Diagnostics for Multivariate Imputations*

Kobi Abayomi†, Andrew Gelman‡, Marc Levy§

December 26, 2006

Abstract

We consider three sorts of diagnostics for random imputations: (a) displays of the completed data, intended to reveal unusual patterns that might suggest problems with the imputations, (b) comparisons of the distributions of observed and imputed data values, and (c) checks of the fit of observed data to the model used to create the imputations. We formulate these methods in terms of sequential regression multivariate imputation [Van Buuren and Oudshoom 2000, and Raghunathan, Van Hoewyk, and Solenberger 2001], an iterative procedure in which the missing values of each variable are randomly imputed conditional on all the other variables in the completed data matrix. We also consider a recalibration procedure for sequential regression imputations. We apply these methods to the 2002 Environmental Sustainability Index (ESI), a linear aggregation of 68 environmental variables on 142 countries, with 22% missing values.

1 Introduction

When considering models to impute missing data, the hypothesis of missing-at-random (MAR) can, inherently, never be tested from observed data.

However, any specific imputation model, whether MAR or not, will be fit to observed data, and that fit can be checked. In particular, we propose checking the fit of multivariate imputations by examining the model for each imputed variable given all the others. From a Gibbs sampling perspective, we are checking the fit to data of each full conditional distribution.

In addition, the completed data sets can be checked for plausibility. Formally, though, this is not a hypothesis test because the plausibility check inherently uses external information or speculation (e.g., that a particular variable should not have bimodal distribution, say, in the complete data).

*Revision for *Applied Statistics*. We thank John Carlin and Jennifer Hill for helpful comments and the National Science Foundation for partial support of this research. In addition, the 2002 Environmental Sustainability Index is the result of collaboration among the World Economic Forum's Global Leaders for Tomorrow Environment Task Force, the Yale Center for Environmental Law and Policy, and the Columbia University Center for International Earth Science Information Network.

†Department of Statistics, Columbia University. kobi.abayomi@columbia.edu

‡Departments of Statistics and Political Science, Columbia University. gelman@stat.columbia.edu

§Center for International Earth Science Information Network (CIESIN), Columbia University. mlevy@ciesin.org

1.1 Missingness

Multiple imputation (MI) has become popular in the twenty-five years since its formal introduction [Rubin 1978], and a variety of imputation methods and software are now available [e.g., Schafer 1997, Van Buuren and Oudshoom 2000, and Raghunathan, Van Hoewyk, and Solenberger 2001]. The development of diagnostic techniques for multiple imputation, though, has been retarded by the belief that the assumptions of the procedure are untestable from observed data.

The argument is that the quality of imputed data cannot be checked; imputed values are guesses of unobserved values, which are unknown. There are at least two responses to this argument:

1. Imputations can be checked using a standard of reasonability: the differences between observed and missing values, and the distribution of the completed data as a whole, can be checked to see if they make sense in the context of the problem being studied.
2. Imputations are typically generated using models (such as regressions or multivariate distributions) fit to observed data. The fit of these models can be checked.

Diagnostic techniques do exist: we can characterize them as *external*—comparisons to outside knowledge—or *internal*—specific to the observations and modeling. This article illustrates how a battery of techniques, of both types, can serve as a comprehensive method for assessing the goodness of imputed data.

We apply these diagnostics to a randomly selected completed dataset constructed using a multiple imputation procedure. The completed data is used to construct an index of environmental sustainability. We believe this approach is appropriate for the broader applied statistics community as well as environmental indexers. On the one hand we seek to introduce our method as a semi-automatic post-imputation procedure. On the other, we recognize that the particular findings are specific to environmental indexing. We hope that researchers in other applied fields will adapt these diagnostic ideas to the specific features of their problems.

1.2 The ESI...

The Environmental Sustainability Index (ESI) was created as a measure of overall progress towards environmental sustainability and designed to permit systematic and quantitative comparison between nations [World Economic Forum 2002]. The ESI is a scaled linear combination of 68 variables of environmental concern. Environmental measures (such as oxide emissions and concentration) are included along with political indicators relevant (such as civil liberty and level of corruption) that are relevant to environmental sustainability [World Economic Forum 2001, 2002].

The ESI, like other indices of environmental concern (such as the environmental wellbeing index (EWI), and the human development index (HDI)) tries to condense dissimilar social and physical metrics into cohesive summaries for national level comparisons [Prescott-Allen 2001, UNDP 2002].

