Kwantitatieve Methoden (1999), *62,* 123-138.

A REVIEW OF CURRENT SOFTWARE FOR HANDLING MISSING DATA

Joop J. Hox[1]

ABSTRACT.

When we deal with a large data set with missing data, we have to undertake two important tasks. First, it is important to inspect the pattern of missingness. This can provide very practical information. For instance, we may find that most of the missing values concern only one specific variable. If this variable is not central to our analysis problem, we may delete it from our analysis, rather than keeping it, and therefore having to delete many cases. We would also like to know if the missingness forms a pattern, or if it is related to some of our observed variables. For if we discover a system in the missingness pattern, we may try to include that in our statistical analyses.

The second task is to produce sound estimates of the parameters of interest, despite the incompleteness of the data. There are two major approaches to this problem. One is to make the data complete by imputing the missing values, and then do the analysis on the completed data. The other is to use a method, typically a likelihood-based procedure, that allows us to model incomplete data directly.

Modern software assists us in both tasks. For inspecting the pattern of missingness, either we can use the standard procedures in a statistical package like SPSS, or the specialized procedures made available in the SPSS procedure Missing Value Analysis (MVA). For imputation and direct modeling of incomplete data, we need specialized software. This contribution reviews some of the options in generally available software like SPSS MVA, SOLAS, and NORM.

1. INTRODUCTION

Data-analysts who find themselves confronted with 'missing data' need to go through an analysis strategy that consists of two steps. The first step is to investigate the pattern of missingness, to get some idea of the process that could have generated the missing data. The second step is to analyze the data, despite the fact that the data set is incomplete. There are several strategies to do the latter. Typical possibilities are: delete cases with missing data, impute

[1] Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University. Email j.hox@fss.uu.nl

the missing data, use multiple imputation to preserve the error, or use some method that directly estimates the parameters of interest on the incomplete data matrix. Recently, software has become available that assists us in coping with missing data. In this article, I will review some of the capabilities available in the SPSS package and two stand-alone programs: NORM and SOLAS. Programs that use direct estimation on the incomplete data matrix will be mentioned only briefly.

To examine the capabilities of these programs, I will use an example data file. This file holds longitudinal data with five time points labeled MT1 to MT5, and one explanatory variable labeled 'sex'. There are 40 cases. The data have been made incomplete using a process that mimics panel dropout. The dropout mechanism is that at each time point after the first, about 5% of the respondents leave the panel, with panel dropout weakly related to having a low score on the previous occasion. In addition to the panel dropout, some data points were made missing by a completely random mechanism. There are no missing values on the variable sex. Since both the missingness process and the observed values are known for all missing data points, we can check how well the various approaches are performing.[1]

SPSS and SOLAS are commercial products. Some other programs I mention are freely available as shareware. In an appendix preceding the Reference section I will provide the relevant Internet locations.

2. EXPLORING THE MISSINGNESS PATTERN

Some basic distinctions concerning missing data are data that are Missing Completely At Random (MCAR), Missing At Random (MAR), or Not Missing At Random (NMAR). The precise meaning of these terms is subtle (cf. Little and Rubin, 1987). For the purpose of this review, MCAR means that the mechanism that produces the missing data is not related to any of the variables (observed or latent) in the analysis. MAR means that the missingness depends on some of the variables in the analysis, but conditional on those variables, is not related to the value that should have been observed for that data point. NMAR means that the missingness does, in fact, depend on the value that should have been observed for that data point. These

[1] Both data files LONGCOMP and LONGMIS are available from the author as portable SPSS files.

distinctions are important, because data that are MCAR will produce unbiased estimates even with rather primitive analysis methods. Data that are MAR will produce unbiased estimates, if a model and estimation technique is used that renders the missingness mechanism *ignorable*. When data are NMAR, an analysis method must be used that includes both a model for the observed data, and a model for the missingness mechanism. For data that are MCAR or MAR, general modeling software is available, that produces unbiased estimates using all the available information. For data that are NMAR, there are no easy solutions.

Exploring the pattern of missingness is important for a variety of reasons. Firstly, when we find that most missing data are concentrated in a few cases or in a few variables, one option is to remove those cases or variables from the analysis. Especially when we have a number of similar variables, such as responses to the items of a multi-item scale, we do not lose much if we omit one variable. The second reason is that data that are MAR do not produce unbiased estimates automatically. We must use the right estimation method, but we also must include the variables that predict the missingness in our analysis. Therefore, we must have a good picture of the patterns of missingness in our data.

The SPSS module for Missing Values Analysis (MVA) includes a variety of techniques to inspect the pattern of missing data. It can display the pattern of missingness grouped for all cases, or individually for all cases that have missings. Missings can also be displayed sorted by the value of another variable, such as sex in our example. Figure 1 below shows the output of SPSS MVA for the longitudinal data.

