



StatSoft®
Business White Paper

Handling Missing Data in *STATISTICA Data Miner*

Last Update: February 2, 2003

U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft South America
Czech Republic: StatSoft Czech Rep. s.r.o.
France: StatSoft France

Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
Israel: StatSoft Israel Ltd.
Italy: StatSoft Italia srl

Japan: StatSoft Japan Inc.
Korea: StatSoft Korea
Netherlands: StatSoft Benelux BV
Poland: StatSoft Polska Sp. z o. o.

Portugal: StatSoft Iberica Ltda.
Russia: StatSoft Russia
Singapore: StatSoft Singapore
S. Africa: StatSoft S. Africa (Pty) Ltd.

Spain: StatSoft Espana
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
UK: StatSoft Ltd.

Handling Missing Data

We have on numerous occasions investigated the issue of missing data, and alternatives to simple casewise deletion of missing data, and we are aware of the methods referenced in the paper that you forwarded to us (Acock, A. C. "Working with missing data"). There are a few of points we'd like to make:

1. Most of the specific algorithms referenced in the paper can easily be applied in *STATISTICA*.

1.1. You can recode missing data into valid data (codes), and then perform various analyses or generate graphs to see whether and how missing data cases are different from those that are observed.

1.2. You can use multiple regression or any other method that will produce predicted values from continuous and/or categorical predictors, to compute predicted values for cases that are missing (many modules of *STATISTICA* will label the sample of cases where no dependent variable observations are available the Prediction Sample).

2. We do not have an implementation of the EM algorithm, as described in the paper. Note that there are many questions that are difficult to answer, in general, when imputing missing data in this manner.

2.1. For example, is it reasonable to use an iterative application of a linear model to find replacements for missing data? In particular in data mining, most data (if not all data!) we see from areas other than social sciences (like the data presented in the paper you sent) are not normal, nor contain linear relationships.

2.2. As a simple example, even in the social sciences: If a person will not report his/her health status or income, do you think that those variables (for those persons) can reasonably be estimated as linear combinations of other variables? Having been involved in survey research quite a bit ourselves, I think that in most cases people won't respond if the response is deemed (by the respondent) to be uncomfortable, perhaps embarrassing, or unusual. (I remember for example discussing with researchers at the survey research center at the University of Michigan the problems involved in surveying "rich households"-- they won't tell you what they have, how they got it, or how they spend it; so all you have is missing data which are likely not well predicted by simple models from other variables).

2.3. The point here is that, for data mining purposes, where strong apriori models usually don't exist, using these methods (to impute missing data based on certain a-priori model assumptions) can be quite misleading.

3. SPSS has a designated Missing Data Analysis program, with the stated goal to "Improve the likelihood of finding statistically significant results" (direct quote from their brochure! This statement is a direct and obvious misunderstanding of what statistical significance is all about, i.e., it is NOT related to statistical power or effect size! It is quite amazing to see this in an SPSS brochure.)

4. In addition to offering many methods to estimate (predict) missing data (if you decide that's what you really want to do), STATISTICA offers tremendous flexibility regarding missing data:

4.1. There are missing data plots, to study the patterns of missing data.

4.2. You can quickly recode missing data into a valid code, and hence perform tests comparing missing data cases to those without missing data.

4.3. It is easy to specifically recode missing data using distribution functions, and so on.

5. The most useful way to handle missing data is to treat them in a way most meaningful in relation to the analysis at hand:

5.1. In most linear models, mean substitution is a good unbiased way to replace missing data, because a value at the mean does generally not contribute to covariances.

5.2. In other modules, you can simply ignore the issue of Missing Data; e.g., run Interactive Trees, and you will see that missing data are only excluded when necessary, i.e., if the respective variable is chosen for a split, and a substitute cannot be found (or none were requested). The same mechanism is used in Generalized Clustering, where missing data are only excluded for each specific variable, when estimating means or distribution parameters, i.e., all available info is used, and nothing is eliminated or substituted.

5.3. In Time Series, we have various methods for interpolating missing data, to create a smooth "transition" over "holes" in the time series.

5.4. In our upcoming implementations of support vector machines and naive Bayesian classifiers, we will also always use all valid data on a variable-by-variable basis, for example, to estimate the kernels for naive Bayesian.

6. We are somewhat more restrictive in our competitive evaluation of models project. If we permitted different numbers of valid cases for different analyses, the implementation of that project would be very complicated and inefficient (e.g., to generate the final summary spreadsheet with all predictions); nevertheless, and in practice, you can run individual analyses and take advantage of the typically method-specific types of procedures for dealing with missing data...

I hope we were able to address the questions you raised. At this point, we don't see any good way to deal with missing data in a general way, e.g., via simple imputation and linear models. Instead, missing data really need to be handled in a manner consistent with each method, to ensure unbiased results (e.g., we have an entire suite of survival analysis methods that will handle censored data, which are a "special kind" of missing data).