# Item Imputation with the Discrete Edit System

**Yves Thibaudeau, Bor Chung Chen, William E. Winkler**
Yves Thibaudeau, U.S. Census Bureau, Washington, DC 20333

**Key Words**: Nearest Household Type; Nearest Neighbor, Discriminant Analysis.

## 1. Introduction

The Fellegi-Holt algorithm (Fellegi and Holt 1976) provides a framework for item imputation by identifying for each record with one or more edit failures a minimal set of fields that must be imputed in order to satisfy a cohesive set of edits. The set of fields is minimal in the sense that there exists at least one joint value for the fields in the set such that, when this joint value is substituted, it results in a record with no edit failure and there does not exist a smaller set of fields that also provides an edit solution. In the paper, we develop an imputation strategy based on a framework similar to Fellegi-Holt. The input of the imputation process consists of a quasi-minimal set of fields for each record that fails the edits, in addition to distributional information on the fields of the records not failing the edits. The goal of the paper is to show how DISCRETE, an edit-impute system developed at the Census Bureau, process this distributional information to retrieve joint values for a set of fields in such a way that: first, these joint values resolve all the edit conflicts, and second the solution provided is optimal relative to a decision rule based on a likelihood function. In other words, we show how DISCRETE seeks to resolve edit conflicts with a solution that is *minimal* in the Fellegi-Holt sense, but also with a solution that is *probable* in reference to a likelihood function.

The supporting statistical model for our imputation system is the multivariate correlation model with mixed discrete and continuous variables discussed by Olkin and Tates (1961) – See also Schaffer (1997). Based on this model, categorical variables divide a population in categories of households, in the case of a demographic survey or a census. Then, conditional on the categorical variables, the ages of the persons in a household, or suitable transformations of these ages, jointly follow a multivariate normal distribution. This mixed continuous-categorical model provides natural distance functions to identify optimal imputations for the household failing edits, in terms of their reported information. In the paper we focus on the imputation of the categorical items when the ages of persons in a household are reported. We give priority to that situation for two reasons: First reported of ages are more reliable than other information because respondents are asked their age, and those of the other members in the household, twice. First, the ages are requested and second the birth dates are requested. Therefore, in situations where there are edit failures, given the choice between changing age and changing a categorical item to resolve the failure, when the reported ages are corroborated with a matching birth date, DISCRETE is geared to impute the categorical item, rather than changing age. The second reason to focus on the imputation of the categorical items is that the imputation of age, when needed, is relatively straightforward. Indeed, it is implemented with multivariate regression tools and follows naturally from the model.

Categorical item imputation with DISCRETE integrates some of the basic concepts of nearest neighbor item imputation (Chen and Shao 1997, 2000), which are implemented in edit-impute systems such as NIM (Masson, Bankier and Poirier 2002). Like nearest neighbor imputation DISCRETE uses a distance function to identify optimal item imputations. But the distance function of DISCRETE is different of those defined for traditional nearest neighbor systems in that it serves to identify entire classes of survey units, rather than single units. For instance, in the case of demographic surveys and censuses, instead of identifying a single household to serve as an "imputation donor", as nearest neighbor algorithms do, DISCRETE identifies an entire class of households, which acts as a surrogate for a specific household failing edits. The class is optimal in the sense that it is closer to the household than any other class in terms of a distance function. The selected class is unambiguously characterized in terms of the items to be imputed, and thus it determines the values of the imputations. More specifically, the imputation of categorical items with DISCRETE is equivalent to resolving a problem of classification into one of several populations (Anderson, 1958 page 142), which is resolved with discriminant analysis.

DISCRETE requires a large amount of observations to fully catalogue the patterns of joint values of the fields subject to edits. Because of its size, the American Community Survey (ACS) is a good test bed for DISCRETE. In the next section, we describe the general approach of DISCRETE and we describe how imputation with DISCRETE is articulated on the concepts of likelihood and frequency, in addition to the concept of proximity, unlike the traditional nearest neighbor approach, which depends on proximity only. In section 3 we give the specifics of the methodology in the context of categorical household item imputation. In section 4 we present an example of household record editing in the context of the ACS to illustrate the methodology.