The MVA display shows that we have 17 cases with a pattern of no missing data at all. There are five cases with MT5 missing, two with both MT4 and MT5 missing, and so on. The column labeled ‘complete if…’ informs us how many complete cases we will have if we remove the variables included in the missingness pattern from the analysis. If we omit variable MT5 from the analysis, we have 22 complete cases. If we omit both MT4 and MT5 from the analysis, we have 26 complete cases: the 2 that are incomplete on both MT4 and MT5 plus the 5 that are incomplete on MT5 only. The output also presents the means on the MT variables for all patterns of missingness. Inspection of these means can lead to valuable clues about the missingness pattern.

Tabulated Patterns

Number of Cases	Missing Patterns [a] SEX	MT1	MT2	MT3	MT4	MT5	Complete if ... [b]	MT1[c]	MT2[c]	MT3[c]	MT4[c]	MT5[c]
17							17	52.24	53.24	53.47	55.06	56.35
5						X	22	49.80	51.00	52.80	51.40	.
2					X	X	26	49.50	51.00	48.00	.	.
2					X		19	53.00	57.50	55.00	.	56.50
1		X					18	.	61.00	63.00	70.00	71.00
1			X				18	57.00	.	58.00	57.00	56.00
2				X			19	44.50	46.00	.	51.50	50.50
5				X	X	X	33	46.40	48.80	.	.	.
5			X	X	X	X	39	47.00	.	.	.	.

Figure 1. SPSS MVA output for missing data pattern

In our example, inspection of the missingness pattern suggests that if a case goes missing at a specific time point, there is a good chance that it will also be missing on subsequent occasions. This pattern indicates panel dropout in our data. An inspection of the means suggests that cases that fit the panel dropout pattern tend to have lower means, while other missingness patterns do not. For instance, the complete cases have a mean of 52.2 on variable MT1. The five cases in the last or next-to-last rows of the table show a clear dropout pattern, and the have a mean of 47.0 and 46.4 on MT1. The two cases in the third row, who are missing on MT4 (occasion 4) only, do not fit a dropout pattern, and they have a mean of 53.0 on MT1. Thus, careful inspection of the pattern in Figure 1 reveals in fact the two missingness mechanisms that have been put in these data.

Another useful display that SPSS can generate is a t-test for MCAR. Here, groups are defined by having or not having a missing value on a variable, and their means on other variables are compared by a t-test. The (partial) output in Figure 2 shows this test.

		MT1	MT2	MT3	MT4	MT5
MT1	t	.	.	.	.	.
	df	.	.	.	.	.
	# Present	39	33	27	25	22
	# Missing	0	1	1	1	1
	Mean(Present)	50.13	51.91	53.22	54.12	55.82
	Mean(Missing)	.	.	.	.	.
MT2	t	.8	.	.	.	.
	df	6.8	.	.	.	.
	# Present	33	34	27	25	22
	# Missing	6	0	1	1	1
	Mean(Present)	50.39	52.18	53.41	54.64	56.50
	Mean(Missing)	48.67	.	.	.	.
MT3	t	4.6	3.4	.	1.0	1.8
	df	27.1	13.7	.	1.2	1.2
	# Present	27	27	28	24	21
	# Missing	12	7	0	2	2
	Mean(Present)	51.81	53.26	53.57	55.00	57.05
	Mean(Missing)	46.33	48.00	.	51.50	50.50

Figure 2. Spss MVA output, comparing data points with and without missings by t-test.

The SPSS output in Figure 2 is close to unreadable. It is best read column by column. The first column after the table legend, under MT1, tells us that there are 39 cases present, for which we can compute a mean for MT1. Of the missing cases, there are none for which we can compute a mean for Mt1 (which is logical). In the second row, we see that for the cases present on MT2, there are 33 for which we can compute a mean on MT1, which is 50.39. For the cases missing on MT2, there are six for which we can compute the mean on MT1, which is 48.67. Similarly, for the cases present on MT3, there are 27 for which we can compute a mean on MT1, which is 51.81. For the cases missing on MT3, there are 12 for which we can compute the mean on MT1, which is 46.33. For MT4 and MT5 (not shown in the Figure) we have the same pattern. Thus, Figure 2 shows that cases that are missing on MT2 (occasion 2) have a lower mean on MT1 than complete cases. This pattern is stronger with MT3 (and also with MT4 and MT5, which are not in the Figure). This confirms our earlier suspicion, based on inspection of the missingness pattern in Figure 1, that panel dropout is influenced by the score; cases with a low score tend to drop out selectively.

