NONRESPONSE IMPUTATION IN PUBLIC HEALTH RESEARCH

S. van Buuren*, J.L.A. van Rijckevorsel**

* Department of Psychometrics, Faculty of Social Sciences, University of Utrecht (presently at the Department of Statistics and Computer Science, TNO Institute of Preventive Health Care)

** Department of Statistics and Computer Science, TNO Institute of Preventive Health Care

Abstract

This paper suggests a method to supplant missing categorical data by `reasonable' replacements. These replacements will maximize the consistency of the completed data. Consistency is measured by a between-total variance ratio. The idea is that similar profiles obtain comparable imputations. The text outlines the optimization problem, describes relationships to the relevant psychometric theory and studies some properties of the method. Some examples are presented. The main application fields are in the analysis of survey data, rating scales and questionnaires.

Keywords: *missing data, imputation, categorical variables, consistency, least squares, MISTRESS, multiple imputation*

3.1 INTRODUCTION

Missing data are common and costly. Public health data with up to 30% missing are no exception, and attempts to do something about missing data are as old as missing data themselves. Popular ways to accommodate for missing data - pairwise - and listwise - deletion - may amount to wasting labour-intensively collected material. An alternative is to fill in missing entries with `appropriate' replacements. The advantage of this is that standard multivariate techniques can be applied to the completed data.

Sometimes, external information is available that may help. For example, suppose that a subject is unwilling to tell his age, thereby producing a missing value. If we have the actual person in front of us we may nevertheless infer his age by using other clues, and subsequently fill in our prediction as if it had been observed. Another source of external information could be a previous score on same question or test. Unfortunately, such situations are more an exception than a rule.

In general, if we want to complete the data we should look for other sources of information. An obvious alternative is to consider the data that *are* available for the subject, in combination with the data that are collected on other sample units. Under the assumption that observations with a similar response pattern are likely to score identically on any remaining, unobserved variables, we may try to interpolate missing values. This type of imputation strategy is known as *hot deck* imputation. The basic `reasonable imputation assumption' is: objects with almost similar profiles have the same distribution on any missing responses. Thus, the idea is to borrow the observed score from a closely related profile. Some arguments that are often mentioned in favour of hot deck methods are: the reduction of the response bias, the preservation of the distribution of the population, and - most important of all - the production of a complete data set. Particularly in large surveys, computational ease and speed are highly evaluated.

In this paper, we consider a hot deck imputation technique based on the within-homogeneity of all variables simultaneously. The method consists of two ingredients: a *donor variable* and an *imputation rule*. The donor variable measures how much individual data profiles differ from each other. So, a donor is not a sample unit or a data profile, but a latent variable. The imputation rule states how blank entries should be filled, given the values on the donor variable. As we will see, these two components are closely

intertwined in the method. Changing the donor variable may cause a modification of the imputations. The converse is also true; changing an imputation has an effect on the donor variable.

The donor variable is equal to a weighted average of all variables. We use the familiar between - total sum of squares - ratio to indicate how well the donor represents the total variation in the data. It follows that the `best' donor variable is equal to the first principal component of the completed data. The position, or score, of each observation on the donor reflects how much sample units have in common. The function of the donor is thus much like the partitioning of observations into homogeneous classes employed by traditional hot deck procedures. The difference with existing hot deck procedures is that all (categorical) variables act simultaneously as donor.

Imputation is based on comparing donor scores. If an incomplete profile resides closely to a completely observed profile (in the sense that their donor scores differ little), then its missing entries can replaced by the known values of the observed unit. We measure 'closeness' by the squared Euclidean distance between donor scores. Mathematically, we look for imputations that minimize a sum-of-squared-distances function.

In practice, just one donor may not be representative. Primary reasons for using just one is that it is simple, and that it has some attractive analytical properties. If necessary, the extension to multiple, orthogonal donors is possible. We will indicate where this is appropriate.

The donor is the most homogeneous replacement for all variables simultaneously, and hence it is most homogeneous with regard to the complete and incomplete data. This satisfies the, according to Ford (1983), most important principle in the construction of any hot deck procedure: the `imputation model' and the `data model' must be the same. Both imputations and quantifications maximize the homogeneity of the completed data set. As such they are relevant to the observed as well as to the missing observations.

The present imputation method is similar to missing data estimation by the EM algorithm (Little and Rubin, 1987) in that both methods optimize an objective function over the imputations. Moreover, both methods consist of the two main steps: an Expectation (E) step that completes an incomplete data matrix, and a Maximization (M) step that estimates the model parameters. However, there are also substantial differences. We use Least Squares instead of Maximum Likelihood, we do not make any distributional

assumptions, and we provide discrete instead of fractional imputations , which are spread out over more than one cell of the contingency table.

We like to emphasize at this point that maximizing consistency is by no means the only valid or useful criterion to find missing information. Suppose that we are interested in demonstrating that two variables are independent of each other. In that case, it will be clear that maximizing consistency is a bad idea since it moves us further away from the independence model. A more natural alternative here would be to do the opposite, that is, to minimize homogeneity. There exists yet no unbiased technique nor a general purpose strategy for dealing with missing data. Multivariate optimality for imputation of missing data is about impossible to define without violating some statistical model or another. Nonetheless, *if* we believe that the observed data tell us something about the missing data - and this is a fundamental assumption of all hot deck methods - then maximally consistent replacements will be attractive in general.

Good reviews imputation techniques for categorical data are Kalton and Kasprzyk (1982) and the three volumes edited by Madow, Olkin and Rubin (1983). The annual proceedings of the Section of the Survey Research Methods of the American Statistical Association offer a continuing story on the handling of missing data in survey research. The primary source for Maximum Likelihood models for missing categorical data is Little and Rubin (1987). For applications of multiple imputation in public health see Rubin and Schenker (1991). Missing data in experimental designs are discussed in Dodge (1985). For multiple imputation, in which not just one but many replacements are searched, see Rubin (1987). Hedges and Olkin (1983) give a selected and annotated bibliography on incomplete data. Ford (1983) summarizes many hot deck strategies. Little and Rubin (1990) provide a recent overview of missing data strategies in the social sciences.

The structure of this paper is as follows: first, we discuss a small imputation example. After this, we define the consistency measure, and we introduce the loss function. Subsequently, we relate the consistency criterion to other psychometric theory, and indicate a number of similar approaches. Practical use of the method is illustrated by some examples, one of them concerning multiple imputation. Finally, we summarize the main results and we discuss some practical implications and future work.

3.2 EXAMPLE

To be able to grasp the nature of consistent imputations, we discuss the small artificial data listed in Table 3.1. This table contains 10 observations on three categorical variables. There are three missing values, indicated by *a*, *b*, and *c*.