## 2. Item Imputation with the Nearest Household Type Approach for the ACS

Discrete is based on a more explicit approach than a generic nearest-neighbor algorithm in the sense that all the observed *household types* that comply with the edits are identified and catalogued before proceeding to any item imputation. Furthermore, the frequency of each household type is recorded. The imputation problem boils down to deciding which one of these types will serve as the surrogate for a particular household failing the edits. As with the traditional Nearest Neighbor approach, we want to select a household type that is "close" to the household to be imputed. But, unlike with the traditional Nearest Neighbor approach, we explicitly embed frequency considerations in the distance function, which is akin a likelihood function. In that context, choosing the household type minimizing the distance function is equivalent to selecting the item imputations needed for a household based on the likelihood of their joint occurrence for a generic household of that same type.

The first measurement that enters in the characterization of the various household types is household size. Household size is also used in generalized nearest neighbor systems, such as NIM (Mason, Bankier, Poirier 2002): In addition to household size, the joint categorical measurements relating to selected members of the household also serves to classify the population in household types. If we assume only categorical measurements need to be imputed, DISCRETE implements the imputation of items for a household by identifying the household type that best fits the age pattern exhibited by its members given the available categorical information on the household.

## 3. Imputation of Relationship, Marital Status, and Sex Based on the Age Pattern with Discriminant Analysis

We center the attention on the imputation of the categorical items and we show how DISCRETE uses a discriminant function to process this imputation. Linear or quadratic discriminant analysis based on age (after a square-root transformation) can be applied to impute items by classifying a household with edit failures in a population represented by a household type. Specifically, DISCRETE classifies a household as being of type $(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s})$, where $\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s}$ are respectively the vectors of relationships, marital statuses, and sexes of the member of the household whenever $D\left(\boldsymbol{a}^{H}, (\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s})\right)$, the distance function between household $H$ and the household type with relationship, marital status, and sexes given by $(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s})$, is minimal. An explicit form for the distance function is:

$$D\left(\boldsymbol{a}^{H}, (\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s})\right) \propto$$

$$q^{-1}(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s}) \left|\Sigma_{\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s}}\right|^{1/2} \times$$

$$\exp\left[\frac{1}{2}\left(\boldsymbol{a}^{H} - \mu_{\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s}}\right)' \Sigma^{-1}_{\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s}} \left(\boldsymbol{a}^{H} - \mu_{\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s}}\right)\right]$$

$$(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s}) \in \boldsymbol{Q} \qquad (1)$$

Note that $D\left(\boldsymbol{a}^{H}, (\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s})\right)$ is the inverse of the discriminant function associated with household type $(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s})$ evaluated at $\boldsymbol{a}^{H}$, which is the vector of the ages corresponding to household $H$. Furthermore, $\mu_{\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s}}$ and $\Sigma_{\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s}}$ are respectively the mean age and the age covariance matrix corresponding to the population represented by "household type" $(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s})$. $q(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s})$ in (1) can be thought of as the "prior probability" of household type $(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s})$. In practice we set $q(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s})$ to be equal to the observed frequency of the occurence of households of type $(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s})$. $\boldsymbol{Q}$ represents the set of edit constraints on $(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s})$. Therefore we have

$$(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s}) \notin \boldsymbol{Q} \Rightarrow q(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s}) = 0 \qquad (2)$$

(2) expresses the fact that the distance between any household and a household type which would itself have a structure of categorical variables that fails the edits is infinite. We also assume that $\boldsymbol{Q}$ includes the constraint $\boldsymbol{r}_1 = 1$, which means that the pre-edit program always identify one householder per household whose relationship to the householder (self=1), marital status and sex are represented by the elements at the first position of the vectors $\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s}$ . The householder is typically an adult and the parent of some of the children in the household, whenever there are children.

The offshoot is that the distance function in (1) integrates together the concept of closeness between a household and a household type, in terms of the proximity of $\boldsymbol{a}^H$ to $\boldsymbol{\mu}_{r,m,s}$, and the concept of frequency, through the inverted likelihood in (1), and in particular through the inverted frequency $q^{-1}(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s})$.