It should be noted that all the analyses reported so far could also have been done using

standard procedures available in SPSS and other statistics packages. We must add to the data file a set of indicator variables, one for each substantive variable, which is 1 if that variable is missing, and 0 if it is not. Using these indicator variables, we can add some useful analyses, which are not included in SPSS MVA. For instance, we could do a principal components analysis of the indicator variables to look for multivariate patterns of missingness. Or we could recode the original variables, and treat the missing value code as a real value (categorize any continuous variables) by including the missing value code as an extra category. This can be followed by a homogeneity analysis or correspondence analysis (SPSS CATEGORIES: HOMALS) to look for relationships between patterns of missingness and real data values.

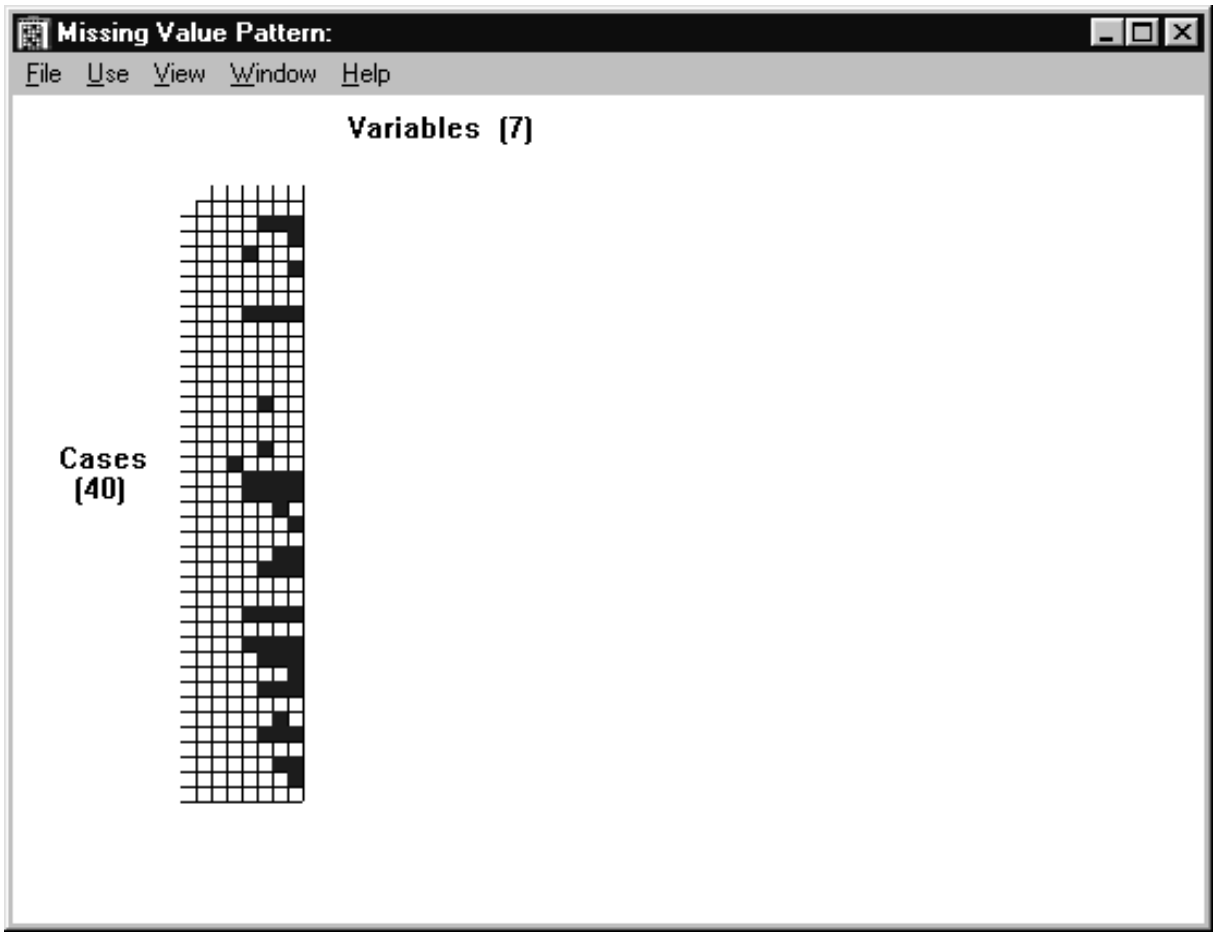

Figure 3. Missing data patterns as shown by SOLAS.

SOLAS does not include diagnostic tests for the pattern of missingness, but it is possible to inspect the data visually. The section of SOLAS output reproduced in Figure 3 shows this.

