Imputation and Unbiased Estimation: Use of Centered Predictive Mean Neighborhood Method

A.C. Singh, E.A. Grau, and R.E. Folsom, Jr.
RTI International, NC
egrau@rti.org

Abstract

Methods for determining the predictive distribution for multivariate imputation range between two extremes, both of which are commonly employed in practice: a completely parametric model-based approach, and a completely nonparametric approach such as the nearest neighbor hot-deck (NNHD). A semiparametric middle ground between these two extremes is to fit a series of univariate models and construct a neighborhood based on the vector of predictive means. This is what is done under the predictive mean neighborhoods (PMN) method, a generalization of Rubin's (1986) and Little's (1988) predictive mean matching method. Because the distribution of donors in the PMN neighborhood may not be centered at the recipient's predictive mean, estimators of population means and totals could be biased. To overcome this problem, we propose a modification to PMN which uses sampling weight calibration techniques such as the GEM (generalized exponential model) method of Folsom and Singh to center the empirical distribution from the neighborhood. Empirical results on bias and MSE, based on a simulation study using data from the 2002 National Survey on Drug Use and Health, are presented to compare the centered PMN with other methods.

1. Introduction

Under a suitable data model, imputations for missing items can be obtained from the predictive distribution for unbiased estimation of parameters of interest. However, in practice, the datasets are often large with many outcome and auxiliary variables, and it is desirable to have multivariate imputation to preserve associations between variables. In these cases, there is a need of joint modeling of outcome variables (say, y's which typically have some missing values), and the covariates (say, x's which are typically complete).

An alternative to the above parametric approach is the commonly used nonparametric approach of nearest neighbor hot deck (NNHD) imputation. Here a neighborhood consisting of donors (deemed close to the recipient in terms of a metric based on observed x's) gives rise to an empirical distribution, that approximates the joint distribution of y's conditional on the observed x's of the recipient. A single donor is selected at random from the neighborhood and its y-values are used for multivariate imputation whenever needed. Although this approach takes away the burden of modeling, it allows for donors having x-values different but close to those of the recipient, and so the conditional distribution given x's is somewhat tenuous. Also, the empirical distribution may be unreliable because of the difficulty in finding sufficient donors in the neighborhood. Thus unbiasedness of the estimated parameters using imputed values may be questionable.

In principle, the modeling route is probably the best approach provided the underlying assumptions are reasonable as it lends itself to different types of analyses. In fact, if the percentage of missing values is high and the goal of analysis is varied (quantiles, regression models and multivariate relationship between variables besides the usual means and totals), it may be better to use the predictive distribution under a model to impute missing values. This would, of course, require considerable time and effort to check adequacy of the model especially for high dimensional data. However, if the percentage of missing values is not too high and the goal is primarily estimation of means and totals (and some regression parameters such as contrasts), then it is often desirable to have a simple and expedient alternative to meet time constraints without relying heavily on strong modeling assumptions. Also, in this case, it may be desirable as a robust alternative to choose imputed values from real donors satisfying certain constraints which may be hard to incorporate in the model. Such a situation arose in the context of NSDUH which motivated the development of the imputation method presented in this paper.

A possible compromise of the parametric and nonparametric approaches described above is to use a semiparametric approach which postulates a univariate predictive mean model for each y given the x's which can be fit using a working covariance structure if necessary. The vector of predictive means for the recipient is used to find a neighborhood of donors using Mahalanobis distance such that donors' predictive means are close to the recipient's vector of predictive means within a certain margin. Once the neighborhood is defined, then a donor is selected at random for multivariate imputation. This is the idea underlying the PMN (predictive mean neighborhood) method of Singh, Grau, and Folsom (2001) which is a generalization of the PMM method of Rubin (1986)

and Little (1988). PMN, unlike NNHD, uses more information from y's and x's as summarized by the predictive mean vector in finding the donor. PMN is simple to implement because of univariate modeling, in that it relies on weak modeling assumptions. Moreover, it can also provide approximately unbiased estimates under the model for population means and totals (this forms a very widely used class of parameters), provided the empirical mean of the donors in the neighborhood is close to the recipient's vector of predictive means. This is true even though the values of the x's for the donors need not be close to those of the recipient, which is often the case in practice due to attempts to have a sufficient number of donors. In the case of Hot Deck with imputation classes, a simple class mean model is postulated for the predictive mean. Weighted Hot Deck (see e.g. Rao and Shao, 1992) provides an example where the mean of the distribution used for selecting a donor matches the recipient's vector of predictive means.