Table 3.1 Example Data

Person	Income	Age	Car
1	*a*	young	jpn
2	middle	middle	am
3	*b*	old	am
4	low	young	jpn
5	middle	young	am
6	high	old	am
7	low	young	jpn
8	high	middle	am
9	high	*c*	am
10	low	young	am

The problem is to find replacement values that are reasonable in some way. For *a* this is easy; the most consistent estimate is *low*, because this makes the profiles 1, 4, and 7 identical. A young owner of a Japanese car will have a low income simply because this is a recurring pattern. Moreover, the profile contains all Japanese cars in the data. Analogously, we find *high* for *b* and *old* for *c*. Both imputations make the remaining two incomplete profiles identical to row 6. So, the missing scores are interpolated from other profiles. We simply look for similar rows. This is the same as saying that variables must be as homogeneous as possible, i.e., measure the same thing. So here we end up with two homogeneous groups with three members each.

Since there are 3 missing values, each with 3 categories to choose from, the total number of different solutions is $3 \times 3 \times 3 = 27$. Table 3.2 lists the amount of consistency for each of these solutions. The exact definition of consistency, expressed as fit, can be found in the next section. Because the example is deliberately easy and somewhat trivial, the most consistent solution `*l h o*' can be derived by eye-balling alone. In more realistic situations, eye-balling is usually not enough. First, because the best solution may contain new, previously unobserved, profiles. It will be difficult to find such combinations. Second, because optimal consistency becomes hard to detect for more than three or four missing values.

For categorical data, Wilks procedure - filling in the average- boils down to selecting the modal category. The corresponding solutions in the example are '*l l y*' and '*h h y*'. These imputations have consistencies of 0.70104 and 0.68827 respectively, which illustrates the well known fact that Wilks method tends to discard between-groups variance.

Table 3.2 Consistency of all possible imputations for Table 1

a b c	Fit	*a b c*	Fit	*a b c*	*Fit*
l l y	.70104	m l y	.63594	h l y	.61671
l l m	.77590	m l m	.72943	h l m	.66458
l l o	.76956	m l o	.72636	h l o	.65907
l m y	.78043	m m y	.70106	h m y	.70106
l m m	.84394	m m m	.77839	h m m	.74342
l m o	.84394	m m o	.77839	h m o	.74342
l h y	.78321	m h y	.73319	h h y	.68827
l h m	.84907	m h m	.80643	h h m	.74193
l h o	.84964*	m h o	.80949	h h o	.74198

The obvious difficulty with categorical data is that distances between profiles cannot be easily derived; it makes little sense to substract Japanese from American cars. We deal with categorical variables by first transforming them into numerical data, by quantifying each category separately. Subsequently, the resulting numerical variables combine into the donor. Table 3.3 lists those donor scores for each observation and the scale values of the categories, both before and after imputation.

Table 3.3 Donor Scores and Scale Values for the Optimal Imputation

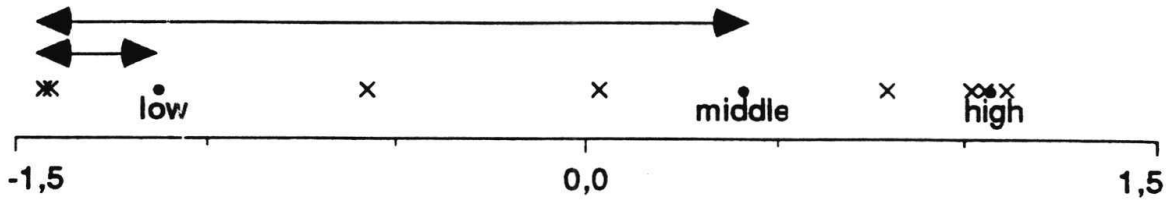
		Donor Scores		Variable	Scale Values	
		Initial	Final		Initial	Final
1	*l* y j	-1.43	-1.33	Inc low	-1.13	-1.15
2	m m a	0.79	0.66	middle	0.41	0.33
3	*h* o a	1.02	1.00	high	1.06	0.98
4	l y j	-1.41	-1.33			
5	m y a	0.04	-0.01	Age young	-0.96	-0.92
6	h o a	1.11	1.00	middle	0.92	0.79
7	l y j	-1.41	-1.33	old	1.07	1.00
8	h m a	1.05	0.92			
9	h *o* a	1.02	1.00	Car jpn	-1.41	-1.33
10	l y a	-0.58	-0.59	am	0.63	0.57

The initial solution ignores missing data, an option known as 'missing passive'

(Meulman, 1982; Gifi, 1990, p. 136). In this case, donor values for subjects 1, 3, and 9 are based on two, instead of three, observed categories. As will be shown below, a scale value for a category is equal to the average of all donor scores that fall into that category. For example, the initial value of *low* is equal to $-1.41 - 1.41 - 0.58 / 3 = -1.13$, i.e., the average donor score of observations 4, 7, and 10.

Let us now try to impute the incomplete entry in profile 1. The initial donor score of the profile is –1.43. To complete the data, we may pick any of the three income low, middle or high categories: The scale values of these categories are –1.13, 0.41, and 1.06. The joint scale of the initial donor scores and the category points is plotted in Figure 3.1.

Figure 3.1 Joint plot of objects (x) and the categories of income (•)



The most consistent imputation is that category whose scale value is closest to the donor. Here, –1.13 is closest to –1.43, so we choose *low*. Apparently, subject 1 has most in common with profiles 4, 7, and 10. So, when compared to the other two income classes, the low income group is most similar to profile 1. Consequently, we borrow the replacement value from this group. Since middle and high income groups are more distinct, imputing *middle* and *high* will increase the within-groups variance more than necessary, and so, these values should not be used as stand-ins.

We execute the same steps for the missing data in rows 3 and 9. After all missing entries have received an initial imputation, new donor scores are computed, but now using the completed data. The entire process is repeated until the consistency of the solution does not change anymore. The values of the final solution are also given in Table 3.3.

3.3 METHOD

Let the data be coded into m binary indicator vectors g_j of length k_j for j=1, ..., m, and let random vector y_j contain k_j category quantifications for the j-th variable, where k_j is the number of categories. The quantified variables $x_j = g_j'y_j$ have zero means and the random variable z contains their average, i.e., $z = 1/m \sum x_j$. The total variation of the data can be decomposed as

$$\Sigma_j (g_j'y_j)^2 = m\, z^2 + \Sigma_j (z - g_j'y_j)^2.$$

This is a between-within partitioning of the form $T = B + W$. The *correlation ratio*, denoted by η, and defined by $\eta^2 = B / T$, measures how well the average can be considered as a representative of each x_j. The ratio ranges from 0 to 1. The coefficient equals 1 if all variables are proportional.

The donor variable z tells us something about the similarity among profiles that belong to distinct replications. Let z_i for i = 1, ..., n denote the score of the i-th profile on the donor z. The difference between z_i and $z_{i'}$ is equal to some distance norm between profile i and profile i'. The donor variable z defines a metric in which observational units can be represented, which we use to compare different data profiles. The correlation ratio η can also be interpreted as a measure of how well the entity $z_i - z_{i'}$ reflects the multivariate differences between rows i and i' over all x_j. Obviously, the larger η becomes, the better the difference $z_i - z_{i'}$ portrays the similarity between i and i'. We might say that z is a satisfactory donor variable in this case.