Although the graphic display in Figure 3 it is generally less informative than the various tests described above, the pattern displayed by SOLAS also suggests that the dominant missingness mechanism is panel dropout. Whenever a case has a missing value, there is a strong tendency that al the values to the right (which are the subsequent time points) are also missing.

3. SINGLE IMPUTATION

One way to cope with missing data is to fill the holes in the data set with plausible values. Many methods exist, which differ by how they define 'plausible'. The imputation methods can also be distinguished as being model based or parametric, as opposed to data based or nonparametric.

Parametric imputation methods include replacing the missing value by the variable's mean, or by the value predicted by a regression analysis on the available variables for that case. If a missing value is replaced by a value derived from information hat is not contained in the data matrix, we have *cold deck* imputation, which is also considered a model based method. An example of cold deck imputation is filling in the known population mean for each missing value, or assigning a zero to all missed items in an intelligence test. A model based technique called EM-estimation can also be used for parametric imputation. EM-estimation is a method that is often used to estimate a mean vector and a covariance matrix on incomplete data, for instance prior to a factor analysis. As such it is one of the direct estimation methods I discuss in a later section. However, once we have these means and covariances, we can use them to calculate multiple regression equations to predict the value of missing data points, using the variables that are present for each specific case. The EM method assumes that the data are MAR, while the other methods generally assume MCAR. Since the EM method also uses all the information in the available data, it is the most effective method to impute missing data points.

The EM method is available in SPSS MVA. Strangely enough, the EM procedure in SPSS produces a correlation matrix in a table, but this cannot be output on a file. However, the EM procedure can be used to impute the missing data points. The correlations can then be computed on the completed file, which is the same.[1] EM estimation is not available in SOLAS version 1, but

[1] After publication, I have received a note from Jeroen Pannekoek, of Statistics Netherlands, and David Nicholls, of SPSS Inc., who informed me that this is *not* the same. Imputing missing values using EM still underestimates the variance of the variables. <JH: note inserted in PDF version>

it has been promised for SOLAS version 2. EM estimation is also available in two free shareware programs: a DOS program called EMCOV by John Graham and a Windows program (or Splus module) called NORM by Joe Schafer. EMCOV takes incomplete data, and returns both the EM estimated covariance matrix, as a completed data set using single imputation. NORM returns the EM covariance matrix, and can be used for multiple imputation, which is treated in the next section. Mean substitution is available in standard SPSS (not in MVA, but in the Transformations menu). SOLAS has an option for either overall or group mean imputation, where missing values are imputed by the calculated from an appropriate subgroup. In our example data, we could use SOLAS to impute the mean separately for male and female respondents.

Nonparametric or data based imputation replaces the missing values by an appropriate value that is observed in the data set. By using a donor case to provide the imputed value, nonparametric imputation insures that the imputed value is a value that can actually exist. *Hot deck* imputation sorts the respondents and nonrespondents on a particular variable (or variable set if we are imputing multivariate data) into imputation classes. Missing values are then imputed, using randomly chosen observed values from donors that are in the same imputation class. To create imputation classes, we need auxiliary variables that are related to the missingness mechanism. Hot deck imputation is available in SOLAS. *Regression hot deck* is a specific variant of this method, which specifies imputation classes by predicting the missingness mechanism itself. It is not available in SPSS or SOLAS, but it exists in a limited form in the program PRELIS, which is a preprocessor for LISREL. Hot deck imputation assumes an ignorable missingness mechanism, which translates to MAR data with the relevant predictors included in the imputation procedure.

The program window in Figure 4 shows some of the SOLAS options for imputation. SOLAS allows the user to define series of variables as a single longitudinal variable. The option chosen is Last Value Carried Forward, which is for specifically for longitudinal data. It will impute a panel attrition mechanism by filling in the missing data points with the last observed value for that case. As O'Callaghan (1999) points out elsewhere in this issue, this method is generally not recommended.