The problem of unbiased estimation of population means and totals in PMN arises when the empirical mean of a neighborhood for a given recipient is not the same as that recipient's predictive mean. We propose a modification of PMN, termed centered PMN (CPMN), which enforces the mean of the empirical distribution obtained from the neighborhood to be equal to the recipient's predictive mean, through adjustment of the probability of selection for each donor within the neighborhood. This change allows the (unweighted) empirical mean to differ from the recipient's vector of predictive means. Under CPMN, estimates of means and totals using imputed values would be unbiased under the mean model. However, unbiasedness may not hold for parameter estimates for regression models in general as in the case of commonly used NNHD methods.

For CPMN, we assume all y's are categorical. Some y's, if necessary, could be categorized for this purpose. We propose to use sampling weight calibration techniques such as the GEM (generalized exponential model) method of Folsom and Singh (2000) to center the marginal means of the empirical joint distribution obtained from the neighborhood for each recipient. For this purpose, the observed cell proportions in the joint categorical distribution from the neighborhood are treated as initial weights in the calibration step, and the recipient's predictive mean vector provides the calibration controls. It may be necessary to enlarge the neighborhood to make GEM converge, i.e., satisfy the calibration controls.

In this paper we discuss a simulation study based on 2002 NSDUH to compare different methods of imputation such as PMN, CPMN, weighted sequential hot deck, and unweighted sequential hot deck. Estimators of domain means are evaluated in terms of Mean Squared Error (MSE).

2. Review of Several Imputation Methods

The random NNHD is a commonly used imputation method (Little & Rubin, 1987, p. 65). With this method, donors and recipients are distinguished by the completeness of their records with regard to the variable(s) of interest (the donor has complete data, the recipient does not). A donor set deemed close to the recipient with respect to a number of covariates is used to select a donor at random. To further ensure that a donor matches the recipient as closely as possible, discrete variables (or discrete categories of continuous variables) strongly correlated with the variable being imputed can be used to restrict the set of donors. In NNHD, a distance function is used to define closeness between the recipient and a donor. So there is less of a problem of sparseness of the donor class, but the distance function involving categorical or nominal variables is typically ad hoc and often hard to justify.

The predictive mean matching method (PMM) represents an important development in the area of imputation which was introduced by Little (1988) as a generalization of the original Rubin's (1986) method. It was introduced in the context of continuous outcome variables although the idea could be generalized to discrete variables as is done by PMN. With PMM, a distance function is used to determine distances between the predictive means for the recipient and potential donors are obtained under a model. The respondent with the smallest distance is chosen as the donor. Unlike the NNHD, the donor is not randomly selected from a neighborhood. The advantages of PMM include

1. Model bias in the predictive mean can be minimized by using suitable covariates.

2. The PMM method is not a pure model based method because the predictive mean is only used to assist in finding a donor. Hence, like NNHD, it has the flexibility of imposing certain constraints on the set of donors. However, the choice of donor is nonrandom. This nonrandomness leads to bias in the estimators of means and totals. It also tends to make the distribution of outcome values skewed to the center.

Predictive Mean Neighborhoods (PMN) was developed for the National Survey on Drug Use in Health (NSDUH) in 1999. PMN is a combination of the two commonly used imputation methods, the NNHD and PMM. PMN enhances the predictive mean matching (PMM) method in that it is not necessarily deterministic and the natural generalization of its applicability to discrete variables. PMN also enhances the nearest neighbor hot deck (NNHD) method in that the distance function used to find neighbors is no longer ad hoc.