Procedures for finding optimal η over y_1, ...,y_m are known as homogeneity analysis, multiple correspondence analysis, dual scaling and others (see Gifi, 1990). These techniques usually consider several orthogonal sets of z's, with corresponding η's. We define the donor variable z as the set of numbers that maximizes η.

We now discuss missing data. Let Ω denote the set of all *observed* variables and let the $x_j^* = g_j^{*\prime}y_j$ stand for an imputed value. Obviously,

$$x_j = \begin{cases} x_j, & \text{if } j \in \Omega \\ x_j^*, & \text{if } j \notin \Omega \end{cases}$$

It is possible to partition the variation into three independent quadratic components:

$$\Sigma_j (g_j'y_j)^2 = m\, z^2 + \Sigma_{j\in\Omega} (z - g_j'y_j)^2 + \Sigma_{j\notin\Omega} (z - g_j^{*\prime}y_j)^2.$$

Since $T = B + W$ the maximum of η^2 coincides with the minimum of $W / T = 1 - \eta^2$. Maximal homogeneity among the imputed variables can be found by minimizing this W / T-ratio over $z, y_1, \ldots, y_m$ *and* over the imputations $g_1^*, \ldots, g_m^*$. The corresponding loss function can be written as

$$\sigma(z; y_1, \ldots, y_m; g_1^*, \ldots, g_m^*) = \Sigma_{j\in\Omega} (z - g_j'y_j)^2 + \Sigma_{j\notin\Omega} (z - g_j^{*\prime}y_j)^2.$$

Let $\sigma(\cdot)$ stand for this loss function. The imputation problem is where to impute the '1' in the missing vector g_j^*. This is a combinatorial optimization problem.

Since larger η lead to more consistent imputations, it seems logical to look for imputations that will maximize η. We thus strike two flies at one blow: imputations will not only amplify the structure of multivariate row differences as summarized by z, but, at the same time they cause z to be a more adequate composite of those differences. This principle induces imputations such that similar looking units become even more alike, while plainly different units grow even more distinct under imputation. Dependencies in the data are thus extrapolated to the missing entries.

The solution can be computed by an iterative algorithm based on a combination of homogeneity analysis and the *k*-means algorithm. It works by comparing donor scores and category quantifications. The exact procedure can be found in Van Buuren and Van Rijckevorsel (1991, 1992a). See Van Buuren and Heiser (1989) for a related optimization problem.

3.4 MAXIMIZING CONSISTENCY BY IMPUTATION

The search for scores that maximize consistency is deeply rooted in psychometrics, and the following results are mainly due to this development.

It is known that η^2 is proportional to the largest eigenvalue of the correlation matrix R

of (quantified) variables. We also know that η^2 is equal to the averaged squared correlations between the quantified variables and donor variable. The average correlation among variables is another well-known measure for internal consistency. This measure is also used for categorical data, if computed from the optimal scores instead of the raw data. The average correlation is proportional with Cronbach's α (Cronbach, 1951), This is a very popular statistic in item analysis and questionnaire research. Cronbach (1951) showed that α is very similar to the average correlation. Lord (1958) showed that η^2 can be written as a function of α, so maximizing η over the missing data also maximizes Cronbach's α, the average correlation, the largest eigenvalue of the correlation matrix, and related measures.

We finish this section with the following. The idea to maximize consistency by imputation is not entirely new. Gleason and Staelin (1975) replace correlations between numerical variables by estimates that maximize the consistency of the completed data. This method is a modification of the imputation techniques proposed earlier by Dear (1959) and Buck (1960). Gleason and Staelin treat categorical data by an ad hoc rounding procedure (p. 244). Unlike the numerical case, they do not present any simulation results for their discrete imputation method. In an analysis of variance context, Hartley and Hocking (1971) identify the so-called (*X*, *m*, *d*) model in which one tries to find estimates for missing classifications on the experimental variables. This is a combined estimation and classification problem. They note some difficulties with the model, but they do not pursue the matter any further. Nishisato (1980) wants to impute and quantify categorical data, just like in this paper, but does not present a practical solution to the problem of selecting the optimal category to be imputed.

A difficulty with imputing categorical data in general is that one has a limited set of donor categories to choose from and no distance measure between them. One can quantify categories and use the Euclidean distances - as we do - or try to find margins that optimize consistency. Greenacre (1984, p. 237) does the latter by imputing 'consistency optimizing' rounded estimates of marginal frequencies.

3.5 MULTIPLE IMPUTATION

A drawback of any imputation method that imputes a single value is that the precision of the imputations is unknown, i.e., the variance is not estimated. In MISTRESS, one could say that the imputation variance is equal to zero, since there is only one imputation that maximizes consistency. This shows much confidence in the appropriateness of consistency as a criterion, and in the reliability of the data. According to Rubin: "It is of no use looking for the `best' or `most appropriate' imputation. Such a thing simply doesn't exist." (Rubin, 1987). What is best for one model doesn't work for another. So one has to make a distinction between the optimal value in terms of the one closest to the real, but unobserved, value and an imputation that is best in some model sense. Such values coincide if we succeed in finding that only model that generated the data; a desirable but rarely attained state of affairs, as every data analyst knows. Only in simulation studies, where indeed reality is artificially simulated and thus grossly simplified, one can hope for and achieve the coincidence of such imputations.

A different approach is to estimate the variance of imputation by generating not one, but several, say 3 to 5, completed matrices. Imputations are to be drawn from a posterior predictive distribution, or from decent approximations thereof. The spread of the imputations then conveys roughly how imputations vary. Rubin (1987) shows for a large class of statistical models that, after a model is separately fitted on each completed data matrix, simple pooling procedures can be used to obtain unbiased estimates of model parameters and the associated variances. The individual imputations do not have to be very precise, as long as together they estimate the variance. Rubin and Schenker (1991) discuss various applications of multiple imputation in health-care databases. Because multiple imputation involves a lot of work, it is worth the effort if it concerns a large body of data that is to be used by several researchers applying different models and different subsets of the data on various occasions. See also Schnell (1986, p. 227). In psychometrics, the combination of multiple imputation and sampling in various combinations is discussed by Rubin (1991).

For categorical data multiple imputations are to be drawn from a predictive distribution of categories. One can define such a distribution in several ways. The dominant distinction lies between *implicit* and *explicit* models. If we use a specified distribution

to this purpose like the normal, we use an explicit model. Often there exists no proper argument to select an explicit model, and thus an implicit model, or implicit distribution is used. The most implicit model is the traditional hot deck method, where the value of the preceding observation is imputed. Multivariate simultaneous consistency is a less implicit model.