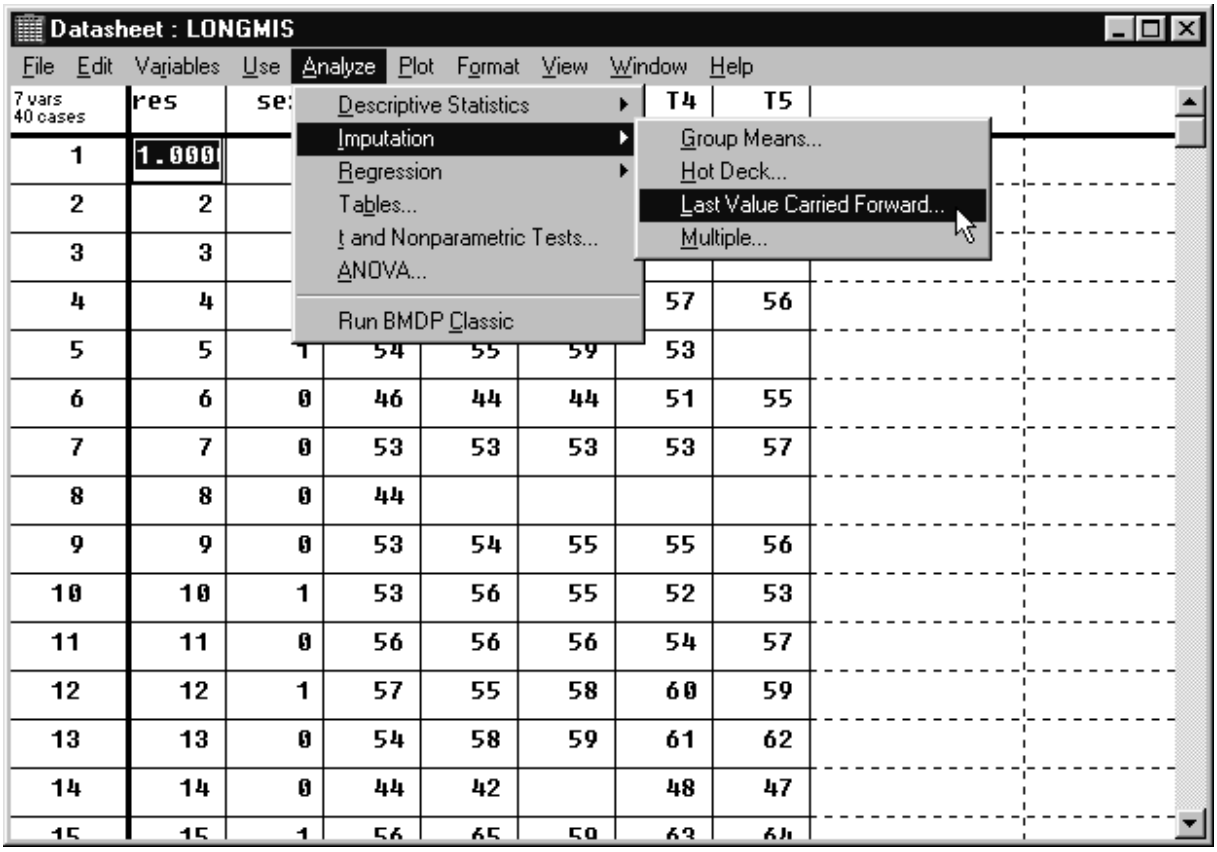

Figure 4. Solas 1.0 imputation options.

Most of the imputation methods used above will impute a value that is optimal, according to some criterion. As a result, they will generally underestimate the variance, and thus lead to biased significance tests. A remedy is to add a random error to the imputed value. This random error can come from a model or from the observed data. SPSS MVA includes regression imputation plus an added error term, which is either from a normal or Student distribution, or a randomly chosen residual from the complete cases. The hot deck method in SOLAS imputes observed values, and thereby includes the error term that is included in that observation. Since adding error terms or using hot deck incorporates a chance mechanism, the results will vary. Actually, when error is added to the imputed values, multiple imputation is the logical next step.

Table 1 below presents the means computed on the complete data set, and the means computed on the incomplete data, using various methods. It is clear that using only the complete cases, as in SPSS Listwise Deletion, produces biased estimates. The reason is that this assumes MCAR, while the missing information in our data is mostly MAR. The available cases method, known in SPSS as pairwise deletion, also assumes MCAR. Mean imputation is unbiased when

the data are MCAR, but it always decreases the variance. The hot deck procedure presented in Table 1, and the regression estimation are in fact MAR for these data. If we calculate the correlations between the five time-variables (not presented here), all methods perform rather poorly. It should be noted that hot deck, because it draws replacements at random from a class of similar complete cases, assumes good auxiliary variables and a large data set, neither of which we have in the example data.

Table 1. Means computed using different methods to analyze missing data

Means for:	T1	T2	T3	T4	T5
Complete data (no missings)	50.5	51.5	52.5	53.6	54.6
Complete cases	52.2	53.2	53.4	55.0	56.4
Available cases	50.1	52.2	53.6	54.7	56.5
Mean imputation	50.1	52.2	53.6	54.7	56.5
Regression imputation	50.2	52.0	52.7	54.3	55.7
Regression + error	50.3	51.8	52.6	54.0	55.6
Hot deck	50.1	51.4	52.2	53.7	54.4

4. MULTIPLE IMPUTATION

Multiple imputation means that the missing data are imputed a number of times, typically 3 to 5 times, with a different randomly chosen error term added in each imputation. The completed data sets are analyzed using standard methods, and the results are combined. The variance of the analysis results across the data sets provides an estimate of the imputation error (cf. Rubin, 1987). In multiple imputation, a major problem is to construct imputed data sets. Again, there is a parametric and a nonparametric approach. SPSS MVA has no provision for multiple imputation. The shareware program NORM by Schafer provides a parametric method, and SOLAS version 1.0 a nonparametric method.

