for small businesses is larger than the variability of ratios for large businesses (Sugavanam 1983). This fact is expressed by the following graph:



This drawback has the effect of identifying too many small units as outliers and not enough large units. This effect will be referred to as the "size masking effect".

2.1.2 Proposed Procedure

For two occasions t and $t + 1$, the overall trend for the data pair given by

$$(x_i(t),\ x_i(t+1)),\ i = 1,\ \ldots,\ n$$

is

$$R = \sum_{i=1}^{n} x_i(t+1) \Big/ \sum_{i=1}^{n} x_i(t).$$

Now, R may be expressed as

$$R = \sum_{i=1}^{n} I_i\, r_i$$

where

$$I_i = x_i(t) \Big/ \sum_{i=1}^{n} x_i(t)$$

and

$$r_i = x_i(t+1)/x_i(t).$$

I_i is a measure of the relative importance of the *i:th* unit amongst the n units at time t. The individual trends r_i must be transformed in order to ensure that outliers are detected at both tails of the distribution. This transformation is:

$$s_i = \begin{cases} 1 - r_M/r_i, & \text{if } 0 < r_i < r_M \\ r_i/r_M - 1, & \text{if } r_i \geq r_M \end{cases}$$

where r_M is the median of the ratios.

In order to bring in the magnitute of the data, the following transformation is required (Berthelot 1983):

$$E_i = s_i \{\text{Max } (x_i(t),\ x_i(t+1))\}^U$$

where $0 \leq U \leq 1$. The E_i's will be referred to as effects and the exponent U in the transformation provides a control on the importance associated with the magnitude of the data. This transformation allows us to place more importance on a small change associated with a "large" unit as opposed to a large change associated with a "small" unit. The values of the median and quartiles as used by Sande (1981) will be applied to the transformed, E_i's, in order to detect potential outliers. Denoting as E_{Q1}, E_M and E_{Q3} as the first quartile, the median and the third quartile respectively, define the following two deviations:

$$d_{Q1} = \text{Max } (E_M - E_{Q1},\ |AE_M|),$$

$$d_{Q3} = \text{Max } (E_{Q3} - E_M,\ |AE_M|).$$

Outliers will be defined as all those units whose associated effect E_i lies outside the interval $(E_M - Cd_{Q1}, E_M + Cd_{Q3})$. The purpose of the AE_M term is to avoid difficulties which arise when $E_M - E_{Q1}$ or $E_{Q3} - E_M$ are very small. That is, the problem which may arise when the effects E_i are clustered around a single value with one or two modest deviations may produce false outliers. The parameter C controls the width of the acceptance interval. The parameter U controls the shape of the curve defining upper and lower boundaries. The effect of increasing U is to attach more importance with fluctuations associated with the larger observations. A value of 0.05 is suggested for A as it has proved to be adequate in practice.

2.1.3 Treatment For Outliers

Once units have been identified as possible outliers, they are flagged as such and brought to the attention of the survey takers. A decision must then be taken on how these abnormal observations are treated. Their existence may have arisen as a result of several factors. These factors include measurement error, incorrect interpretation of the questionnaire by the responding unit, or intrinsic variability of the population being surveyed. For units which have measurement error due to incorrect transcription of the data or incorrect responses, a simple follow-up will clear up the majority of these errors. For units which display intrinsic variability as a result of rapid growth, the reported values are correct but dominate too much the resulting summary tables. For those units, techniques, which reduce the sampling weight as suggested by Hidiroglou and Srinath (1981) or change the values themselves as suggested by Ernst (1980), must be used in order to accomodate (minimize) the effect of outlying observations. For units having unrepresentative data which cannot be verified, their data must be substituted with other data based on imputation techniques. The different kinds of corrective actions taken on outlying units must be flagged as well.

3. IMPUTING PERIODIC DATA

The information collected by periodic business surveys, such as sales and employment are collected via samples using mail questionnaires or telephone interviews. Non-responding units are followed up as much as possible within allotted budgets in order to improve the response rates. The follow-up is usually done by mail in the case of the smaller to medium sizes non-responding companies and by telephone for the larger or dominating companies. Although following up delinquent companies improves response rates for a given reference period, there will be nevertheless, a group of non-responding companies which may be classified into either hard-core or late respondents. Hard-core non-respondents are units which require a great deal of persuation to respond, if at all. Late respondents are units which respond late with respect to the survey's reference period either because they do not mail back their questionnaire on time or because they need to be prompted by a follow-up questionnaire. The non-responding units must therefore be imputed in order to make up for their contribution to the particular estimator being used by the survey. In the case of Monthly Business Surveys, such as the Monthly Retail Trade Survey, totals (e.g., sales) are being estimated. Imputation procedures can also be used to generate values for units declared as outliers. These imputed values can be used in lieu of these outlying observations, if no valid explanation can be provided for their presence.