Because there is only one optimal imputation per missing value, it is impossible to generate multiple imputations by just maximizing consistency. MISTRESS yields a crisp `all-or-none' predictive distribution for each incomplete response pattern, which is not very useful in the context of multiple imputation. For multiple imputation, we must have some way to even out the predictive category distribution so that all categories are candidates for imputation, though with varying probabilities. It would require another paper to discuss MISTRESS as a way to create posterior predictive distributions of missing data. Here we only mention some possibilities of doing so as a way to apply the method.

Let p_{ijk} denote the probability that category k of variable j is the imputation for object i, then we can specify the following:

a) a density distribution based on inverse distances

We assume that p_{ijk} is inversely related to the squared distance between the scale value y_{jk} and the donor score x_i.

b) a density distribution based on multiple donors

We mentioned that multiple orthogonal z's with corresponding y's can be considered as well. We can use each column of Z to generate a successive, separate imputation instead of using the first column of Z with the largest consistency only. The columns of Z are ordered by their respective contribution to the overall consistency, denoted by η^2. The inverse of the relative contribution defines the probability that an imputation is sampled from the imputations corresponding to the s-th column of Z. These probabilities are independent of i and j. The range of potential categories to be chosen is then restricted by their occurrence in one of the k_j imputation values. If the same category value occurs more than once as an imputation the probabilities for different s add up. One could sample as many times as one likes, but just as many draws as the number of columns of Z seems reasonable.

c) a density distribution based on conditional frequencies

One of the oldest methods assumes that p_{ijk} is proportional to the observed frequency of category k of variable j. This is a very simple way to define a predictive distribution, but it uses only univariate information. If we crosstabulate the data, each cell corresponds to a possible response pattern, and one may use the conditional frequencies instead. Note that this way of deriving the distribution will only be effective if the cells in the multidimensional crosstabulation contain a sufficient number of observations. In practice, this implies that the number of variables is limited.

These alternatives yield different predictive distributions. We do not know how this effects the results. We expect that differences will be relatively small, but more research is needed to confirm this idea. In the next section, we apply the option of inverse donor distances, with in this case quite satisfactory results.

3.6 DUTCH LIFE STYLE SURVEY

This example is taken from the Dutch Life Style Survey (Leef Situatie Onderzoek) conducted by the Netherlands Bureau of Census. The data were collected at different time points during the years 1977-1986. The data are compiled and made available to us by Anneke Bloemhoff of NIPG-TNO. As is often the case in large surveys, not all questions were posed at each occasion. Consequently, when taken together, the data contains many systematic missing entries. This example illustrates how MISTRESS can be used to find imputations for those unknown values.

The analysis sample consists of 7332 individuals. For each person, we have scores on five labour conditions. These are labelled *dirty* (**D**), *heavy* (**H**), *risky* (**R**), *stench* (**S**) and *noise* (**N**). Each subject responded whether the attribute was applicable to his, or her, job. For a subgroup of 5750 people we also know the type of job, classified into 7 categories: *management* (**MAN**), *administrative* (**ADM**), *commercial* (**COM**), *scientific* (**SCI**), *service* (**SER**) *agrarian* (**AGR**) and *industrial* (**IND**). The classification by profession is missing for 7332 – 5750 = 1582 observations. The results for single imputation, ordered by donor scores, are presented in Table 3.4.

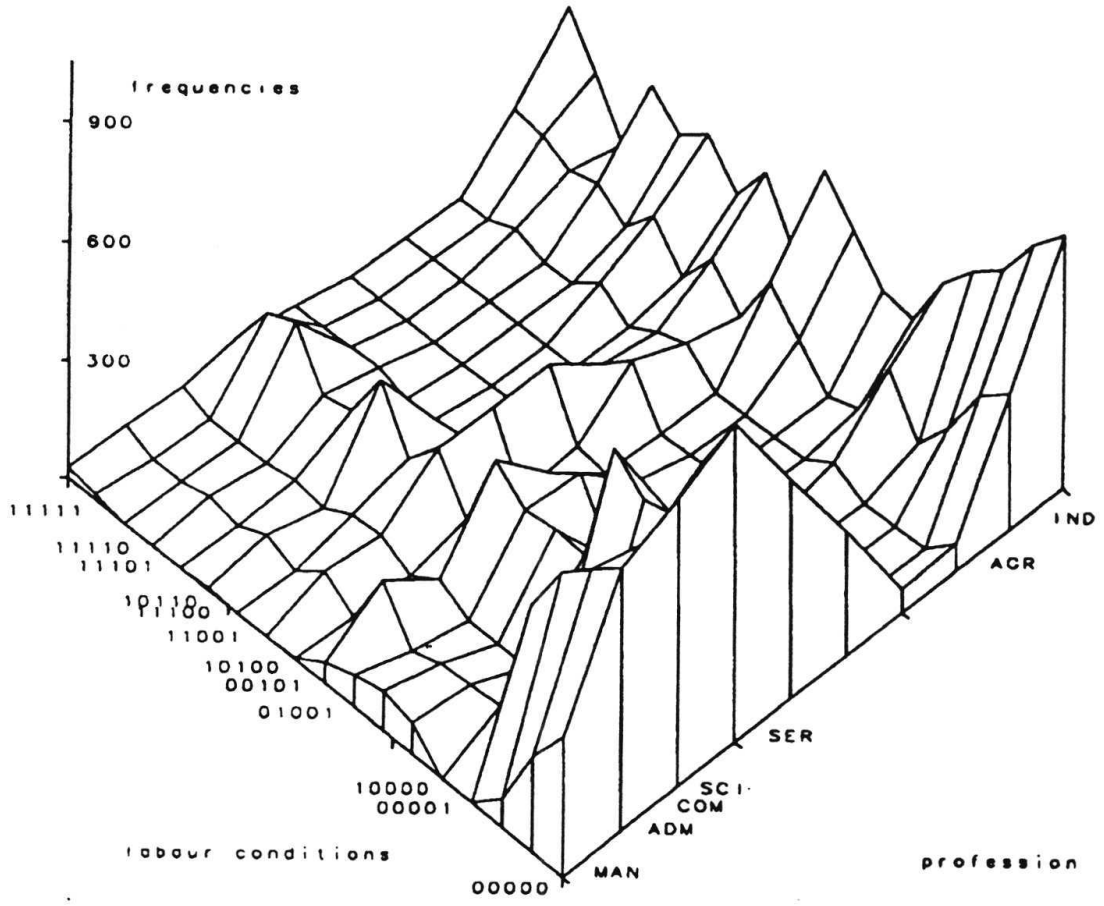
Table 3.4 Single Imputation LSO Table (**FAT** = Imputation)