In the parametric approach programmed in NORM, the Multiply Imputed (MI) data sets are simulated draws from a Bayesian predictive distribution of the missing data. This requires a model for the complete data, and properly adding uncertainty about both the missing values, and the parameters of the predictive distribution. NORM assumes that the data come from a multivariate normal distribution. Schafer (1996) describes similar procedures for other data

models, including categorical and mixed normal-categorical data, and panel data. For these data models, there are only Unix-based S-plus procedures, aptly called CAT, MIX, and PAN. Windows versions of these programs have been promised, but are not yet available. NORM uses a data augmentation algorithm to generate the imputed data. The Markov Chain Monte Carlo (MCMC) procedure NORM uses requires that the analyst makes two decisions. The first decision is, how many iterations are needed for the MCMC algorithm to converge on the correct predictive distribution. The second is, how many iterations are needed between imputations to achieve independent MI's. NORM provides diagnostic plots to assist in these decisions.

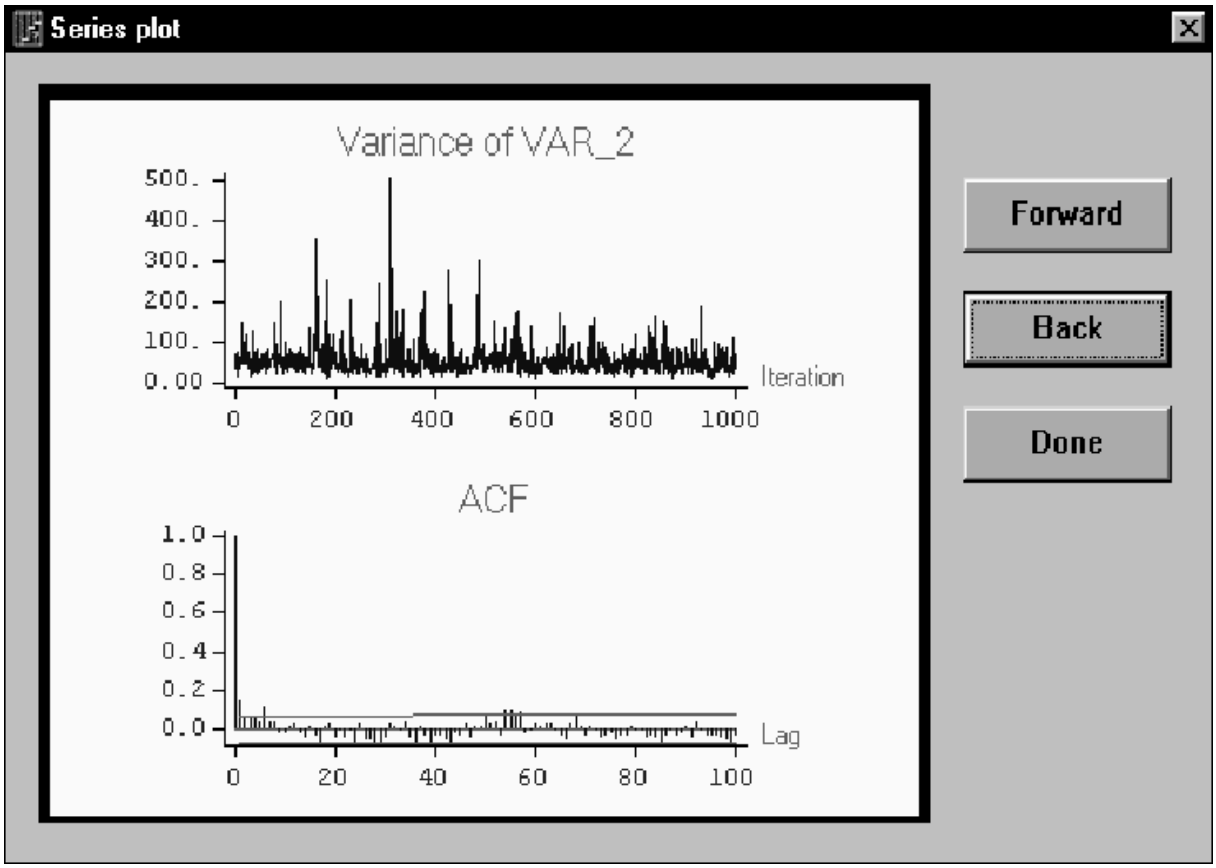

Figure 5. MCMC diagnostics plot from NORM.