The units with no response whatsoever, will be termed as total non-respondents and those with some, but not all, required data items, will be termed partial non-respondents. Desirable features of an imputation system should include the following properties (Berthelot and Hidiroglou 1982):

- it must automatically determine the most reasonable imputation procedure possible under the existing circumstances,
- the imputation cell, the level at which the computation of trends and means (medians) is performed, will usually correspond to the finest level of stratification of the sample,
- a minimum number of units must participate in the computation of trends or means (medians), otherwise, the imputation cells are automatically collapsed (using a pre-determined pattern), until the minimum requirement has been satisfied,
- it will recognize through the use of status codes that there are units which must not be imputed. These include seasonal units during the period that they are not operating, units temporarily out of business, or units which are no longer active,
- births which have no previous business history will have their data imputed using the means (medians) of similar responding births,
- units will be re-imputed for a number of periods previous to the current period: this is done in order to improve the strength of the imputations if the previous periods have been updated with data,
- backward imputations will be applied to units which have been continuously imputed using a forward imputation procedure as soon as a good response is obtained for a given period,
- imputation status codes will be associated with imputed units in order to provide a history of the procedure used for imputation,
- the ranking for imputing non-responding units is as follows: trends (monthly, quarterly, annual), means (medians) with the most recent trends being given priority. For instance, in the case of a monthly system, monthly trends are used for units which have data (response or imputed) in the month prior to the one to be imputed. Annual trends are used mostly for units which are seasonal and which fail to provide a response as they emerge from their out of season period and for which a last year value existed for the month to be imputed. Imputations based on the trends are obtained by multiplying the trends by the unit's last month or last year value. In the event that trends cannot be applied, the mean (median) of the cell is used as an imputation.

In order to formalize the preceding paragraphs in a mathematical fashion, let the number of units which are expected to respond for a given cell and given month be n. Let the number of non-respondents with total non-response be n_3, the number of respondents with total response be n_1 and the number of respondents with partial response be n_2. It is assumed that the sample design is stratified with the sampling being simple random without replacement. Let the size for the follow-up sample of the non-respondents be m_3 ($2 \leq m_3 \leq n_3$, with m_3 having been selected from n_3 according to a randomized mechanism). Note that $n_4 = n - \Sigma_{i=1}^{3} n_i$ units are not expected to provide any response to the survey process for a number of possible reasons. At a time t, they may be out of season, inactive, dead, or out of scope to the survey. For these units, the system will automatically associate zero values for all relevant fields in the given period.

The imputation process will then be done in several different ways according to the type of non-response.

3.0 Total Non-Response

The imputation process for the total non-respondents will first be discussed. Bearing in mind that either the whole vector $\underline{x}_i(t)$ or that some of its elementary vectors as given in

Section 2.0 must be totally imputed, denote as $(x_{i1}(t), \ldots, x_{ip}(t))$ one of the elementary vector within $\underset{\sim}{x}_i(t)$ where the editing and imputation process is independent from other elementary vectors within $\underset{\sim}{x}_i(t)$. Assuming that

$$x_{ip}(t) \geq \sum_{j=1}^{p-1} x_{ij}(t),$$

(which implies that the sum of the first $p-1$ data elements of the elementary vectors are smaller than the *p:th* datum element, the total) $x_{ip}(t)$ will first be imputed as

$$I_{ip}^{(1)}(t) = \sum_{k=1}^{6} [z_{ip}^{(k)}(t)\ \delta_i^{(k)}]$$

where $\delta_i^{(k)}$ refers to the procedure used for imputation and $z_{ip}^{(k)}$is the associated imputed value. One of the six $\delta_i^{(k)}$ values will be one and the other five must be zero ($\Sigma_{k=1}^{6}$ $\delta_i^{(k)} = 1$). The imputed $z_{ip}^{(k)}(t)$ values will be as follows:

