

9th National Convention on Statistics (NCS)
EDSA Shangri-La Hotel
October 4-5, 2004

**Treatment of Missing Data in the
Annual Survey of Establishments**
by:
Lourdes V. Homecillo

For additional information, please contact:

Author's name:	Lourdes V. Homecillo
Designation:	Statistician V
Agency:	National Statistics Office
Address:	Sta. Mesa, Manila

Treatment of Missing Data in the Annual Survey of Establishments¹

by:
Lourdes V. Homecillo

I. Introduction

The case of non-response, or any type of missing data for that matter, is a common occurrence in any survey. Managing non-responses then becomes necessary and this comes in two stages: prevention and treatment.

During the prevention stage, all strategies and efforts to achieve a high response rate and quality data are put into action such as proper training, persistent follow-ups thru personal visits and phone calls, sending of reminder letters, including the gathering of data from secondary sources, aggressive publicity campaign, and editing, among others.

The treatment stage comes when all efforts have been exhausted to produce a complete data set and yet non-responses still occur. The need to treat non-responses arises because the statistician **has** to tackle the problem on how estimation shall be done for samples with missing data.

This paper focuses on the latter stage and discusses the treatment process applied to missing data in the annual survey of establishments.

Missing data in the annual survey of establishments are primarily due to non-reporting by some sample establishments. Non-responses in the annual survey can be classified, as follows: unit non-response, item non-response, and inconsistent or unusable reports. **Unit non-responses** are those samples whose reports have not been collected because of refusals (but in operation) and those which have closed/moved out. **Item non-response** of responding samples refers to particular data items for which required entries are not reported. **Inconsistent or unusable reports** are those for which information are collected but are not usable (e.g. out-of-scope units or stratum jumpers), including those reported entries in an item which are unacceptable, invalid or appear inconsistent.

II. Treatment Process

The treatment of these missing data in the annual survey of establishments vary by stratum and the traditional practices currently adopted are the following:

- i. **Weight adjustment** – This method is applicable only for unit non-responses in the non-certainty stratum. In this method, the weights ($w=N/n=$ *inverse of probability of selection*) of the responding establishments in the stratum are increased in the estimation process by an adjustment factor equivalent to n/n' . Thus, the adjusted weight, w' ,

¹ Paper presented by Lourdes V. Homecillo, Statistician V, National Statistics Office, Manila during the 9th National Convention on Statistics, October 2004

compensates for the non-sampled units and the unit non-respondents in the stratum.

The adjusted weight is shown as: $w' = N/n * n/n' = N/n'$.

where: N = total number of establishments in the stratum
n = number of samples in the stratum
n' = number of responding samples in the stratum.

- ii. **Imputation** – Imputation, as defined, is replacing missing values with ‘credible’ data from a donor. In the annual survey, imputation is done for unit non-response and item non-response.

For unit non-response, imputation involves the preparation of a dummy questionnaire. The extent of the application of this method varies by stratum. In the non-certainty stratum, particularly the lowest employment size of ATE less than 20, imputation is generally done for all non-responding samples, especially those that fall under the ‘minimum of 3’ rule.

For larger employment sizes like ATE 20-99 and 100-199, it would be desirable to impute all, or majority of non-responding samples to be able to derive unbiased estimates of levels. Weight adjustment is also applied in the estimation process if not all non-responses are imputed.

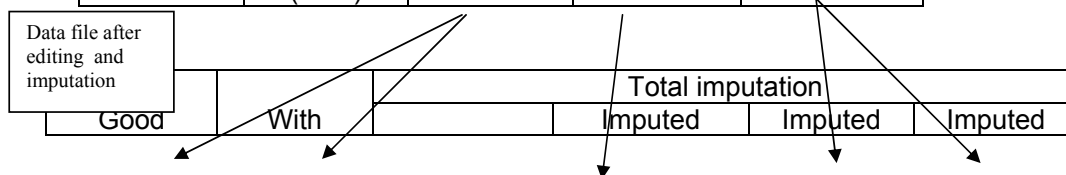
There are, however, instances when imputation is done only for a few non-responding samples until a sufficient number of reports (original and imputed) in the non-certainty stratum is obtained. Estimation still makes use of weight adjustment to derive the levels.

Likewise, imputation of all non-responses in the certainty stratum is required provided these are known to be in operation at the survey reference date.

- iii. **‘Do Nothing’** - This approach is rather simple and applicable only for non-responding samples, which are reported as closed or out-of-scope, in the certainty stratum.