Environmental Systems (13 variables) Measurements on the state of natural stocks such as air, soil, and water.

Environmental Stresses (15 variables) Measurements on the stress on ecosystems such as pollution and deforestation.

Vulnerability (5 variables) Measurements on basic needs such as health, nutrition, and mortality.

Capacity (18 variables) Measurements of social and economic variables such as corruption and liberty, energy consumption, and schooling rate.

Stewardship (13 variables) Measurements of global cooperation such as treaty participation and compliance.

Figure 1 Components of the 2002 Environmental Sustainability Index (ESI).

The ESI can be partially disaggregated across measurably similar groups of variables (components). See Figure 1.

1.3 . . . and missingness

As noted in the ESI [2002] report, "missing data are an endemic problem for anyone working with environmental indicators." Environmental data are often dissimilarly reported across regions or nations—rendering the data quality poor, missing, or so incomparable that variables need to be treated as missing. Index constructors tend to use simple missing-data methods such as casewise deletion and column averaging. For example, the 2001 ESI set missing values to the minimum of three univariate regressions. Broadly, index constructors are less concerned with the point estimate of a missing value and more with the final value of the index—a complete-data statistic. Within social science literature writ large, however, multiple imputation—the process of combining a set of missing value estimates—is becoming a popular tool [see Rubin 1996]. Multiple imputation allows inference on a complete-data statistic, by fitting a complete-data model to the observed data.

A variable is *missing completely at random* (MCAR) if the probability of missingness is the same for all units. Missingness is generally *not* completely at random, as can be seen from the data themselves. For example, in the ESI, some countries are much more likely than others to have missing observations. A weaker condition is *missing at random* (MAR): that the probability a variable is missing depends only on available information. For example, if a variable is more likely to be missing for countries with low values of per capita GDP, and this GDP predictor is available for all countries, then this pattern could be missing at random but not missing completely at random. Lastly, both assumptions are violated if the probability of missingness varies and cannot be characterized by an available predictor: this condition is called *not missing at random* (NMAR) [Rubin 1976, Little and Rubin 2002].

There are imputation procedures that do not require the MAR assumption, such as selection or pattern-mixture models (see Heckman [1976] and Little and Rubin [2002]). It is common in practice, however, to impute regression-type models fitted to the available data under the missingness at random assumption, with the understanding that these imputations, while imperfect, may be useful, especially if the fraction of missingness in the dataset is small.

In principle, it is impossible to test the assumption of missingness at random without additional data collection, since the information that would be used to make such a test is, by definition, unavailable. We suspect that this theoretical difficulty has discouraged researchers and practitioners from developing diagnostics for imputations.

However, there can be indirect evidence of problems relating to the missingness assumptions, and thusly the imputation model. For an example, consider the observed and imputed data for the BODWAT variable—a measure of the industrial and organic pollutants per available freshwater (Metric Tons of BOD Emissions per Cubic Km of Water) [ESI 2002]. Most of the observed data is on the order of 10^{-1} to 10^{1}. The exception is Kuwait, which, as a net importer of freshwater, is at 10^{9}. Under the general MAR assumption, one imputation draw is at the right tail of the observed distribution. The imputation model is sensitive to this outlier even after the inclusion of an indicator as a covariate; the completed data distribution is bimodal. In the absence of extra information (e.g. knowledge of water policy in Kuwait) it would be natural to suspect the model underlying the imputations, and it would be appropriate to examine the observed data more closely.

We illustrate in Figure 2. In this example, the assumed normal distribution for the complete-data distribution of BODWAT is clearly wrong. This is our point: one might naively think that missing-data models are inherently uncheckable, but here we can see that the normal model, if valid, would lead to implausible conclusions about the observed and missing values of this variable.

In general, evidence of departure from the missingness assumptions are not necessarily apparent as problems in the residual distribution of the imputation model. Even if the residuals appear correct, the completed data may look implausible. The model itself may not fit; the model may fit and a bimodal distribution (like that in Figure 2) is correct; the model may fit the observed data but not the missing. In these cases, the diagnostic in Figure 2 will flag variables which deserve further inspection.

