Mass Imputation

Karol Krotki, Stephen Black, and Darryl Creel
Research Triangle International, 1615 M St., Suite 740, Washington, DC 20036

Key Words: Nonresponse, large-scale data processing and adjustment

1. Introduction

This paper gives an overview of the methods used to handle missing data in the 2004 National Postsecondary Student Aid Survey (NPSAS:04). More generally, this paper deals with the concept of mass imputation – or the process of simultaneously filling in large blocks of missingness in a data file.

As surveys become larger and data sets get bigger, dealing with large amounts of missing data is quickly becoming an issue to data producers. Holes in the data, whether caused by unit or item nonresponse, are beginning to become particularly problematic. Another motivating factor is the requirement by some agencies to provide complete data sets.

The statistical literature is replete with research on single variable imputation with very little on mass imputation whose underlying theory and guidelines have not been widely discussed, established, or understood.

2. Definition

Previous use of the term "mass imputation" has referred to the multiple use of a single donor, imputing from a sample back to the frame, and imputing for a large block of unobtainable data. These uses of the term seem rather esoteric and limited and we would propose that the term be reserved for a more generic application.

As we use the term in this paper, it refers to the simultaneous imputation of a very large number of variables, in the hundreds. The ultimate objective is to fill in large blocks of missingness in a data file.

3. Motivation

Traditional reasons for imputing data include the need to produce a complete data set, the fact that data producers are best qualified and have the most relevant resources, and, finally, the empirical observation that if data producers do not impute, users will, and in multiple ways. It has been widely recognized that having at least one definitive version is highly desirable.

More specifically, in the case of mass imputation, there are several additional and important reasons. The first is that data sets are becoming larger and more complex and holes in the data pose bigger challenges for the users. Second, mass imputation is needed to adjust for nonresponse, specifically item nonresponse. Third, it is increasingly the case that clients, especially federal government agencies, require complete data sets particularly to adjust for item nonresponse. And, finally, users should always keep in mind that the unimputed data are available in case there are serious questions about the imputation process.

Widely available and popular imputation methods can be used on an ad hoc basis to assist in mass imputation but there is a need for more formal study of alternative strategies. One objective of this paper is to lay out some of the foundation for this research.

4. Background

The general setting being dealt with here is a data set consisting of hundreds of variables of varying type, continuous and categorical, and nonresponse rates ranging across a very broad spectrum. For NPSAS:04, the following shows the distribution of nonresponse rates for the entire dataset:

Table 1: NPSAS:04 Nonresponse Rates

Percent Missing	**Number of Variables**
None	23
Less than 5%	37
Between 5% and 29.9%	56
Between 30% and 49.9%	59
Between 50% and 79.9%	51
More than 80%	31
Total	257

These data are cross-sectional but there is little reason for not expanding these methods to time series data. In fact, such data, particularly ones involving many time points, could easily lend themselves well to mass imputation since we can leverage data from previous timepoints for more accurate imputation.

5. Strategy

The general strategy employed is to block the variables into relatively homogeneous blocks/clusters and carry out the imputation procedure sequentially, both within and across the blocks of variables. Once variables are imputed, they are used in subsequent stages. Vector imputation is used when several variables exhibit similar patterns of missingness. When vector imputation is not possible, we still aim to impute blocks of variables simultaneously in order to save time and computing resources.

6. Procedure

Once the variables are grouped into clusters, the percent missing and pattern of missingness are used to determine which method to employ – sequential or vector imputation. Variables with less than 5% missing are imputed initially using a logical (deductive/deterministic imputation) approach. In fact, all attempts are made to maximize the use of logical imputation before implementation of any stochastic process. Remaining variables are studied to determine the optimal ordering in which imputation would occur. Variables with similar patterns of missingness and percent missing are candidates for vector imputation. The remaining variables are sent through the imputation process sequentially in blocks in order to save time/resources.

Imputation classes are formed based on a CHAID analysis of likely candidates for variables related to those being imputed. Once the imputation classes are defined, the data are sorted within each class in the spirit of implicit stratification before implementing the actual imputation. A weighted sequential hot-deck imputation procedure is used to impute the missing data. Figure 1 (see below) lays out a summary schematic of the processing strategy.

7. Programming Strategies

Mass imputation imposes heavy demands on computer capacity and execution time and, whereas this does not pose an insurmountable challenge, it is important to design the programs and data processing with care to ensure maximum efficiency and avoid bottlenecks due to processing demands.

As a minor example, for the processing of these NPSAS data we relied on a CHAID-based algorithm to generate optimal weighting classes which were then input into the hot-deck imputation program. The fact that the first was handled in SPSS and the second in SAS created considerable logistic problems.