The window in Figure 5 is an example of an MCMC diagnostic plot from NORM 1.0. It first shows a plot of the sequence of variances generated for one of the variables. What is needed here is absence of a pattern. From the plot, it would seem that after 500 iterations the process is

stable. The second plot is the autocorrelation function. It shows the correlations between data generated that are separated by 1,2, …, 100 iterations, with the 95% confidence limits. It would seem that about 50 iterations are enough to achieve independence here. NORM requires that we specify how many iterations the algorithm must go through before each imputation. Since we cannot be too generous here, we should choose at least 500 here. Since the program is actually rather fast, we prefer 1000 iterations here.

The nonparametric approach in SOLAS version 1.0 uses randomly chosen donors to provide the imputations. There are several randomization methods involved to properly include the uncertainty about both the missing data and the distribution from which replacements should be taken. First, a bootstrapped logistic regression is used to predict the nonresponse. Next, the regression equation is used to compute the propensity to have missing data for each case in the sample. The complete and incomplete cases are sorted into five subgroups based on their propensity scores. Finally, missing data are replaced by values taken from a randomly chosen donor in the same imputation class. This procedure is repeated a number of times, to generate multiple imputations. (For a more detailed description see O'Callaghan, 1999, in this issue.) The advantage of the nonparametric approach in SOLAS is that it does not require an assumption about a data model; the model for the missing data is provided by the observed data. It will also always generate imputed values that actually exist. A disadvantage is that the procedure assumes a reasonably large data set, for which Bernaards (1999) suggests a sample size of 500 cases.

Both NORM and SOLAS contain procedures to combine the results from the multiple analyses into a single set of results, including proper standard errors and significance tests. The current versions do not include procedures to combine the chi-squares or p-values from overall model tests, such as produced for example by LISREL. These procedures are described in Schafer (1996), and are promised for later versions of SOLAS.

Table 2. Combined resiults, using parametric multiple imputation

quant	estimate	std.err.	t	df	p	% missing info
r_{12}	0.78	0.17	4.61	625.5	0.000	6.0
r_{13}	0.76	0.17	4.51	760.6	0.000	5.4
r_{14}	0.64	0.21	3.08	14.4	0.009	44.5
r_{15}	0.63	0.22	2.90	11.2	0.015	50.3

Table 2 is adapted from the NORM output (SOLAS produces the same results in a different format). It shows the combined estimates from three imputed data sets for some selected correlations.

Note that the percentage missing information is not equal to the percentage missing data. This is logical, for if we have 50% missing values for a variable that is perfectly correlated with another variable, we actually have no missing information. The percentage missing information in Table 2 is based on the imputation error variance (cf. Schafer, 1996).

5. DIRECT ESTIMATION

When the data are MAR, two Likelihood based procedures are generally available to estimate a model directly on incomplete data: the *EM-method* and the *factored likelihood* approach. The EM-method derives its name from the two computation steps that are involved: an **E**xpectation step and a **M**aximization step. The expectation step computes the expected values for the sufficient statistics, given a model and values for the model parameters θ. The maximization step estimates the model parameters by maximizing the likelihood using standard procedures, given complete data. Thus, the **EM** algorithm, described loosely as follows. Fill the holes in the data with plausible start values, estimate the model parameters in θ on the completed data using standard ML procedures, estimate missing data again using the model and the current θ, and repeat until convergence (Dempster, Laird & Rubin, 1977). This is a form of direct estimation, which has compared to multiple imputation methods the disadvantage that we need a specialized program for each model, and that the ME method does not directly produce standard errors. ME is sometimes employed in a two-step procedure. For example, the ME method is used to estimate a correlation matrix on incomplete data, which is then analyzed using standard factor analysis. Used this way, it underestimates the variability of the data. As stated above, the EM method is available in SPSS MVA, promised for SOLAS version 2, and available in shareware form in EMCOV and NORM.

Factored likelihood is based on the principle of separating the likelihood function in different parts for different groups. Assume that the data Y have a probability function $f(Y/\vartheta)$, in which the ϑ are the model parameters, and the Likelihood function for the observed data is given by

$L(\vartheta/Y_{obs})$. For each distinct missingness pattern, we form a separate data group. This results in 1…G distinct groups. The likelihood function for all our data can then be written as $L(\vartheta/Y_{obs}) = L_1(\vartheta_1/Y_{1obs}) \; . \; L_2(\vartheta_2/Y_{2obs}) \; . \; . \; . \; L_G(\vartheta_G/Y_{Gobs})$, which gives the log-likelihood as $l(\vartheta/Y_{obs}) = l_1(\vartheta_1/Y_{1obs}) + l_2(\vartheta_2/Y_{2obs}) + \ldots + l_G(\vartheta_G/Y_{Gobs})$. In this formulation, the components of the log-likelihood each l_G have complete data for a subset of the variables. Since the total log-Likelihood is a simple additive function, we can use standard methods to maximize the Likelihood separately in each group, using the same model.