For another context in which missing-data models can be checked, consider selection models, which are sometimes used for sensitivity analysis of imputation procedures. For any example, the constructed completed dataset given any selection model can be examined. If, for example, it looks bimodal, with observed data in one mode and missing data in the other mode, this may go beyond believability—thus suggesting limits to the range that sensitivity must be tested. This is related to the index of sensitivity to non-ignorability [Troxel, Ma, and Heitjan 2004].

The graphical displays just described are *external* (in the sense of the observed dataset) diagnostics of an imputation procedure. There is no *internal* test of missingness at random (or, for that matter, of whatever non-missing-at-random model might be used). However, internal tests can be performed of the imputation model itself, in the context of the observed data used to fit the model. We shall focus on sequential regression imputation models, so that standard regression diagnostics

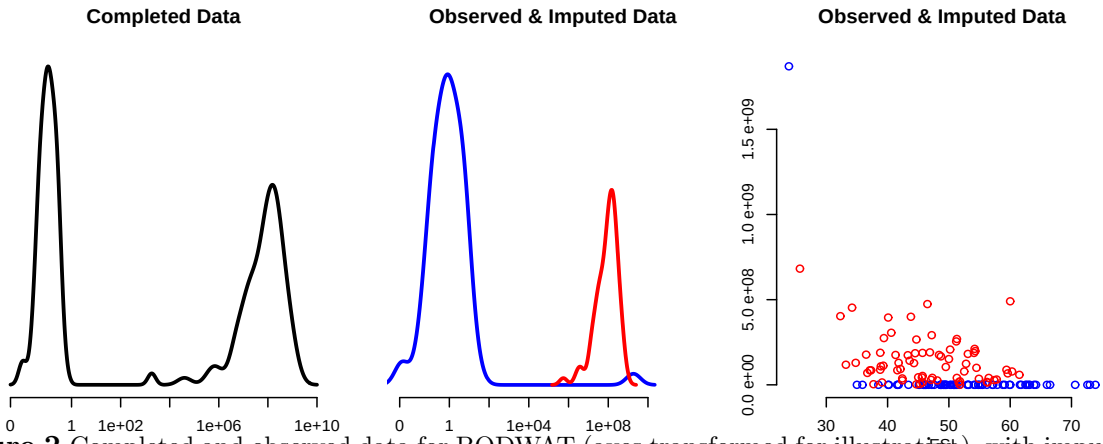

Figure 2 Completed and observed data for BODWAT (axes transformed for illustration), with imputations based on a fitted normal distribution. The completed data in the leftmost graph are bimodal. Observed data are shown in blue, imputations in red, completed data in black. The histogram (center) has the imputed data, from one draw, at the right tail of the distribution. The observed outlier is rightmost and blue. Imputations generated under this model are incorrect. The model would be flagged because the imputed data markedly differ from the observed. A post hoc plot of the observed data illustrates the problem: the influential outlier in the imputation model (large blue dot) is Kuwait. Available observed data for cases where BODWAT is imputed may be similar to the Kuwait elsewhere; the imputation model at this variable has, incorrectly, low precision.

can be used to check model fit and recalibrate if residuals do not have mean value zero conditional on available predictors. Our general procedure is to use external tests to flag possible problems which then must be checked using subject-matter knowledge. Internal tests can be performed more automatically, by analogy to regression diagnostics.

These examples illustrate where and how external tests motivate inspection of the multivariate model used to generate the imputed data. Remember that the goal is not data modeling, but generation of a (completed) data statistic. In both of the illustrated cases, a poor imputation procedure could easily be obscured by the completed data. As well, violations of the random missingness assumptions could be hidden behind a completed data statistic. In MI, the multivariate model, even when implicit, can and should be checked using comparisons of observed and imputed distributions, Under a default assumption modeling idiosyncracies are distinguishable. Indeed, *a fortiori*, using the completed dataset to check the MI model should flag, at least, where the modeling may be inappropriate—if not explicitly where the missingness assumptions are not met.

1.4 . . . and missingness in the ESI

As is shown in Figure 3, the countries with low environmental sustainability indexes and low incomes tend, unsurprisingly,[1] to have more missing items in the ESI. (ESI and per-capita GDP are positively correlated, but this correlation is only 0.4.) Figure 4 displays the overall pattern of

[1]Data collection is usually an expensive task. In the context of non-random missingness, poorer countries may have less ability, as well as lesser motivation, to collect and report environmental data broadly.