$$z_{ip}^{(1)}(t) = [\sum_{r \epsilon s_1} w_r\, x_{rp}(t) / \sum_{r \epsilon s_1} w_r\, x_{rp}(t-1)]\ x_{ip}(t-1),$$

$$z_{ip}^{(2)}(t) = [\sum_{r \epsilon s_2} w_r\, x_{rp}(t) / \sum_{r \epsilon s_2} w_r\, x_{rp}(t-Q)]\ x_{ip}(t-Q),$$

$$z_{ip}^{(3)}(t) = [\sum_{r \epsilon s_3} w_r\, x_{rp}(t) / \sum_{r \epsilon s_3} w_r\, x_{rp}(t-1)]\ x_{ip}(t-1),$$

$$z_{ip}^{(4)}(t) = [\sum_{r \epsilon s_4} w_r\, x_{rp}(t) / \sum_{r \epsilon s_4} w_r\, x_{rp}(t-Q)]\ x_{ip}(t-Q),$$

$$z_{ip}^{(5)}(t) = [\sum_{r \epsilon s_5} w_r\, x_{rp}(t) / \sum_{r \epsilon s_5} w_r],$$

$$z_{ip}^{(6)}(t) = [\sum_{r \epsilon s_6} w_r\, x_{rp}(t) / \sum_{r \epsilon s_6} w_r],$$

w_r = inverse selection probability of unit r for the given cell. The subsets s_i $(i=1, \ldots, 6)$, will be determined by selecting the units which have provided a response for the *p:th* variable at time t and which have passed the edits. The conditions for each subset is

s_1 = all units which have provided edited responses between times t and $t-1$,

s_2 = all units which have provided edited responses between times t and $t-Q$,

s_3 = units in the follow-up subsample which have provided edited responses between times t and $t-1$,

s_4 = units in the follow-up subsample which have provided edited responses between times t and $t-Q$,

s_5 = all units which have provided edited responses at time t,

s_6 = units in the follow-up subsample which have provided edited responses at time t.

The choice of the imputation procedure will be governed by the following considerations.

(i) Procedures 1 (or 2) will be used if there is a response or imputed value at time $t-1$ (or $t-Q$) and that it is believed that the trends for the non-respondents is the same as the one for the respondents, within the given cell,

(ii) Procedures 3 (or 4) will be used if there is a response or imputed value at time $t-1$ (or $t-Q$) and that it is believed that the trends for the non-respondents differs from the one for the respondents within the given cell.

(iii) Procedure 5 will be used if there is no response at either times $t-1$ or $t-Q$ and that is believed that the mean of the non-respondents is equal to the mean of the respondents within the given cell,

(iv) Finally, procedure 6 will be used if there is no response at either times $t-1$ or $t-Q$ and that it is believed that the means of the respondents and non-respondents are different.

The choices between the different procedures can be made using decision tables which determine the conditions and, given the condition, choose the best imputation procedure according to pre-determined rules. Once that $x_{ip}(t)$ has been imputed for an elementary vector, its remaining components can be imputed using the procedures for partial non-response.

3.1 Partial Non-Response

For an elementary vector $(x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))$ which is part of $\underset{\sim}{x}_i(t)$, let δ_{ij} be the indicator variable which is equal to 1 if $x_{ij}(t)$ is present and zero otherwise at time t. Some additional notation is introduced at this point in order to ease the development. To this end, define

$$s_{i,R}(t-1) = \sum_{j=1}^{p-1} \delta_{ij}\, x_{ij}(t-1)$$

= the sum of responses at time $t-1$, for which there is a response at time t

$$s_{i,NR}(t-1) = \sum_{j=1}^{p-1} (1-\delta_{ij})\, x_{ij}(t-1)$$

= the sum of responses at time $t-1$, for which there is no response at time t,

$$s_{i,R}(t) = \sum_{j=1}^{p-1} \delta_{ij}\, x_{ij}(t).$$

The partial imputation will be based on the assumptions that $x_{ip}(t) \geq \Sigma_{j=1}^{p-1} x_{ij}(t)$ and that the distribution of the elements within $\underset{\sim}{x}_i(t)$ is similar to the distribution of the elements within $\underset{\sim}{x}_i(t-1)$. Two separate cases will be discussed.