In the univariate version of PMN (denoted as UPMN), as in the case of predictive mean matching, the prediction model is fit to the data from complete respondents, but the predictive means for both recipient and the donor are computed to obtain the distance between the two predictive means. With this innovation, the distance is also well defined for discrete variables, and a delta neighborhood (delta chosen to be a small positive number signifying that donors in the neighborhood are within delta-distance from the recipient, i.e., their predictive means are almost equal to that of the recipient) is defined to pick one donor at random. In the multivariate version of PMN (denoted as MPMN), a set of predictive means can be obtained from a single multivariate model, or from a series of univariate models. Fitting a multivariate model requires the specification of a covariance structure, which may not be straightforward, while fitting univariate models may preclude the ability to incorporate the correlations between the outcome variables in the models. A middle ground is to fit a sequence of univariate models, where models later in the sequence are conditioned on outcomes from models earlier in the sequence. This is the strategy pursued in the NSDUH; however, due to time constraints that strategy was not pursued in this paper. Rather, a series of univariate models are fitted that do not incorporate the full covariance structure of the explanatory and response variables, except in the final assignment of imputed values.) Regardless of the method used to obtain the vector of predicted means, a neighborhood for picking a donor is defined using a vector delta or the Mahalanobis distance or both.

In addition to PMN and the proposed CPMN, other methods evaluated in this study include a regression imputation and a weighted sequential hot deck (WSHD). For the regression imputation, the missing value is replaced by the predicted value, with no added error term. In the WSHD, the choice of classing variables and order of the sorting variables was crucial, with variables lower down in the sorting order having minimal impact in cases where the number of respondents with such attributes is sparse within the given imputation class. The order of the sorting variables was determined by looking at the levels of significance of the variables in the predicted mean models. Specifics about the WSHD, and the software used for its implementation, are discussed in more detail in the appendix.

3. Proposed Method

The PMN method provides unbiased estimates if the empirical mean of the observed responses in the neighborhood of potential donors is the same as the vector of predicted means for the recipient. In practice, an exact match is usually impossible, though the empirical mean is often very close. However, there are instances where insufficient donors are available to provide an empirical mean that is sufficiently close to the recipient's predicted means, leading to bias in the estimates. With the centered PMN method, unbiasedness is guaranteed if the selection of a donor in the neighborhood is altered so that each donor does not have equal weight, but rather donors are weighted so that the weighted empirical mean of the donors in the neighborhood is equal to the recipient's predicted mean. The weights are calibrated using the GEM technique of Folsom and Singh (2000), where the observed cell proportions in the joint categorical distribution from the neighborhood are treated as initial weights in the calibration step, and the recipient's predictive mean vector provides the calibration controls. In the example where three binary outcome variables require imputation, the three outcomes lead to 8 categories, each of which appears with a certain frequency within the neighborhood. These frequencies, converted to proportions, serve as the initial weights in the calibration.

4. Methodology

The bias and variance for each of the imputation methods were evaluated using a simulation. The simulation study was conducted by setting up the model in several stages, then comparing imputation techniques under the known model.

Data model

A loglinear model was based on a set of demographic covariates fitted to the complete respondents aged 18 to 25 from the 2002 NSDUH. The covariates in this loglinear model included a three-level categorical variable representing age, a three-level race-ethnicity variable, gender, a three-level categorical variable

representing last grade completed in school, a binary marital status variable, and a binary employment variable (full-time employed vs. otherwise), as well as 2- and 3-factor interactions involving these variables, resulting in a total of 216 expected cell frequencies. A multinomial logistic model was then fitted, conditional on these x-values, with the 8-category response variable resulting from the combination of the three binary lifetime usage indicators for cigarettes, alcohol, and marijuana.

Generated population

Data from a finite population were generated from which samples would be drawn. A finite population of 1,000,000 cases was generated using the integer count obtained from the loglinear expected values. Responses were randomly allocated to the 8 categories corresponding to the binary lifetime usage indicators, using the multinomial logistic model.