8. Evaluation

A large number and variety of methods can be used to evaluate the imputed data. These include checking the number of times individual donors are used and distribution percentages overall and by imputation class (see Table 2 below). Flags are switched on for differences that exceed a prespecified threshold. The distributions that are compared are both unweighted and weighted. We also use logical consistency checks to minimize the number of cases with inconsistent data.

More sophisticated evaluation techniques include measures of bias reduction and variance estimates using multiple imputation. These were not included as part of this project.

9. Recommendations

We would recommend that future mass imputation efforts focus on the following areas:

- Develop a unified, streamlined, and efficient computer system to handle all aspects of the imputation process.
- Derive imputation actions automatically from the edit rules thus avoiding the possibility of imputing inconsistent data.
- Incorporate regression imputation as part of the system, for handling continuous variables with known relationships.

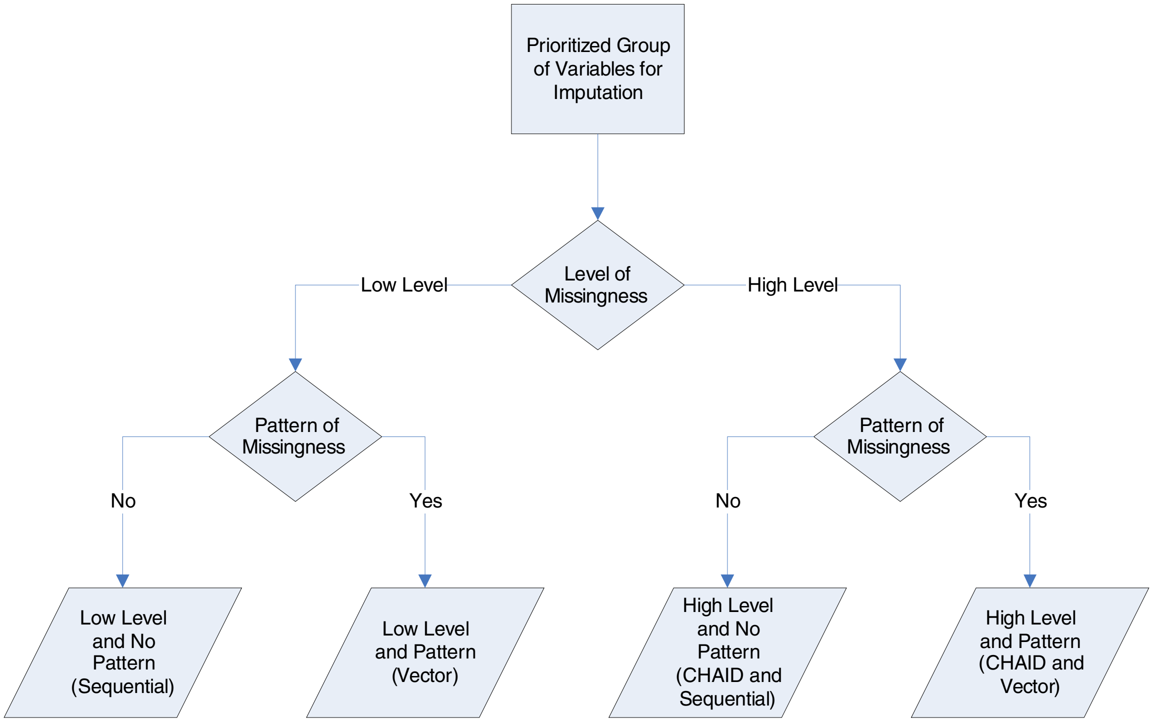
10. Conclusion

Mass imputation is becoming one more tool in the survey statistician's toolkit for which there is an ever-increasing demand. As a result there is a need to develop theory and application guidelines for implementing this methodology effectively and efficiently. Finally, we believe that a major part of the implementation strategy will be empirically determined, that is, very much data-driven.

TABLE 2: COMPARE WEIGHTED DISTRIBUTIONS BEFORE AND AFTER IMPUTATION.

Weighting Class	Variable 1	Variable 2	Before Percent	After Percent	Difference	Flag Diff	Before Count	After Count
1	2	1	55.6	55.7	0.097	0	155,993	225,902
2	2	2	39.1	38.8	0.260	0	109,704	157,540
3	2	3	5.3	5.5	0.163	0	14,864	22,150
4	4	1	30.3	27.8	2.488	0	4,250	7,070
5	4	2	47.3	49.8	2.472	0	6,639	12,661
6	4	3	22.3	22.4	0.016	0	3,133	5,683

FIGURE 1: IMPUTATION DECISION FLOW CHART


Prioritized Group of Variables for Imputation
Level of Missingness
Low Level
High Level
Pattern of Missingness
Pattern of Missingness
No
Yes
No
Yes
Low Level and No Pattern (Sequential)
Low Level and Pattern (Vector)
High Level and No Pattern (CHAID and Sequential)
High Level and Pattern (CHAID and Vector)