Case 1: Parts of the elementary vector missing and $x_{ip}(t)$ present

Two subcases are possible: $x_{ip}(t) = \sum_{j=1}^{p-1} x_{ij}(t)$ or $x_{ip}(t) > \sum_{j=1}^{p-1} x_{ij}(t)$.

(i) $x_{ip}(t) = \sum_{j=1}^{p-1} x_{ij}(t)$

If all the elements of $\underset{\sim}{x}_i(t)$ excluding $x_{ip}(t)$ are missing, that is $\Sigma_{j=1}^{p-1} \delta_{ij} = 0$, then we must have that $s_{i,NR}(t) = x_{ip}(t)$. If some of the elements of $\underset{\sim}{x}_i(t)$ excluding $x_{ip}(t)$ are missing, that is $\Sigma_{j=1}^{p-1} \delta_{ij} > 0$, then $s_{i,NR}(t) = x_{ip}(t) - s_{i,R}(t)$.

(ii) $x_{ip}(t) > \sum_{j=1}^{p-1} x_{ij}(t)$

If all the elements of $\underset{\sim}{x}_i(t)$ *excluding* $x_{ip}(t)$ are missing, then $s_{i,NR}(t) = s_{i,NR}(t-1)$ $x_{ip}(t)/x_{ip}(t-1)$. If some of the elements of $\underset{\sim}{x}_i(t)$ excluding $x_{ip}(t)$ are missing, the choice of $s_{i,NR}(t)$ is not so obvious. In any event, one must have that $s_{i,R}(t) + s_{i,NR}(t) < x_{ip}(t)$. To this end, four separate possible imputations for $s_{i,NR}(t)$ will be given in order of preference.

(a) $s_{i,NR}(t) = [s_{i,NR}(t-1) + s_{i,R}(t-1)]\, x_{ip}(t)/x_{ip}(t-1) - s_{i,R}(t)$ provided that $s_{i,NR}(t) \geq 0$. Note that the condition $x_{ip} > \Sigma_{j=1}^{p-1} x_{ij}(t)$ is met if $s_{i,NR}(t) \geq 0$.

(b) $s_{i,NR}(t) = s_{i,NR}(t-1)\, [s_{i,R}(t)/s_{i,R}(t-1)]$

(c) $s_{i,NR}(t) = s_{i,NR}(t-1)\, [x_{ip}(t)/x_{ip}(t-1)]$

(d) $s_{i,NR}(t) = x_{ip}(t) - s_{i,R}(t)$.

The preferred imputation will be the first one that does not violate the inequality condition. For all the above cases, the imputed (actual values) will then be

$$I_{ij}^{(2)}(t) = (1-\delta_{ij})\, [s_{i,NR}(t)/s_{i,NR}(t-1)]\, x_{ij}(t-1) + \delta_{ij}\, x_{ij}(t);\ j=1, \ldots, p-1$$

Case 2: Parts of the elementary vector missing and $x_{ip}(t)$ is missing

As in case 1, two subcases are possible:

(i) $x_{ip}(t) = \sum_{j=1}^{p-1} x_{ij}(t)$

If $\Sigma_{j=1}^{p-1} \delta_{ij} = 0$, then $s_{i,NR}(t) = I_{ip}^{(1)}(t)$ where $I_{ip}^{(1)}(t)$ has been obtained using the imputation for total non-response. The imputation $I_{ij}^{(2)}(t)$ is then used. If $\Sigma_{j=1}^{p-1} \delta_{ij} > 0$, $I_{ij}^{(2)}(t)$ will be used provided that $s_{i,NR}(t) = I_{ip}^{(1)}(t) - s_{i,R}(t) \geq 0$. Otherwise, the following imputation must be used

$$I_{ij}^{(3)}(t) = (1-\delta_{ij})\, [s_{i,NR}(t)/s_{i,NR}(t-1)]\, x_{ij}(t-1) + \delta_{ij}\, x_{ij}(t);\ j=1, \ldots, p-1$$

and $I_{ip}^{(1)}(t)$ is replaced by $I_{ip}^{(3)}(t) = \Sigma_{j=1}^{p-1} I_{ip}^{(3)}(t)$

(ii) $x_{ip}(t) > \Sigma_{j=1}^{p-1} x_{ij}(t)$

For this case, the $x_{ip}(t)$ in case 1(ii) is replaced by $I_{ip}^{(1)}(t)$ and the methods given for this case are used, provided that the above inequality condition is satisfied. If the condition cannot be met, $I_{ip}^{(3)}(t)$ must be used and $I_{ip}^{(1)}(t)$ is replaced by $I_{ip}^{(3)}(t) = \Sigma_{j=1}^{p-1} I_{ip}^{(3)}(t)$.