The table below presents the status of establishment’s reports used in the tabulation of the 2002 ASPBI for manufacturing (reference year 2001).

Total Samples	Receipts	Good reports	Closed/CB L/ OS	Non-response (all stratum)
3,496	2,969 (85%)	2,308	661	527



reports QR = 1	partial imputation QR = 2	Total QR = 3	closed/CBL/O S (<i>non-certainty stratum</i>)	non- responding (<i>non- certainty stratum</i>)	non- responding (<i>certainty stratum</i>)
2,110	186	854	403	288	163

Note: The remaining non-responses and closed/CBL establishments in the non-certainty stratum were not imputed. Their estimates were handled by weight adjustment.

III. Imputation Methods

The matrix of imputation methods used for unit non-response and item non-response is shown in the following table.

	Hot Deck	Cold Deck
1. Historical imputation		x
2. Mean value imputation	x	x
3. Nearest neighbor imputation	x	x
4. Ratio imputation	x	x
5. Use of data from other survey/ external source		x
6. Combination of any of the above	x	x

i. **Hot deck** – In this method, the missing data is replaced by a (randomly chosen) value from the ‘clean’ respondent/s’ in the current survey.

1. Mean imputation – This method makes use of the mean value of the responding establishments in the stratum. A good application of this method is when imputing employment variable in the non-certainty stratum.
2. Nearest neighbor imputation for unit non-response – This method makes use of the values of all items of a responding establishment in the same stratum as the non-responding sample. This assumes that the non-responding sample has similar characteristics for all items as the respondent donor. This method is best applied when no previous information is available about the non-responding sample. (It is not advisable to use this method for item imputation as the correlation between variables may be lost.)
3. Ratio imputation – This method makes use of the value ratio of two items from responding establishment/s in the stratum to impute the missing data of an item (for item imputation).

4. Combination of any of the above.

ii. Cold deck – In this method, the missing data is replaced by a value from previous survey or other data collected from external sources.

1. Use of historical value – This method makes use of values reported by the same sample establishment in previous survey, with no adjustment for trend.

2. Use of historical value with trend adjustment – This method makes use of values reported by the same sample establishment in previous survey, and adjusted for trend (growth rate). Growth rate may be based on growth rate of responding establishment/s (nearest neighbor/s) in the stratum, or from other external sources, i.e. Y-Y growth.

3. Ratio imputation – This method makes use of value ratio of two items from responding establishment/s in previous survey to impute missing value of an item (for item imputation). The donor establishment should be in the same stratum as the sample for which an item is imputed.

4. Nearest neighbor imputation for unit non-response - This method makes use of the values of all items of a responding establishment in the previous survey. This is usually adopted when there are no other responding establishment/s in the current survey. The donor establishment should be in the same stratum as the non-responding sample.

5. Use of data from other survey or other external sources for the same period, i.e. MISSI, QSPBI.

6. Combination of any of the above.

iii. Combination of any hot deck and cold deck method.

IV. Recommendations

- Establish guidelines on acceptable stratum response rate when applying either the weight adjustment or imputation method for missing data.
- Aggressive implementation of preventive measures to minimize cases of non-responses. This requires good planning for the survey, including the availability of updated and comprehensive frame of establishments and effective collection strategies.
- Establish guidelines on the use of various types of imputation methods. Study the possible use of other imputation methods, i.e. Regression models, automated imputation, etc.

- Estimates of variance should consider the effect of imputed values. Whenever possible, separate variances shall be estimated for **actual** values and imputed values.

References

Kovar, John and Rancourt, Eric. *Workshop on Editing and Imputation of Survey Data*. Berlin 2003.
 Kalton, Graham and J. Michael Brick. *Methods of Handling Missing Survey Data and Their Effects*. Berlin 2003

Definition of terms

A **stratum**, in the recent annual survey of establishments, is defined in terms of region, industry (3-digit PSIC) and employment size for ATE 20 and over. For ATE less than 20, a stratum is defined in terms of industry at the national level.

A **non-certainty stratum** is one in which probability sampling was applied in the selection of samples.

A **Certainty stratum** is one in which all units in the stratum are selected as samples.

'**Minimum of 3 rule**' requires that the minimum number of samples in a non-certainty stratum (with $N \Rightarrow 3$) be set at 3, regardless of the sampling rate. The **maximum number**, on the other hand, is 10.

Comment:

QR or Quality of Report is an identifying characteristic of an establishment report. QR with code=1 refers to a good report, code=2 refers to a report with some partially imputed items, and code=3 refers to a report that is totally imputed.

Comment: !!!