Obtain samples with induced missingness

Twelve strata were defined in the population based on age, race/ethnicity, and gender, including three levels of age (18 to 20, 21 to 22, and 23 to 25), two levels of race/ethnicity (non-Hispanic white and Hispanic/non-Hispanic nonwhite), and the two levels of gender. Five hundred stratified samples were obtained based on proportional allocation, each with approximately 500 observations (each stratum had a minimum of two observations). Item nonresponse of approximately 10% missingness for all three lifetime indicators was induced in each sample through randomization. Missingness depended upon employment status, so that a missing-at-random mechanism was generated. The following mechanism was used: for full-time employed respondents, 8% of the values for all three indicators were set to missing, and for respondents who were not full-time employed, 12% of the sample had missing values for all three indicators. Nonresponse that depended upon the response was also evaluated, though the conclusions did not differ appreciably.

Impute missing values

To obtain imputed values, univariate logistic imputation models were fitted for each of the three binary indicators. Sampling weights were not adjusted for item nonresponse. The covariates used in the models were the same as those used to define the population (though 3-factor interactions were excluded). Missing values were imputed using the following methods: Centered PMN, PMN, weighted sequential hot deck (WSHD), and regression. For the WSHD, imputation classes were defined by the two most highly significant variables in the predicted mean model, employment status (full-time employed vs. not full-time employed) and education level (highest grade completed: less than high school, high school grad, at least some college). Sorting variables included race/ethnicity (non-Hispanic white, non-Hispanic black, Hispanic/non-Hispanic other), age (18-20, 21-22, 23-25), marital status (married vs. not married), and gender.

Evaluation

Bias was evaluated by looking at the prevalence for each of the three drug variables in question: cigarettes, alcohol, and marijuana. The prevalence of the three drugs in question was compared overall and within domains (males vs. females, and three race/ethnicity categories). Evaluation criteria included a comparison of the bias as well as the simulation variance. Bias was evaluated by comparing the sample prevalence with the imputed values against the finite population prevalence. Table 1 gives the overall population prevalence for each drug, obtained from the generated population. These were the values against which the sample prevalence values were compared.

Table 1. Drug prevalence in generated population, 18-25 year olds

Drug	**Domain**	**Prevalence**
Alcohol	Overall	0.8670
	Males	0.8762
	Females	0.8584
	Non-Hispanic whites	0.9078
	Non-Hispanic blacks	0.7930
	Hispanics & Other	0.8113
Cigarettes	Overall	0.6997
	Males	0.7311
	Females	0.6703
	Non-Hispanic whites	0.7575
	Non-Hispanic blacks	0.5652
	Hispanics & Other	0.6367
Marijuana	Overall	0.5328
	Males	0.5674
	Females	0.5006
	Non-Hispanic whites	0.5920
	Non-Hispanic blacks	0.4917
	Hispanics & Other	0.4164

5. Results

Although it had been theorized that bias could be a problem for existing methods, this simulation study was unable to detect any meaningful bias, with any of

the methods. Furthermore, none of the methods showed a consistent pattern of higher or lower variance beyond what was expected.

Tables 2-4 give the overall mean and variance of prevalence estimates, and Tables 5-7 give the mean and variance of prevalence estimates by gender. For the sake of brevity, tables giving the mean and variance of prevalence estimates by race/ethnicity group are omitted. The "Method" column summarizes the various imputation methods that were compared in this study, where the "Population" row is the prevalence obtained from the generated population. The columns corresponding to "Average mean", "Standard error of mean" and "Average variance" refer to the mean and variance of the prevalence estimates across the 500 iterations. Bias is determined by comparing the average estimate obtained using the imputation method under the "Method" column with the estimate corresponding to "Population." Before approximately 10% of the sample had values for the lifetime usage of cigarettes, alcohol, and marijuana set to missing, the original generated values were saved. The mean and variance of the prevalence estimates with these original generated values (where no imputation was required), were calculated. The row corresponding to "Nonmissing" gives the estimates when no imputation was required. The difference between the average variance in the "Nonmissing" row and the average variance for the other methods gives a sense of the contribution of imputation to the overall variance. The column named "Sig. bias" is an indicator of whether the prevalence estimate obtained for a given method was significantly different from the overall prevalence estimate at the 95% level (i.e., whether the bias differed significantly from zero). No effort is made to account for a multiple comparison effect.