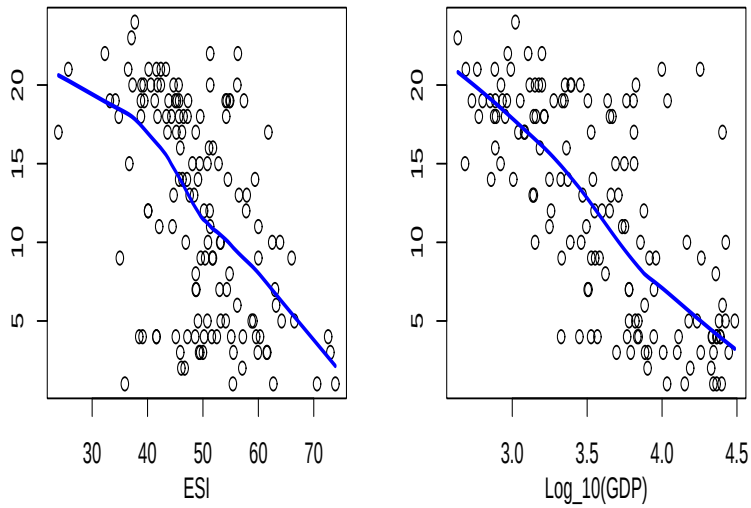

Figure 3 For each country, percent of variables missing, plotted vs. ESI and GDP, with fitted lowess lines [Cleveland 1979]. Countries with higher environmental sustainability indexes and higher incomes tend to have fewer missing items. The graphs clearly demonstrate that the variables are not missing completely at random.

missing data: every country is missing some data, and a total of 22% of all the potential data are missing. Constructing the ESI using available cases would severely restrict its scope.

To generate an index with a useful coverage, the missing values were multiply imputed. We wanted the ESI, and thus its component parts, to be defensible, and thus it was important to check the imputations to see if they were reasonable. With 68 variables in 142 countries, a somewhat automatic method was necessary to screen the imputations and identify potential problems. This motivated the suite of tools developed in this paper.

2 Methods

2.1 Multiple imputation using sequential regressions

We begin with a dataset—a data matrix with missing values—and suppose that the user has already decided on a multiple imputation procedure, fit it to the data, and constructed a set of imputations. We then have several imputed *completed datasets*. Our diagnostics can be applied independently to each completed dataset. These methods are intended for multiple imputation procedures where the imputations are draws from a predictive distribution. For simplicity we shall work with just a single randomly-chosen imputation in our example.

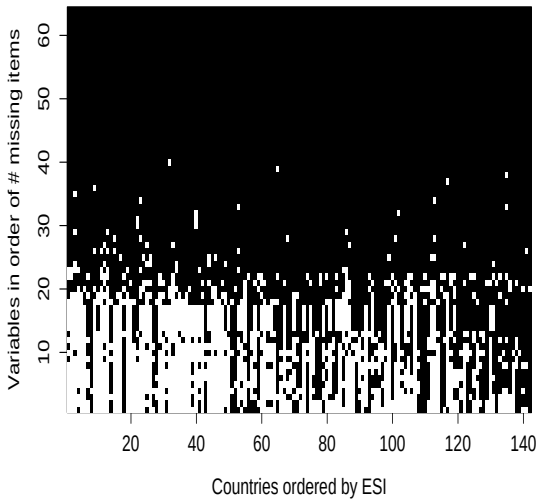

Figure 4 The pattern of missingness—missing values in white. Countries are listed in rank order of ESI. (Kuwait is the first country on the abscissa, Finland is the last.) Variables are listed in order of number of missing items in the ESI. (WATCAP (water capacity) is bottom on the ordinate, GMS.SS (suspended solids) is top.)

We shall assume the imputations have been constructed from a model of the data. Multivariate models that have been used include the normal, t, and general location families [Liu 1995, Schafer 1997]. More generally, Van Buuren, Boshuizen, and Knooket [2000], Van Buuren and Oudshoom [2000], and Raghunathan et al. [2001] define imputations using a set of marginal conditional distributions, a more general—though potentially inconsistent—specification that allows imputation singly at each variable conditional on all the others in the dataset (see Gelman and Raghunathan [2001]). Sequential regression multiple imputation (SRMI) proceeds by partitioning and ordering the dataset by number of missing items, then imputes the least missing variables before the most missing at each round of the procedure. The key idea is to see multivariate imputation as a linked set of regression models, or analogously chained equations, and proceed iteratively until convergence in model parameters is achieved.