Labour Conditions					Professional Category												Donor Score
DHRSM					**MAN**		**ADM**	**COM**	**SCI**	**SER**		**AGR**		**IND**		
1	1	1	1	1	1		1	1	6	2		5		64	**19**	3.18
1	1	1	1	0	0		0	0	6	3		7		11	**10**	2.68
1	0	1	1	1	1		1	0	3	3		1		21	**6**	2.58
0	1	1	1	1	0		0	0	1	1		1		3		2.48
1	1	1	0	1	0		1	1	1	1		3		61	**23**	2.47
1	1	0	1	1	0		1	1	**4**	6		**4**		50	**15**	2.41
1	0	1	1	0	0		1	2	1	0		2		8	**5**	2.08
0	1	1	1	0	0		1	0	0	0		1		2	**1**	1.97
1	1	1	0	0	1		1	0	9	2		9		51	**22**	1.96
1	1	0	1	0	0		1	1	9	2		20		13	**20**	1.90
0	0	1	1	1	0		3	2	2	0		0		12	**4**	1.88
1	0	1	0	1	0		0	2	1	1		1		20	**12**	1.87
1	0	0	1	1	4		3	2	6	3		2		46	**32**	1.81
0	1	1	0	1	0		0	1	2	1		0		8	**7**	1.76
1	1	0	0	1	2		6	**4**	6	9		12		88	**32**	1.70
0	1	0	1	1	0		0	0	1	0		0		5	**2**	1.70
0	0	1	1	0	0		0	0	1	1		0		2	**3**	1.38
1	0	1	0	0	0		1	0	1	4		3		14	**10**	1.37
1	0	0	1	0	0		2	1	2	2		13		17	**10**	1.31
0	1	1	0	0	0		0	3	6	0		1		10	**9**	1.26
0	1	0	1	0	0		1	0	3	4		0		4	**3**	1.20
1	1	0	0	0	2		6	16	21	38		81		95	**81**	1.19
0	0	1	0	1	1		16	3	10	6		2		15	**14**	1.17
0	0	0	1	1	3		19	6	16	6		0	**29**	28		1.00
1	0	0	0	1	8		11	6	20	14		10	**48**	103		0.99
0	1	0	0	1	2		4	12	19	21		4	**16**	40		0.89
0	0	1	0	0	3		5	7	27	12	**19**	2		29		0.16
1	0	0	0	0	4		15	28	32	25	**96**	60		104		0.09
0	0	0	1	0	6		22	3	27	8	**25**	1		17		0.09
0	1	0	0	0	2		12	58	115	87	**71**	16		80		-0.02
0	0	0	0	1	21		133	40	132	54	**122**	3		125		-0.11
0	0	0	0	0	157	**816**	843	373	916	349		54		324		-0.91
					218	**816**	1100	573	1406	665	**333**	318	**93**	1470	**340**	

The majority of employees does not work in any of the disturbing circumstances. Most discomfort is experienced by blue collar workers like labourers, farmers and service personnel. All workers experiencing at least three or more adverse conditions are assigned to the group of industrial workers. So, under maximal consistency, we expect that people with many job-related harassments are labourers. Three out of 10 incomplete profiles with 2 annoyance scores are assigned to farmers. The 816 persons working in a clean environment are all assigned to the management group. This is done because this group is by far the most outspoken group.

The analysis shows that it is possible to find categorical imputations such that the major trend in the data is extrapolated. Clearly, labour conditions are consistent with the type of work people do. This relationship is automatically taken into account when searching for maximally homogeneous imputation.

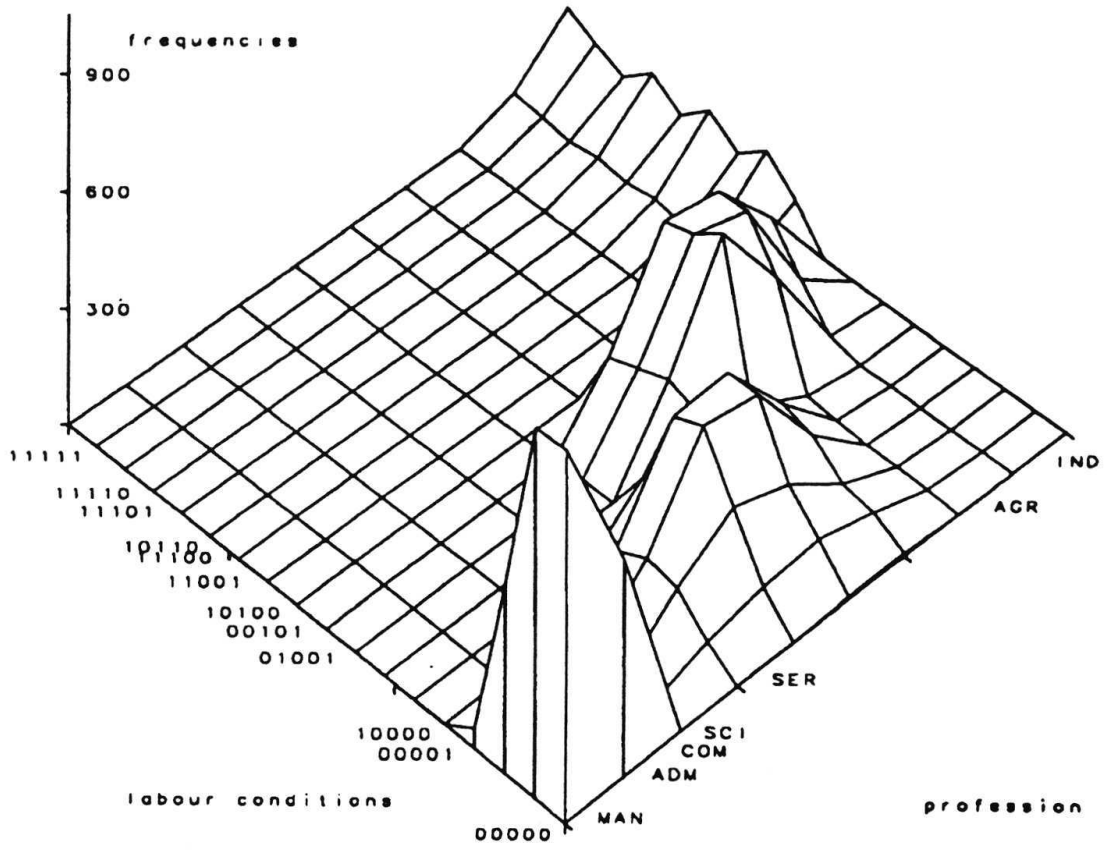
Figure 3.2 Observed frequencies (Z) versus job classes (X) versus labour conditions (Y)



The frequencies of the observed data are also pictured in Figure 3.2 by a slightly smoothed graphical analogue of the cross-tabulation in Table 3.4. The plot shows job classes on the *X*-axis, labour conditions on the *Y*-axis, and vertically on the *Z*-axis the frequencies are shown. The job classes and nuisance patterns are scaled by the consistency maximizing scores obtained by MISTRESS, with blue collar jobs relatively close together on one side and well separated from white collar jobs on the other side of the *X*-axis. The interpretation is that, based on nuisance patterns, we have two homogeneous subgroups of jobs: blue collar and white collar jobs. A similar reasoning is to be applied to labour conditions, although they do not fall apart into two groups. The conditions with few or none nuisance parameters are somewhat separated from the rest on the *Y*-axis. A consistent subset of white collar jobs experiencing hardly any nuisance in labour conditions is thus located in the lower corner pointing towards us. An intuitive

interpretation of the most consistent imputation is that it should disfigure the landscape in Figure 3.2 as little as possible. Like in Figure 3.2, we can picture the frequencies of the imputed data. See Figure 3.3.