Table 2. Evaluation of bias and variance: alcohol (overall)

Method	Average mean	Standard error of mean	Average variance	Sig. bias
Population	0.8670			
CPMN	0.8680	0.00072	0.000260	No
PMN	0.8685	0.00071	0.000254	Yes
Regress.	0.8678	0.00069	0.000238	No
WSHD	0.8680	0.00072	0.000258	No
Non-missing	0.8680	0.00066	0.000215	No

Table 3. Evaluation of bias and variance: cigarettes (overall)

Method	Average mean	Standard error of mean	Average variance	Sig. bias
Population	0.6997			
CPMN	0.6991	0.00104	0.000508	No
PMN	0.6993	0.00101	0.000505	No
Regress.	0.6992	0.00096	0.000461	No
WSHD	0.6994	0.00099	0.000493	No
Non-missing	0.6994	0.00093	0.000434	No

Table 4. Evaluation of bias and variance: marijuana (overall)

Method	Average mean	Standard error of mean	Average variance	Sig. bias
Population	0.5328			
CPMN	0.5331	0.00108	0.000580	No
PMN	0.5334	0.00110	0.000601	No
Regress.	0.5331	0.00103	0.000534	No
WSHD	0.5338	0.00108	0.000583	No
Non-missing	0.5328	0.00097	0.000466	No

Table 5. Evaluation of bias and variance: alcohol (by gender)

Males				
Method	Average mean	Standard error of mean	Average variance	Sig. bias
Population	0.8762			
CPMN	0.8760	0.00103	0.000529	No
PMN	0.8768	0.00104	0.000537	No
Regress.	0.8761	0.00101	0.000507	No
WSHD	0.8755	0.00101	0.000508	No
Non-missing	0.8763	0.00094	0.000437	No
Females				
Method	Average mean	Standard error of mean	Average variance	Sig. bias
Population	0.8584			
CPMN	0.8605	0.00105	0.000550	Yes
PMN	0.8607	0.00103	0.000526	Yes
Regress.	0.8600	0.00101	0.000501	No
WSHD	0.8610	0.00100	0.000498	Yes
Non-missing	0.8602	0.00095	0.000451	No

Table 6. Evaluation of bias and variance: cigarettes (by gender)

Males				
Method	Average mean	Standard error of mean	Average variance	Sig. bias
Population	0.7311			
CPMN	0.7310	0.00141	0.000991	No
PMN	0.7312	0.00139	0.000964	No
Regress.	0.7314	0.00134	0.000895	No
WSHD	0.7293	0.00133	0.000878	No
Non-missing	0.7310	0.00126	0.000794	No
Females				
Method	Average mean	Standard error of mean	Average variance	Sig. bias
Population	0.6703			
CPMN	0.6693	0.00144	0.001035	No
PMN	0.6695	0.00144	0.001030	No
Regress.	0.6692	0.00137	0.000944	No
WSHD	0.6714	0.00136	0.000929	No
Non-missing	0.6699	0.00133	0.000884	No

Table 7. Evaluation of bias and variance: marijuana (by gender)

Males				
Method	Average mean	Standard error of mean	Average variance	Sig. bias
Population	0.5674			
CPMN	0.5676	0.00153	0.001177	No
PMN	0.5671	0.00158	0.001234	No
Regress.	0.5674	0.00149	0.001106	No
WSHD	0.5650	0.00150	0.001122	Yes
Non-missing	0.5668	0.00137	0.000932	No
Females				
Method	Average mean	Standard error of mean	Average variance	Sig. bias
Population	0.5006			
CPMN	0.5010	0.00153	0.001168	No
PMN	0.5018	0.00151	0.001143	No
Regress.	0.5011	0.00145	0.001057	No
WSHD	0.5047	0.00146	0.001069	Yes
Non-missing	0.5009	0.00137	0.000941	No

As is clear in these tables, the bias is not a major factor for any of the estimates; clearly, the bias ratio (bias/standard deviation of estimate) is much less than 10%, the rule of thumb given by Cochran (1977). The bias is significant in only a small number of cases, some of which may be due to the multiple comparison effect. (There are 45 calculated biases. With a 95% test, if none of the biases were nonzero, we would expect 2 or 3 to be significant just from random variation, and we see 6 significant results.) This is particularly true since there doesn't appear to be a strong, consistent pattern across drugs or methods, though one could argue that there is a marginally larger bias with WSHD.