We used the Raghunathan et al. [2001] software, in the end imputing approximately 19% of the data for the ESI. Several variables were considered unfit for imputation before the imputation procedure, some after, and these were removed from the definition of the ESI.[2] We imputed a total of 10 complete datasets and constructed an estimated ESI on the average of those 10.

2.2 Flagging: tests of difference between the observed and imputed data

The task here is to identify where imputations markedly differ from observed values. Differences can originate from the model used to generate the imputations or can indicate a more serious

[2]Prior conditions for exclusion from imputation varied, including dearth of available cases. *Post hoc*, highly variable distributions were removed.

violation of the missingness assumptions. In both cases the flagging compares the imputed values to the observed. In the sense that the completed dataset is model generated, these are tests of the imputation mechanism. A raised flag indicates a potential problem with the imputation mechanism which could be specific to the generation model, or, more broadly, an inability of the model to capture violations of the missingness assumptions.

There are no foolproof tests of the assumptions of the imputation procedure. We will judge the propriety of the imputed values by comparison with observed. Again, we cannot actually test unobserved values for agreement with an unknown true distribution. We claim that the fit of the multivariate model, in this case an imputation model, must always be checked: it is natural to check the model against the observed data. Chained equation approaches such as SRMI are particularly amenable to multivariate model checking. It is a misconception that the possibilities of non-ignorable missingness implies that imputations are uncheckable. Every model, in general, has untestable aspects—imputation modeling is not uniquely characterized by untestability. For imputations the end result is the complete dataset, which suggests the existence of hypotheses about characterizations of a complete dataset. The point is that imputation modelers usually have a notion about what this complete dataset looks like, and can use these notions to frame their flagging procedures. Restated: We, or any imputation modeler, can do better than guessing about guesses, by using the observed to flag possible problem imputations.

We can discard the imputed values in cases where they pathologically differ from expectation—in a few cases, we did just that. In many others, however, our expectations remained uninformed and pathology in the imputations was ill-defined. Our goal was, again, to test the propriety of the imputations, flag potential problems, and fix or refine our imputation model.

We emphasize that differences in distribution between the imputed and the observed *do not necessarily* indicate violations of the missingness assumptions or problems with the imputation model. Some deviations between observed and missing values can be expected under MAR, but extreme departures require assessment for plausibility. In the absence of true tests, though, we can—and must—exploit the dependence between the completed dataset and the missingness: the observed values provide a basis.

2.2.1 Density comparisons

We can numerically compare the empirical distributions of the observed and the imputed using the Kolmogorov-Smirnov test for each variable, raising the flag when we find statistically significant differences.[3] We also examine empirical densities visually.

Differences in distribution do not necessarily signal a problem with the imputations: the distributions of missing data can differ from the distributions of the observed data while still being missing at random. In fact, if the data have been imputed using this assumption, then any differences in distributions are necessarily explainable by other variables in the dataset. Nonetheless, as discussed in the hypothetical examples of the appendix, dramatic differences between the imputed and ob-

[3]The p-values for these tests are approximate; the imputations are generated from the observed data, thus the empirical distributions of the imputations are not independent of the observed.

served data can suggest a *potential* problem, and in a context with many imputed variables, it is helpful to have some screening devices to identify these potential problems.

We treated the empirical density plots as flags for potential problems with the imputed estimates—in a sense the empirical density plots are visual representations of the KS tests.

Classical statistical significance provides a convenient cutoff rule that seems to work well in our example. More generally, a procedure for deciding which discrepancies to further examine should reflect the cost of performing the further examination along with the potential costs of skipping over a variable. In general there is no reason to suppose that setting a 5% significance level will be appropriate, but we present this rule here as a starting point, worthy of further examination. To the extent that we can examine the distributions visually, this is not necessarily a crucial issue in practice. But in a general implementation we would at the very least allow other thresholds to be considered and perhaps have alternative rules such as selecting for initial examination the 10% of variables whose KS statistics are the most extreme.