Figure 3.3 Single imputation frequencies (Z) versus job classes (X) versus labour conditions (Y)



The imputations follow a curved and peaked range of frequencies from the origin {*white collar, no nuisance*} up to the far upper corner {*blue collar, maximal nuisance*}. The albeit `reasonable' imputations are nevertheless very `single'. All missing data with the non-nuisance pattern are singularly attributed to managers. This is a bit peculiar since other white collar workers are also reasonable candidates.

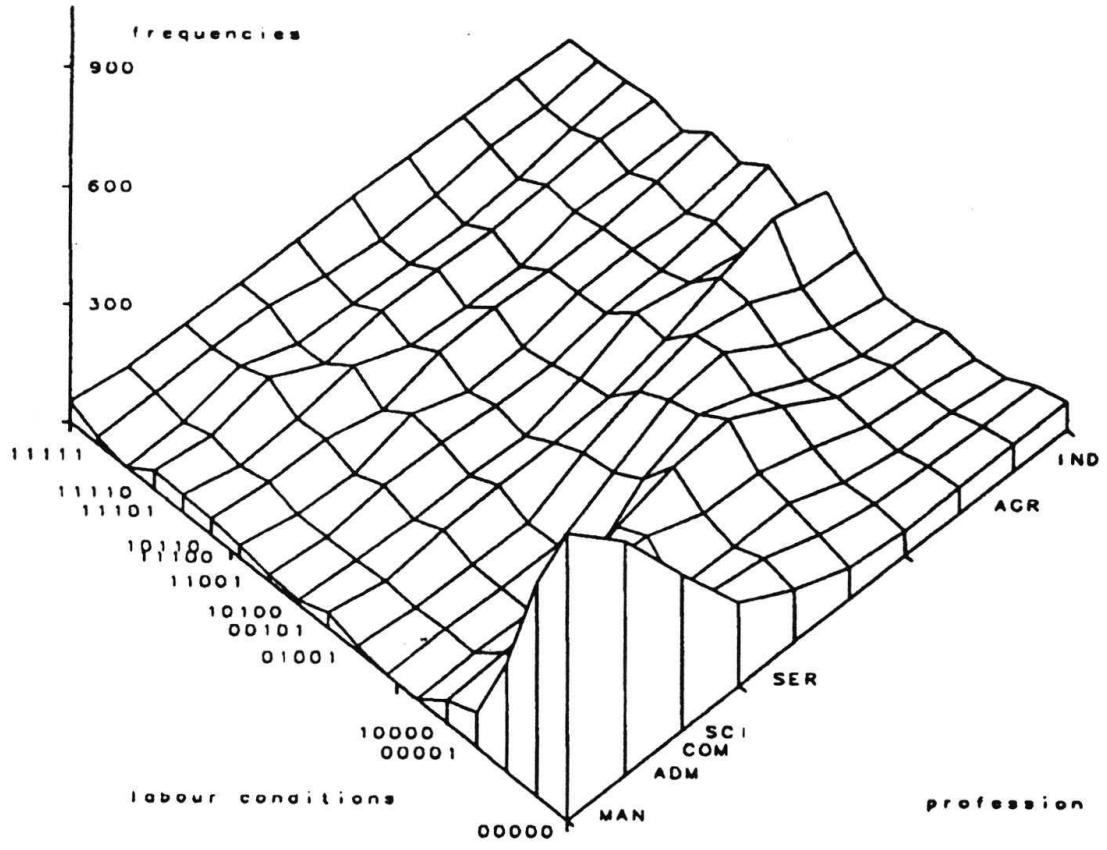
The latter observation leads automatically to the possibility of multiple imputation, where the mass of frequencies is more equally spread over other reasonable candidates. The

data are completed five times by drawing imputations randomly from the predictive distributions based on inverse distances, defined in the preceding section. Although we still use the rather heuristic definition of a predictive distribution, in this example multiple imputation seems to work quite well. The multiple imputations are shown in Table 3.5 and in Figure 3.4.

Table 3.5 Multiple Imputation LSO Table (**FAT** = Imputation)

Labour Conditions					Professional Category												Donor Score
DHRSM					**MAN**		**ADM**		**COM**		**SCI**		**SER**		**AGR**		**IND**		
1	1	1	1	1	1	**2**	1	**1**	1	**2**	6	**1**	2	**2**	5	**5**	64	**5**	3.18
1	1	1	1	0	0		0	**1**	0	**1**	6		3		7	**2**	11	**5**	2.68
1	0	1	1	1	1	**1**	1		0		3		3		1	**1**	21	**3**	2.58
0	1	1	1	1	0		0		0		1		1		1		3		2.48
1	1	1	0	1	0	**2**	1	**2**	1	**3**	1	**1**	1	**2**	3	**6**	61	**7**	2.47
1	1	0	1	1	0	**2**	1		1	**2**	4	**1**	6	**1**	4	**3**	50	**6**	2.41
1	0	1	1	0	0		1		2			**1**	0		2	**2**	8	**2**	2.08
0	1	1	1	0	0		1		0		0		0		1		2		1.97
1	1	1	0	0	1	**1**	1	**1**	0	**2**	9	**1**	2	**2**	9	**5**	51	**11**	1.96
1	1	0	1	0	0		1		1	**1**	9	**1**	2	**1**	20	**5**	13	**12**	1.90
0	0	1	1	1	0		3		2		2		0		0	**1**	12	**2**	1.88
1	0	1	0	1	0		0		2		1	**1**	1	**1**	1	**4**	20	**7**	1.87
1	0	0	1	1	4	**1**	3	**2**	2	**1**	6	**1**	3	**1**	2	**7**	46	**18**	1.81
0	1	1	0	1	0		0		1		2	**1**	1		0	**2**	8	**4**	1.76
1	1	0	0	1	2	**1**	6	**1**	4	**1**	6	**1**	9	**2**	12	**9**	88	**17**	1.70
0	1	0	1	1	0		0		0		1		0		0		5	**1**	1.70
0	0	1	1	0	0		0		0		1		1		0		2	**2**	1.38
1	0	1	0	0	0		1		0		1		4		3	**3**	14	**7**	1.37
1	0	0	1	0	0		2		1		2		2		13	**2**	17	**7**	1.31
0	1	1	0	0	0		0		3		6		0		1	**3**	10	**6**	1.26
0	1	0	1	0	0		1		0		3		4		0	**1**	4	**2**	1.20
1	1	0	0	0	2	**1**	6	**1**	16		21	**1**	38	**2**	81	**39**	95	**37**	1.19
0	0	1	0	1	1		16		3		10		6		2	**7**	15	**6**	1.17
0	0	0	1	1	3		19		6		16		6	**1**	0	**14**	28	**14**	1.00
1	0	0	0	1	8	**1**	11		6	**1**	20	**1**	14	**1**	10	**22**	103	**22**	0.99
0	1	0	0	1	2		4		12		19		21	**1**	4	**10**	40	**5**	0.89
0	0	1	0	0	3		5	**1**	7	**3**	27	**2**	12	**12**	2	**1**	29		0.16
1	0	0	0	0	4	**4**	15	**7**	28	**14**	32	**13**	25	**51**	60	**4**	104	**2**	0.09
0	0	0	1	0	6	**1**	22	**3**	3	**3**	27	**4**	8	**13**	1	**1**	17		0.09
0	1	0	0	0	2	**3**	12	**6**	58	**11**	115	**15**	87	**32**	16	**2**	80	**1**	-0.02
0	0	0	0	1	21	**6**	133	**11**	40	**28**	132	**31**	54	**42**	3	**3**	125	**1**	-0.11
0	0	0	0	0	157	**314**	843	**276**	373	**98**	916	**86**	349	**32**	54	**5**	324	**5**	-0.91
					218	**341**	1100	**314**	573	**172**	1406	**164**	665	**200**	318	**171**	1470	**220**	