Factored likelihood is the principle that is involved when we subdivide a data set in different groups according to missingness pattern, and use the multigroup option in Lisrel or Eqs to estimate a structural model for all groups combined. If we have many missingness patterns, multigroup SEM is unwieldy, but some SEM programs (e.g., Amos, Mplus, Mx) allow ML estimation on directly on observed part of raw data. This is the identical to the multigroup solution, with each case as a different group, only such a model would not run in standard SEM software. Since regression analysis and tests on means are possible in SEM, using SEM software for direct estimation on missing data is a flexible tool that includes many simple analyses as a submodel of the general SEM. Table 3 below presents the results of using EM and direct likelihood using AMOS (cf. Arbucle, 1996) on the longitudinal data example.

Table 3. Results various estimation techniques, including EM and direct likelihood (DL)

Means for	T1	T2	T3	T4	T5
Complete data	50.5	51.5	52.5	53.6	54.6
Complete cases (Spss Listwise)	52.2	53.2	53.4	55.0	56.4
Hot deck (best single imputation)	50.1	51.4	52.2	53.7	54.4
NORM MI	50.4	51.8	52.0	53.4	54.3
SOLAS MI	50.1	52.1	53.6	55.0	57.2
EM + DL (identical)	50.4	51.9	52.2	53.7	54.6

The likelihood-based methods perform the best, which is to be expected. When the model assumptions are met, ML methods are efficient, and multiple imputation methods are asymptotically equivalent. With an infinite number of replications, multiple imputation produces

the same results as direct ML. In this example, the nonparametric method does not perform well, but the example data set is actually to small to expect good performance. The nonparametric imputation method in SOLAS is proper in the sense of Rubin (1987), and it should work well given a sufficient sample size.

6. CONCLUSIONS

SPSS MVA offers a convenient toolbox for examining the pattern of missingness in a data set. However, an analysis who is well versed in a different statistical package, can emulate most of these procedures by constructing for each variable a missingness indicator, which is coded 0=not missing, 1=missing. Using these indicator variables, most pattern analyses available in SPSS MVA can easily be emulated. The analysis part of MVA is rather limited. SPSS MVA relies on a number of parametric imputation techniques to make the data complete. One of the methods is based on the EM algorithm, which makes it a powerful method. However, there is no multiple imputation, which means that the variability of the completed data will be underestimated.

For multiple imputation, SOLAS offers a nonparametric method, based on the approximate Bayesian bootstrap (Rubin, 1987). Multiple imputation is easy with SOLAS, because it automatically generates several completed data sets for the multiple imputation, with the option to write out the data sets in a large variety of data formats (e.g., SPSS, SAS, Stata, BMDP, Excel). In fact, in buying SOLAS, one obtains a very useful file conversion tool in the bargain. Including parametric multiple imputation in version 2 should make it an even more versatile package.

For parametric multiple imputation, the programs made available by Schafer are extremely helpful. Being shareware, their price can of course not be beat. They are also well written, and supported by explanatory papers and presentation handouts, available from the website. The most important restriction is that the only program available as a stand-alone Windows program is NORM, which multiply imputes data assuming a multivariate distribution. Categorical and longitudinal data are not yet supported. SOLAS, which imputes data by taking them from a similar donor case, does not have this limitation.

7. INTERNET RESOURCES

Below are the websites where information can be viewed about the software reviewed here. Some of the websites, notably <www.methcenter.psu.edu> contain papers about missing value analysis and links to other Internet information. Some commercial sites have a demo-version of their software available.

Amos: www.smallwaters.com

Emcov and Norm: www.methcenter.psu.edu

Lisrel/Prelis: www.scicentral.com

Solas: www.statsol.ie

SPSS: www.spss.com

REFERENCES

Arbucle (1996). Full information estimation in the presence of missing data. In: G.A. Marcoulides & R.E. Schumacker (eds). *Advanced structural equation modeling*. Mahwah, NJ: Erlbaum.

Bernaards, C.A. (1999). SOLAS for missing data analysis. Software review. *Structural Equation Modeling*, 6, 301-304.

Dempster, Laird & Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B: Methodological*, 39, 1-38.

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

Little, R.J.A.& Rubin, D.B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, 18, 292-326.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J.L. (1996). *Analysis of incomplete multivariate data*. New York: Chapman and Hall.