2.2.2 Bivariate scatterplots

Bivariate scatterplots allow us to compare the internal consistency of the missing and observed observations with respect to a continuous predictor. In this diagnostic we look for obvious differences between the distributions of the variable as it relates to the predictor. Coupling these plots with the empirical densities allow us to flag differences in distribution as problematic—we look for unusual patterns in the internal data (observed and imputed) with respect to our external knowledge. Both are important: the *external* knowledge at each variable and the *internal* (KS test type) difference.

In this paper we use a maximum of five comparisons, data which represent *external* knowledge, for each variable in the ESI dataset. We use two closely related combinations of available internal data, one unrelated variable included in the internal dataset, and two closely related variables not included in the internal dataset.

2.3 Fitting: tests of the fit of the imputation model to the observed data

2.3.1 Residual plots

The SRMI software of Raghunahan et al. [2002] does not allow inspection of the imputation model—this is a disadvantage with respect to checking the validity of the second MI assumption. We constructed a proxy for the iterated SRMI models, however, by selecting the best stepwise model at each variable in the completed data (Y_j) regressed on all others ($\mathbf{Y}_{-j}$). We generated predicted values ($\hat{y}_{ij}$) for the 64 variables in the dataset, and we consider these analogs for the unavailable predicted values from the SRMI complete-data models. Each residual r_{ij} is the difference between the observed value in the completed data anbd the prediction of the best stepwise regression. For the imputed data this is the difference between the predicted value of the SRMI model and the best stepwise model. For the observed data this is the traditional residual.

Under the model, the pattern of residuals versus expected values should be random: we generate the imputations from a series of linked, linear regressions.

2.3.2 Fixing the imputations

The aim here is to refine the complete-data model: we believe that we can improve the imputed values by capturing the non-random patterns in the observed and then updating our guess for each imputation.

We fit a lowess curve [Cleveland 1979] to each of the scatterplots of residual differences between an available stepwise model (one we obtain by regressing each variable on all other variables in the completed data) and the SRMI output. In general, where the imputation model is available, we would fit a curve to the observed values vs. the residual differences between the observed and the predicted. We would then update the imputations only, using the curve as the proper residual function. In this paper, we use the lowess curve - in general other functions are possible. Please see the Appendix, section 4.

We applied our method of residual refinement to a sample environmental dataset [Johnson and Wichern 1998] under complete (MCAR), random (MAR) and nonrandom (NMAR) missingness mechanisms. See the Appendix.

When the assumption of random missingness is true, differences in the pattern of residuals indicate a deficiency in the imputation model which the residual calibration corrects. However, when the assumption is false, differences between observed and imputed are not correctable by the residual calibration.

It can be difficult to fix the imputation with the proposed method because fixing is done based on marginal distributions. Marginal adjustments to the imputations in the presence of an incorrect imputation model may introduce incoherence. When problems are found, the imputer should refine the imputation model to create improved imputations that are consistent. With data analysis in general, careful model building is critical when the fractional missing data information is large for subsequent complete-data analysis.

3 Application

We illustrate the proposed methods with the data and imputations for the Environmental Sustainability Index. We look at all variables, first, and then each subset more systematically—tailored to this application. A first step is to look at density plots of variables which are flagged via KS type tests, see Figure 6. A second step is to display the observed and imputed data for all imputed variables, versus the overall index, as shown in Figure 7. We discuss these plots in particular for a group of variables (a 'component') in the ESI. As practitioners, we would investigate all of the data similarly.

Environmental Systems NO2(y)—urban NO2 concentration; SO2—urban SO2 concentration.

Environmental Stresses NUKE(y)—Radioactive waste; WATSTR—Percentage of the country's territory under severe water stress.

Vulnerability DISRES(y)—Child death rate from respiratory diseases; WATSUP—Percentage of population with access to improved drinking water supply.

Capacity SCHOOL(y)—mean years of schooling (age 15 and above); GASPR—Ratio of gasoline price to international average.

Stewardship FSHCAT(y)—total marine fish catch; FSHCON—seafood consumption per capita.

Figure 5 ESI component groupings and variables used to illustrate flagged, and unflagged, differences. The significantly different variable—flagged variable—is indicated with (y).

3.1 A quick look at all the variables

There are plausible explanations for the differences in scatterplot patterns that we see when plotting the data from each variable versus the composite ESI score in Figure 7. Taking the environmental systems group as an example: we may expect that some countries with lower values, in GDP for instance, will have higher emissions—a finding that does not contradict environmental theory.[4]

This sort of information is easy to illustrate but, perhaps equally as easily, can be hidden if the user focuses on the complete-data summaries without checking the imputations.