Figure 3.4 Average frequencies based on multiple imputation (Z) versus job classes (X) versus labour conditions (Y)



The listed imputation frequencies are the average over five multiple imputation. Because of rounding errors, not all imputations exactly add up to the marginal frequencies. Comparing Figures 3.3 and 3.4, it is obvious that the multiple imputations are more spread over jobs and nuisance patterns. Both imputations, single and multiple, follow the same gradient from {*white collar, no nuisance*} on the bottom to {*blue collar, maximal nuisance*} on top.

3.7 CONTINGENCY TABLES

This example compares some aspects of the treatment of missing data in loglinear analysis to the present method. We use the 2 × 2 × 2 table given in Little and Rubin

(1987, p. 187). The data pertain to a partly real life, partly artificial example made up by Little and Rubin. There exist three dichotomized variables: *survival* (**S**), *type of clinic* (**C**) and *amount of prenatal care* (**P**). Type of clinic is unknown for 255 observations (= 26%), which means that 8.8% of the observations in the three-way table is missing. Table 3.6 indicates that the preferred loglinear model for the table based on the 715 complete observations is [**SC, PC**], which means that type of clinic is related to survival and to the amount of prenatal care. Moreover, within the same clinic, survival and prenatal care are not related. Because deletion of the association [**SP**] does not alter the fit, the more parsimonious model [**SC, PC**] is preferred. Model [**SP, SC**] does not fit at all.

Table 3.6 Chi-square and p-values under EM and MISTRESS imputation

Model	Completely Classified	p	Imputation by EM	p	Imputation by MISTRESS	p
[**SP, SC, PC**]	0.044	0.834	0.057	0.810	4.77	0.029
[**SC, PC**]	0.083	0.959	0.031	0.984	9.76	0.008
[**SP, SC**]	169.469	0.000	0.002	0.999	355.16	0.000

The second pair of columns in Table 3.6 contains the χ^2-values that measure the difference between the expected values under the three loglinear models and the imputed contingency tables (cf. Little and Rubin, 1987, p. 190-191). These values are not statistically significant, so all models fit the data. This is caused by, amongst others, the fact that the EM algorithm finds the most favourable imputations *given* the specific loglinear model. In general, loglinear models will fit better as more missing observations are added. In most cases, this will preserve-and even emphasize-the structure among variables as described by the loglinear model.

However, things can also go less well. Observe that model [**SP, SC**] now fits the imputed table (p = 0.999). For the completely classified table this model does not fit at all (p= 0.000), so filling in missing data brought about some real change. But this is a hazardous aspect of EM: suppose that we really had the 255 missing observations as in Little and Rubin, and that we applied EM. Then, we would have been pleased to find a χ^2-value as low as 0.002, and we would have had little reason to question the validity our model. If we compute correlations we see what has happened: the original correlation - actually

a ϕ - coefficient here - of the omitted [**PC**] effect is equal to -0.4924, which is substantial. After EM, it is -0.0130! Imputation corrupted the correlation. Rubin (personal communication) shows that in this case multiple or single imputation makes no substantial difference in estimation of the model parameters for the loglinear models. The same catch, though in the opposite direction, holds for MISTRESS. Because MISTRESS optimizes a different, almost reverse criterion, the imputed tables do not fit the loglinear model as well as the ones produces by EM. None of models fit to the imputed data. On the other hand, the χ^2-statistic clearly signals the important [**PC**] interaction.

Both the EM algorithm and our method have the same basic weakness: *if the model is wrong, imputations will be wrong*. If the model prescribes that a certain interaction does not exist, then EM will do everything to make this true. In the above case, it makes a correlation of 0.49 disappear. Analogously, MISTRESS overemphasizes tiny correlations. Generally speaking, loglinear analysis stresses absence of particular interaction, while maximizing consistency emphasizes presence of overall interaction. If we suspect that the analysis results are heavily biased by imputation, we should use these properties to our advantage.

3.8 CONCLUSION

The technique proposed in this paper is fairly simple. For categorical data, it is a way of selecting the proper category to be imputed. In the context of maximizing consistency, this seems to be new. The method optimizes a well-defined and widespread criterion. In addition, it is fast, flexible and of high practical value. Few assumptions are needed. The method stays close to the data.

It is possible to simulate various hot deck strategies. For example, by (over)weighting one of the variables we simulate a single donor variable. The method then evolves into a traditional hot deck method. In the same way, it is possible to rule out specific variables from the donor. Careful selection of variables may drastically improve the quality of the solution. Non-ignorable models, in which the pattern of nonresponse depends on the values of the data, can be evaluated by adding an indicator variable for

the nonresponse distribution for each variable. Mixes of continuous and discrete data can also be analyzed. Since imputations are determined for each variable separately, mixing does not present any new problems.

There are also situations in which the method will perform less satisfactorily. The main concern is the amount of intercorrelation. If the magnitude of all correlations is below 0.20 then the method may generate imputations that overemphasize small correlations. In this case, random imputation or unconditional mean imputation often work better. It seems preferable to use MISTRESS here only in combination with a resampling method, like the bootstrap, in order to estimate the variability of consistency. Van Buuren and van Rijckevorsel (1992a, 1992b) found that the average correlation should be at least 0.50 before the method becomes practical. At that point, the technique gives reasonable results up to 10-15% missing data.