Although no difference in the variance was apparent among the methods that incorporated an error term (PMN, CPMN, and WSHD) when estimates were determined overall, it is apparent that WSHD actually has a slightly smaller variance than either PMN or CPMN when broken down by gender (and race), so that the mean squared error across the methods did not differ substantially (due to the larger bias of WSHD). Not surprisingly, regression imputation had a lower variance than PMN, CPMN, or WSHD since no error term was added to the predicted mean. The variance associated with the situation where the original values were maintained (no imputation was necessary) was clearly the smallest. This indicates that there is a substantial increase in the variance of the estimates when values have to be imputed, which could loosely be attributed to the variance due to imputation. This provides strong support to the notion that this source of variation must be accounted for, particularly when the amount of missingness is large.

6. Summary

The bias for the existing methods had never been evaluated, but it was theorized that bias could be a problem based on the fact that the predicted means from the imputation models for donors and recipients differed, in some cases substantially. This study was unsuccessful in finding this bias, which either indicates that (1) the study is not sufficiently sensitive to detect this bias; or (2) the bias is not large enough to be of any major concern. In this study, in general, the mean of the empirical distribution of donors was approximately equal to the recipient's predicted mean in most cases. This is not true in general, indicating that a more complicated simulation study may be needed to determine whether CPMN significantly reduces bias in situations where the predicted means of donors and recipients are far apart. This could be done by including more x-values in the definition of the generated population, or in setting up a more complex set of response variables. The difficulty with this, of course, is obtaining a convergent

loglinear model or multinomial logistic model with these more complicated structures.

A problem with CPMN that hasn't been addressed here is that all the focus is placed on the predicted mean. There may be instances, however, where variables that cannot be included in the determination of the predicted mean are strongly related to the response. This may be due to convergence problems that would result from including these variables in the model, or from problems resulting from instability of estimates that would occur by including these variables. The advantage of PMN is that these variables can be included in the hot deck step of PMN. It is unclear how much centering negates the effect of these auxiliary variables. Clearly, more research is needed to resolve these issues.

References

Chromy, J. R. (1979). Sequential sample selection methods. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 401-406). Alexandria, VA: American Statistical Association.

Cochran, W. G. (1977). *Sampling Techniques*. Third Edition. New York: John Wiley & Sons.

Cox, B. G. (1980). The weighted sequential hot deck imputation procedure. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 721-726). Washington, DC: American Statistical Association.

Folsom, R.E., Jr., and Singh, A. C. (2000) A generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *Proceedings of the American Statistical Association, Survey Research Methods Section* pp. 598-603.

Iannacchione, V. (1982). Weighted sequential hot deck imputation macros. In *Proceedings of the Seventh Annual SAS Users Group International Conference*. Cary, NC: SAS Corporation.

Little, R.J.A. (1988), Missing-data adjustments in large surveys, (with discussion), *Jour. Bus. Econ. Stat.*, 6, 287-301.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Rao, J.N.K., and Shao, (1992) Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*. 79, 811-822.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics, 4* (1), 87-94.

Singh, A.C., Grau, E.A., and Folsom, R.E., Jr. (2001) *Proceedings of the American Statistical Association, Survey Research Methods Section for 2001*. need title

Williams, R. L., & Chromy, J. R. (1980). SAS sample selection MACROS. In *Proceedings of the Fifth International SAS Users Group International Conference* (pp. 382-396). Cary, NC: SAS Corporation.

Appendix: The Weighted Sequential Hot Deck

A.1 Introduction

Typically, with the hot-deck method of imputation, missing responses for a particular variable (called the "base variable" in this appendix) are replaced by values from similar respondents with respect to a number of covariates (called "auxiliary variables" in this appendix). If "similarity" is defined in terms of a single predicted value from a model, these covariates can be represented by that value. The respondent with the missing value for the base variable is called the "recipient," and the respondent from whom values are borrowed to replace the missing value is called the "donor."