We demonstrate in this subsection and the next, via (what we believe could be) semi-automatic processes, that methods of exploratory analysis designed for imputation procedures can specifically highlight, address and yield "better" complete-data statistics.

We begin by quickly identifying the variables in which imputed values different greatly from observed data. In all, about half of the imputed variables have KS tests indicating a statistically significant difference between observed and imputed values. The KS tests flag five variables as extremely problematic (approximate $p < .001$): NO2 concentration (NO2), radioactive waste (NUKE), child death rate from respiratory diseases (DISRES), mean years of schooling (SCHOOL), and total marine fish catch (FSHCAT).

For a brief illustration we select a 'flagged' variable within each ESI component grouping, see Figure 5, as well Figure 6. For comparison, in each group we chose one variable that did not significantly differ.

Figure 7 provides a snapshot of the differences between the observed and imputed values for the entire data — in some cases the differences are striking. We might suppose that the differences in distributions are functions of differences in the predictors. In absence of the knowledge of the missingness mechanism that supposition may be indefensible. First, we may expect that some

[4] An example is the BODWAT variable; see Figure 2.

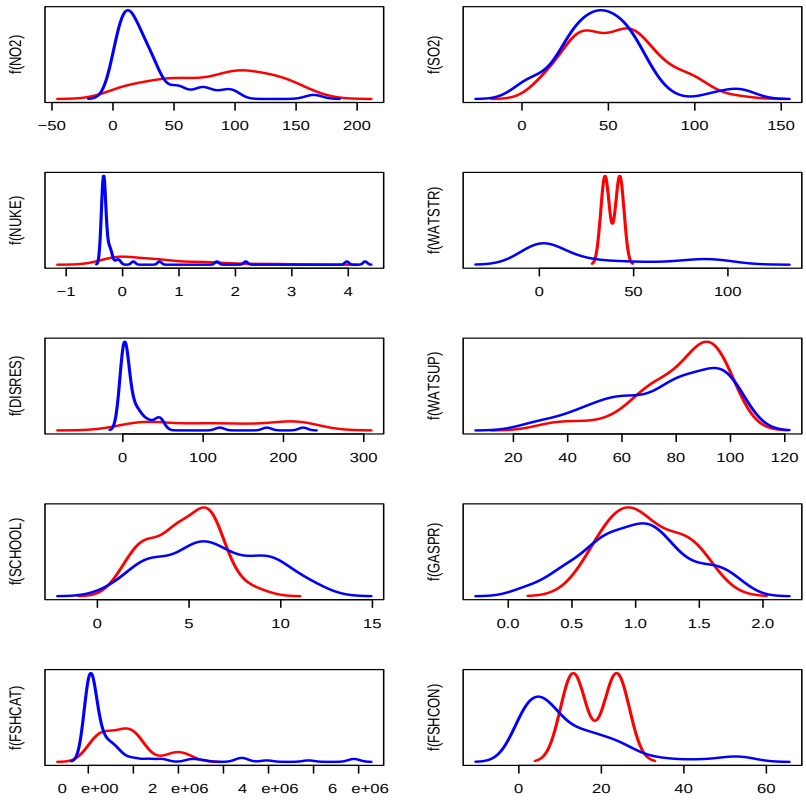

Figure 6 Left side: for each component of the ESI, a variable whose imputed values (red) differ significantly from observed values (blue). Right side: for comparison, a variable from each component which we did not flag. Possible flaws in imputations may appear in the graphs even when not indicated by the KS tests. As well, apparent differences in density plots may not be 'flagged' by the KS test - in particular where there are few imputes: $n = 2$ for WATSTR and FSHCON. Diagnostic by visual inspection is necessary.

countries misreport or restrict—intentionally or not—data (for example air and water particulate concentrations). Second, we may believe that anomalies in distribution, in a few cases, are caused by just a few influential observations. For example, extreme outliers in the distributions of WATCAP and WATINC (internal water capacity and per capita inflow) are idiosyncratic. As discussed earlier, Kuwait imports most of its water. The conclusive statement is that the completed ESI data demands a thorough diagnostic review.