A second cautionary note concerns imputation itself. However attractive the idea may seem, we must never forget that once after we have completed the data, they are partly artificial. The main pitfall is to analyze the filled-in data as if they were real, and thus overstate precision. The sagacious researcher will set up a subconscious alert that signals any peculiarities that might result from imputation. According to Dempster and Rubin (1983), the entire idea of imputation carries one great seductive danger: "it can lull the user into the pleasurable state of believing that the data are complete after all."

We conclude with some words on applications and perspectives. The number of variables or observations hardly influences the computational efficiency of the method. Therefore, the technique can be used with large data matrices. The main application field of MISTRESS is the analysis of surveys, rating scales and questionnaires. Furthermore, the relationship with Cronbach's α makes it attractive for dealing with missing data in psychological testing. It is easy and cheap to reiterate MISTRESS in any fashion, combined with bootstrapping, multiple imputation and the like.

The most spectacular application of the method is multiple imputation. It is ambiguous to call it an application. As a matter of fact one might devote another study to MISTRESS and multiple imputation. Analytical problem number one is to decide how implicit predictive densities are to be derived within the present framework. To this purpose, we mentioned in this paper only some intuitive possibilities. This of course needs further study. It is particularly interesting to examine the shape of the predictive

density function under varying levels of consistency. If the consistency equals zero, a trivial possibility, the predictive distribution should be uniform. Conversely, if the consistency approaches unity, it should have zero variance. And then, we are back at MISTRESS.

REFERENCES

BUCK, S.F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. J. R. Statis. Soc. B22 (1960) 302-306.

BUUREN, VAN, S. & W.J. HEISER. Clustering *n* objects into *k* groups under optimal scaling of variables. Psychometrika 54 (1989) 699-706.

BUUREN, VAN, S. & J.L.A. VAN RIJCKEVORSEL. Fast linear least squares imputation of missing data. RR-01-91, Leiden, University of Leiden, Dept. of Psychology, 1991.

BUUREN, VAN, S. & J.L.A. VAN RIJCKEVORSEL. Imputation of missing categorical data by maximizing internal consistency. Psychometrika 1992A. (to appear)

BUUREN, VAN, S. & J.L.A. VAN RIJCKEVORSEL. Data augmentation and optimal scaling. To appear in: R. Steyer & K.F. Wender (eds) Proceedings of the 7th European Meeting of the Psychometric Society, July 29-31, Trier. Heidelberg, Springer-Verlag, 1992b.

CRONBACH, L.J. Coefficient alpha and the internal structure of tests. Psychometrika 16 (1951) 297-334.

DEAR, R.E. A principal component missing data method for multiple regression models, SP-86. Santa Monica Cal., System Development Corporation, 1959.

DEMPSTER, A.P. & D.B. RUBIN. Overview. In: W.G. Madow, I. Olkin & D.B. Rubin (eds) Incomplete data in sample surveys, vol. 2, Theory and annotated bibliography. New York,
Academic Press, 1983. pp. 3-10.

DODGE, Y. Analysis of experiments with missing data. New York, Wiley, 1985.

FORD, B.L. An overview of hot deck procedures. In: W.G. Madow, I. Olkin & D.B. Rubin (eds), Incomplete data in sample surveys, vol. 2, Theory and annotated bibliography. New York, Academic Press,
1983.

GIFI, A. Nonlinear multivariate analysis. Chichester, Wiley, 1990.

GLEASON, T.C. & R. STAELIN. A proposal for handling missing data. Psychometrika 40 (1975) 229-252

GREENACRE, M.J. Theory and applications of correspondence analysis. New York, Academic Press, 1984.

HARTLEY, H.O. & R.R. HOCKING. The analysis of incomplete data. Biometrics 27 (1971) 783-808.

HEDGES, B. & I. OLKIN. Selected annotated bibliography. In: W.G. Madow, I. Olkin & D.B. Rubin (eds) Incomplete data in sample surveys, Vol. 2, Theory and annotated bibliography. New York, Academic Press, 1983. pp. 3-10.

KALTON, G. & D. KASPRZYK. Imputing for missing survey responses. Proceedings of the Section of Survey Research Methods, 1982, American Statistical Association, 1982. pp. 22-33.

LITTLE, R.J.A. & D.B. RUBIN. Statistical analysis with missing data. New York, Wiley, 1987.

LITTLE, R.J.A. & D.B. RUBIN. The analysis of social science data with missing values. In: J. Fox & T. Scott Long (eds) Modern methods of data analysis. London, Sage, 1990. pp. 374-409.

LORD, F.M. Some relations between Guttman's principal components of scale analysis and other psychometric theory. Psychometrika 23 (1958) 291-296.

MADOW W.G., OLKIN I. & D.B. RUBIN (eds). Incomplete data in sample surveys, Vol. 1, 2 and 3. New York, Academic Press, 1983.

MEULMAN, J. Homogeneity analysis of incomplete data. Leiden, DSWO Press, 1982.

NISHISATO, S. Analysis of categorical data: Dual scaling and its applications. Toronto, University of Toronto Press, 1980.

RUBIN, D.B. Multiple imputation for nonresponse in surveys. New York, Wiley, 1987.

RUBIN, D.B. EM and beyond. Psychometrika 56 (1991) 241-254.

RUBIN, D.B. & N. SCHENKER. Multiple imputation in health-care databases: An overview and some applications. Stat. Med. 10 (1991) 585-589.

SCHNELL, R. Missing Data Probleme in der empirischen Sozialforschung [Missing data problems in the social sciences]. Inaugural dissertation, Bochum, Ruhr Universität, 1986.

BUILDING STATISTICAL DATABASES FOR AN AIDS SURVEY

A.G.C. Vogels*, M.M. van der Klaauw**

* Department of Children and Health, TNO Institute of Preventive Health Care, Leiden, The Netherlands

** Department of Statistics and Computer Science, TNO Institute of Preventive Health Care, Leiden, The Netherlands

Abstract

The study 'Health, Behaviour and Relations among Adolescents' was designed to provide for a strong empirical base for adequate health education with regard to sexually transmitted diseases in general and AIDS in particular; secondly, it should provide a better estimation of the risks for adolescents of HIV-contagion.
In this study, data on sexual behaviour, attitudes and knowledge were collected by means of a questionnaire that is to be completed in the classroom.

The process of data collection has been complicated by several factors. The study aimed to generalize over sex, age, type of education, religious orientation of schools involved and social economic status, for 6 distinct Dutch regions as well as for the country as a whole. Several relevant variables (e.g. homosexuality and sexual experience) were expected to have little variation; to ensure a sufficient number of observations a large sample (n=12.000) had to be approached. Data collection required cooperation of school administrators. Because of the delicate content of the questionnaire a rather high refusal rate was expected related to religious orientation of schools. The data had to be collected in cooperation with more than 45 different organizations (the Sentinel Stations for Youth Health Care); data collection had to be completed within 3 months.
To ensure a controlled collection of data on all relevant variables, several measures had to be taken. These included APRI, a Automatic Project Administration System, by means of which the representativeness of the sample was safeguarded.