If the assumption, that the distributions of the data elements of vectors $\underline{x}_i(t)$ and $\underline{x}_i(t-1)$ is similar, does not hold, then each individual element must be imputed using procedures for imputation for total non-response. These imputations must then be adjusted in order to satisfy the inequality requirement $x_{ip} \geq \Sigma_{j=1}^{p-1} x_{ij}$. Hence, for example, for case 1(i), we would have for $\Sigma_{j=1}^{p-1} \delta_{ij} = 0$,

$$I_{ij}^{(4)}(t) = [x_{ip}(t) / \sum_{j=1}^{p-1} I_{ij}^{(1)}(t)] \, I_{ij}^{(1)}(t)$$

and for $\Sigma_{j-1}^{p-1} \delta_{ij} > 0$

$$I_{ij}^{(4)}(t) = (1-\delta_{ij}) \left[\frac{x_{ip}(t) - \Sigma_{j=1}^{p-1} \delta_{ij} \, x_{ij}(t)}{\Sigma_{j=1}^{p-1} (1-\delta_{ij}) \, I_{ij}^{(1)}(t)} \right] + \delta_{ij} \, x_{ij}(t); \; j = 1, \ldots, p-1.$$

Similarly, cases 1(ii) and 2, could be developed using the imputed values $I_{ij}^{(1)}(t)$.

4. CONCLUSION

For periodic business surveys, it is important to have computer systems which can quickly and accurately monitor the flow of in-coming data in terms of its quality. Conversely, for expected data that are not coming in, the system should impute as well as possible for the non-response given some well specified rules.

The editing will cause the flagging of records in possible error. These errors can be termed as critical and non-critical. All errors should be corrected by either reviewing the questionnaires or checking their authenticity with the respondent. If this is not possible on account of time or budgetary constraints, the most critical errors must be corrected. Given that the errors have been taken care of, the next step of the processing is to impute for the non-respondents. Diagnostic summaries of the actions (edits or imputations) taken by the system, should be printed out in order to inform the survey analyst on the status of his data.

REFERENCES

BERTHELOT, J.-M., and HIDIROGLOU, M.A. (1982). Specifications for imputations in the retail trade survey. Technical report, Statistics Canada.

BERTHELOT, J.-M. (1983). Wholesale-retail redesign, statistical edit proposal. Technical Report, Statistics Canada.

DIXON, W.G. (1953). Processing data for outliers. *Biometrics*, 9, 74-89.

ERNST, L.R. (1980). Comparison of estimators of the mean which adjust for large observations. *Sankhya*, 42, 1-16.

FELLEGI, I.P., and HOLT, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

GENTLEMAN, J.F., and WILK, M.B. (1975). Detecting outliers, II. Supplementing the direct analysis of residuals. *Biometrics*, 31, 387-410.

GRUBBS, F.E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11, 1-21.

GUMBEL, E.J. (1960). Discussion on "Rejection of outliers" by Anscombe, F.J. *Technometrics*, 2, 165-166.

HIDIROGLOU, M.A., and SRINATH, K.P. (1981). Some estimators of population totals form a simple random sample containing large units. *Journal of the American Statistical Association*, 76, 690-695.

KENDALL, M.G., and BUCKLAND, W.R. (1957). *A Dictionary of Statistical Terms*. New York: Hafner.

PRESCOTT, P. (1978). Examination of the behaviour of tests for outliers when more than one outlier is present. *Applied Statistics*, 27, 10-25.

SUGAVANAM, R. (1983). A statistical edit for change. Technical Report, Statistics Canada.

SANDE, I.G. (1981). Estimation in the revised ISPI. Technical Report, Statistics Canada.

TIETGEN, G.L., and MOORE, R.H. (1972). Some Grubbs - type statistics for the detection of several outliers. *Technometrics*, 55, 583-598.

WILKINSON, R.G. (1982). An outlier identification technique designed for the Business Finance Annual Survey. Technical Report, Statistics Canada.