## 4. Example

We give an example of item imputation with DISCRETE for a typical household failing edits. Tables 1 – 4 exihibit the 3 imputation steps required to resolve the edit failures of connected with this household of size four. Table 1 shows the reported household configuration after DISCRETE has identified the edit failures. The relationship of person 3 is flagged because the edit rule that specifies that a parent of the householder must be at least 15 years older than the householder is broken, and changing the relationship to comply to the rule is possible and constitutes a minimal change. In addition, table 1 reveals three ureported items for this household: the sexes of persons 2, 3, and 4. DISCRETE will change the flagged relationship and impute the missing items by identifying household types that are close to this household in terms of the distance function in (1) after matching on some of the reported categorical items. The item imputation proceeds in three steps. The order of the steps follow from the logical construction imposed by the edit rules. Discrete first imputes the sex of person 4, a relative of the householder other than his spouse because the imputation of this item does not impart any structural change in the household.

To impute the sex of person 4., DISCRETE selects the closest household type among all the types sharing the same values as the current household in terms of the marital statuses of the householder and of person 2 (usually the spouse of the housholder when there is one), and of person 4. There is a total of 12 household types that meet these requirements on marital status. Table 5 shows the three closest types among them. Note that, although we use the concept of household type to characterize 4-person households, the type itself is defined in terms of three person only. The idea is that a three-person structure that always include the householder, along with the spouse or partner, if any, is sufficient to determine the value of the imputations, when this structure corresponds to a compatible three-person structure in the household. Common three-person structures are: householder-spouse-child, householder-unmarried partner-roomate, householder-spouse-parent etc…

Table 5 also gives the frequency at which each of the three closest household types occurs, as well as their corresponding values for the discriminant function (the inverse of the distance). From table 5 it is clear that the event that person 4 is female is strongly favored. The value of "male" for the sex of person 4 is a distant third, both in terms of the distance function (discriminant function) and of frequency. Thus sex of person 4 is imputed to be "female", as shown in table 2.

The second step of DISCRETE involves imputing items whose imputed values could have consequences in terms of the fundamental structure of the household. In our example, sex and relationship of person 3 are respectively missing and flagged and DISCRETE imputes these items jointly. This joint imputation must maintain the integrity of the other relationships in the household. For instance person 3 cannot be another spouse, or an "unmarried partner", or a parent (parents must be 15 older than the householder). Moreover, the joint value of the imputation must be compatible with the overall age pattern of the household. DISCRETE ensures this compatibility by considering only household types with matching marital statuses for the householder, person 2, and person 3. There are 29 household types that satisfy these requirements for marital status. Table 6 shows the three closest househhold types among them. Again, note that each household type is defined by a three-person structure. The closest household type yields the values of "child" and "male" for the relationship and sex of person 3.

The third step in our example is the imputation of sex for the spouse. DISCRETE considers household types with matching values for the sex of the householder (male), the marital status of the householder (married), and the relation of the

spouse with the householder (spouse). It is clear then that any legitimate household type will generate a value of "female" for the sex of the spouse.

## 5. Discussion

The advantage of the DISCRETE methodology over a traditional nearest-neighor methodology is that the distance function reflects the likelihood of the possible joint configurations of the imputed and reported items over the conditional domains defined by the matching variables. On the other hand, traditional nearest neighbor edit and/or imputation does not refer to the concepts of likelihood or frequency, but rather it exclusively exploits the concept of distance to justify the imputations. Consequently we expect DISCRETE to generate configurations of reported and imputed items that are "probable", and not only "plausible", conditional to the reported information. In addition, DISCRETE imputes items based on the attempted identification of the minimal set of fields to impute in order to resolve all edit conflicts. Thus we expect DISCRETE to give a solution that typically involves the least changes to the data, a desirable property on the whole. The disaadvantage of Discrete as it stands now is precisely that it allways choose the most likely imputation, which can result in a bias at aggregate levels. A possible remedy would be to randomize. For instance, in the example, at step 2 discrete could the household type by randomizing between the three types shown in table 6, with probabilities proportional to the discriminant functions. On the whole, DISCRETE offer an attractive blend of probabilistic imputation and optimal editing that is geared towar preserving the integrity of the data.

## 6. References


Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis, Wiley.

Chen, J., Shao, J. (2000). Nearest Neighbor Imputation for Survey Data. Journal of Official Statistics, **16**, 2.