A step that is common to all hot-deck methods is the formation of imputation classes, which is discussed in Section A.2. This is followed by a discussion of the sorting methods used for sequential hot decks in Section A.3. Specific details of the weighted method are given in Section A.4. With the weighted hot deck, the identities of the donors are generally tracked. For more information on the general hot-deck method of item imputation, see Little and Rubin (1987, pp. 62-67).

A.2 Formation of Imputation Classes

When there is a strong logical association between the base variable and certain auxiliary variables, the dataset is partitioned by the auxiliary variables and imputation procedures were implemented independently within classes defined by the cross of the auxiliary variables.

A.3 Sorting the File

Within each imputation class, the file is sorted by auxiliary variables relevant to the item being imputed. The sort order of the auxiliary variables is chosen to reflect the degree of importance of the auxiliary variables in their relation to the base variable being imputed (i.e., those auxiliary variables that are better predictors for the item being imputed were used as the first sorting variables). In general, two types of sorting procedures have been used in NSDUH imputation procedures:

- **Straight Sort**. A set of variables is sorted in ascending order by the first variable specified; then within each level of the first variable, the file is sorted in ascending order by the second variable specified; and so forth.

- **Serpentine Sort**. A set of variables is sorted so that the direction of the sort (ascending or descending) changed each time the value of a variable changed.

The serpentine sort has the advantage of minimizing the change in the entire set of auxiliary variables every time any one of the variables changes its value.

A.4 Weighted Sequential Hot Deck

The steps taken to impute missing values in the weighted sequential hot deck are equivalent to those of the unweighted sequential hot deck. The details on the final imputation, however, differ with the incorporation of sampling weights. The first step, as always, is the formation of imputation classes, following by sorting the variables appropriately. The assignment of imputed values is necessarily more complex than in the unweighted case.

The procedure described below follows directly from Cox (1980). Specifically, once the imputation classes are formed, the data is divided into two data sets: one for respondent and one for nonrespondents. Scaled weights $v(j)$ are then derived for all nonrespondents using the following formula:

$$v(j) = w(j)s(+)/w(+);\ j = 1, 2, \ldots n$$

where n is the number of nonrespondents, $w(j)$ is the sample weight for the j^{th} nonrespondent, $w(+)$ is the sum of the sample weights for the all nonrespondents, and $s(+)$ is the sum of the sample weights for all the respondents (Cox, 1980). The respondent data file is partitioned into zones of width $v(j)$, where the imputed value for the j_{th} nonrespondent is selected from a respondent in the corresponding zone of the respondent data file.

This selection algorithm is an adaptation of Chromy's (1979) sequential sample selection method, which could be implemented using the Chromy-Williams sample selection software (Williams & Chromy, 1980). Furthermore, Iannacchione (1982) revised the Chromy-Williams sample selection software, so that each step of the weighted sequential hot deck is executed in one macro run.

Benefits of Weighted Sequential Hot-Deck

With the unweighted sequential hot-deck imputation procedure, for any particular item being imputed, there is the risk of several nonrespondents appearing next to one another in the sorted file. An imputed value could still be found for those cases, since the algorithm would select the previous respondent in the file; however, some modifications are required in the sorting procedure to prevent a single respondent from being the donor for several nonrespondents. With the weighted sequential hot-deck method, on the other hand, this problem does not occur because the weighted hot deck controls the number of times a donor can be selected. In addition, the weighted hot deck allows each respondent the chance to be a donor since a respondent is selected within each $v(j)$.

The most important benefit of the weighted sequential hot-deck method, however, is the elimination of bias in the estimates of means and totals. This type of bias is particularly present when the response rate is low or the covariates explain only a small amount of variation in the specified variable. In addition, many surveys sample subpopulations at different rates, and using the sample weights allows, in expectation, the imputed data for the nonrespondents to have the same mean (for the specified variables) as the respondents. In other words, the weighted hot deck preserves the respondent's weighted distribution in the imputed data (Cox, 1980).