Chen, J., Shao, J. (1997). Biases and Variances of Survey Estimators Based on Nearest Neighbor Imputation. *Proceedings for the Section on Survey Research Methods, American Statistical Association.*

Fellegi I. P., Holt D. (1976). A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*. 71, 353.

Mason P., Bankier M., Poirier P. (2002). Imputation of Demographic Variables From the 2001 Canadian Census of Population. American Statistical Association. CD ROM.

Olkin I., Tate, R. F. (1961) Multivariate Correlation Models with Mixed Discrtte and Continuous variables. Annals of Mathematical Statistics, 32, 448-465.

Schaffer J. L. (1997). Analysis of Incomplete Multivariate Data. Chapman & Hall.

**Table 1 – Example: Originally Reported Household**

| | Age | Relation with Householder | Marital Status | Sex |
|---|---|---|---|---|
| **Householder** | 61 | Self | Married | Male |
| **Person 2** | 49 | Spouse | Married | **Missing** |
| **Person 3** | 21 | **Parent (Flagged by Edits)** | Never Married | **Missing** |
| **Person 4** | 86 | In-Law | Widowed | **Missing** |

**Table 2 – Example: Step 1: Univariate Imputation of Sex for Persons 4**

| | Age | Relation with Householder | Marital Status | Sex |
|---|---|---|---|---|
| **Householder** | 61 | Self | Married | Male |
| **Person 2** | 49 | Spouse | Married | **Missing** |
| **Person 3** | 21 | **Parent (Flagged by Edits)** | Never Married | **Missing** |
| **Person 4** | 86 | In-Law | Widowed | **Female** |

**Table 3 – Example: Step 2: Joint Imputation of Relation and Sex for Persons 3**

| | Age | Relation with Householder | Marital Status | Sex |
|---|---|---|---|---|
| **Householder** | 61 | Self | Married | Male |
| **Person 2** | 49 | Spouse | Married | **Missing** |
| **Person 3** | 21 | **Child** | Never Married | **Male** |
| **Person 4** | 86 | In-Law | Widowed | **Female** |

**Table 4 – Example: Step 3: Univariate Imputation of Sex for Person 2**

| | Age | Relation with Householder | Marital Status | Sex |
|---|---|---|---|---|
| **Householder** | 61 | Self | Married | Male |
| **Person 2** | 49 | Spouse | Married | **Female** |
| **Person 3** | 21 | **Child** | Never Married | **Male** |
| **Person 4** | 86 | In-Law | Widowed | **Female** |

**Table 5 - Closest Household Types in the Imputation of Sex of Person 4 (Step 1)**

| | **Observed Household** | **Closest Type** | **Second Closest** | **Third Closest** |
|---|---|---|---|---|
| **Relationship of Person 2** | Spouse | Spouse | Spouse | Spouse |
| **Relationship of Person 4** | In-Law | Parent | In-Law | Parent |
| **Marital Status of Householder** | Married | Married | Married | Married |
| **Marital Status of Person 2** | Married | Married | Married | Married |
| **Marital Status of Person 4** | Widowed | Widowed | Widowed | Widowed |
| **Sex of Person 4** | **Missing** | **Female** | **Female** | **Male** |
| **Discriminant Function** | NA | .5082 | .4083 | .0334 |
| **Frequency of the Type** | NA | 73 | 53 | 4 |

**Table 6 - Closest Household Types in the Joint Imputation of Relationship and Sex of Person 3 (Step 2)**

| | **Observed Household** | **Closest Type** | **Second Closest** | **Third Closest** |
|---|---|---|---|---|
| **Relationship of Person 2** | Spouse | Spouse | Spouse | Spouse |
| **Relationship of Person 3** | **Parent (Flagged by Edits)** | **Child** | **Child** | **Other Relative** |
| **Marital Status of Householder** | Married | Married | Married | Married |
| **Marital Status of Person 2** | Married | Married | Married | Married |
| **Marital Status of Person 3** | Never Married | Never Married | Never Married | Never Married |
| **Sex of Person 3** | **Missing** | **Male** | **Female** | **Male** |
| **Discriminant Function** | NA | .5425 | .4634 | .0070 |
| **Frequency of the Type** | NA | 6833 | 6319 | 39 |