3.2 A closer look—Environmental systems

As an illustration we look closely at the data in one component group of the ESI. As practitioners, we should repeat this exercise for all the data groupings. Figure 8 is an example of the sort of requisite post-imputation diagnostic plots we produce.

The environmental systems variables in this component are national level measures of the stock, or present state, of environmental quality. The data for environmental systems should be generally

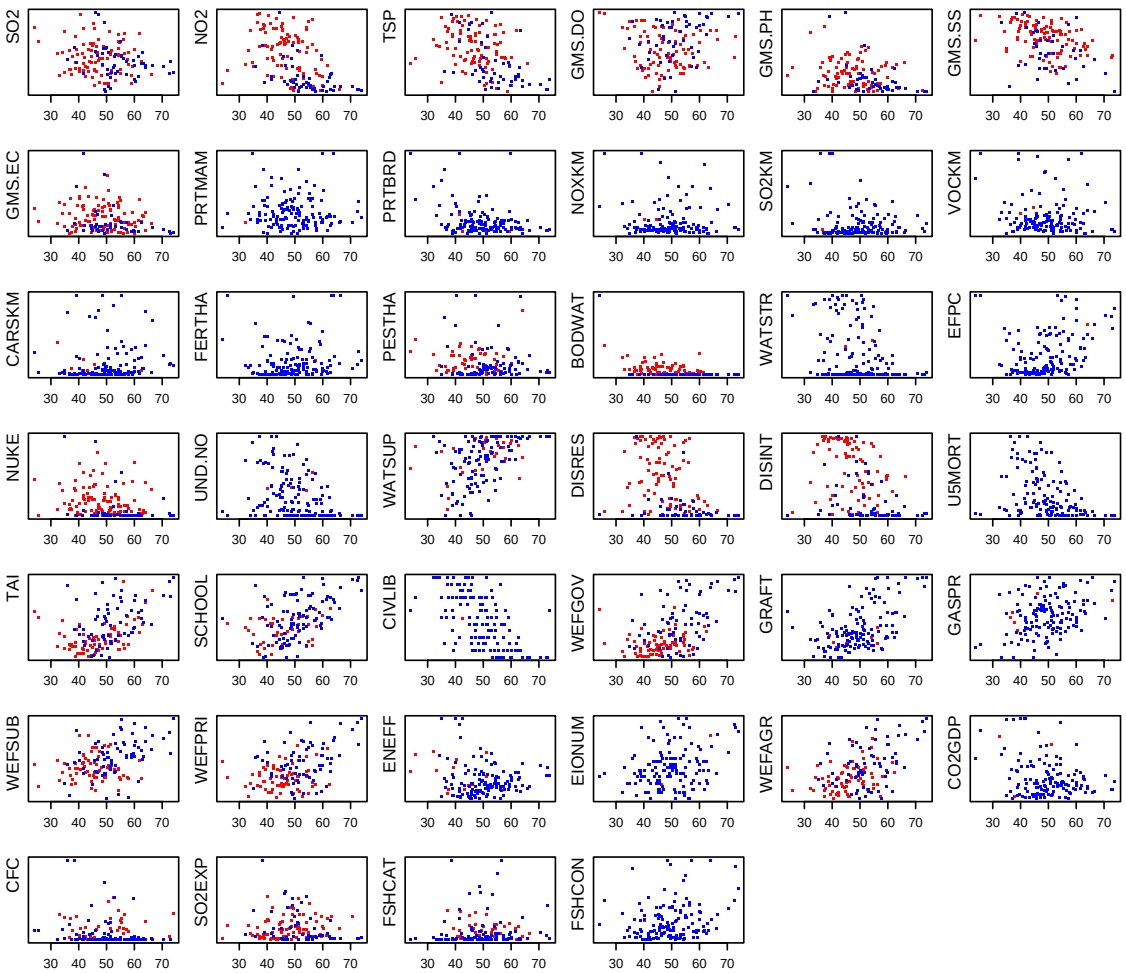

Figure 7 For each variable, its observed and imputed values for the 142 countries, plotted vs. the Environmental Sustainability Index. Imputed values, everywhere, are in red. Observed values are blue. At a glance there is evidence for nonrandom patterns of missingness in many variables, as discussed in detail in the text.

comparable across nations in the sense that the true values are easily observable and calculable. However, this component had the highest rate (36%) of missingness.

The KS test flagged the imputation of NO2 as significantly different, but not that of SO2. Excluding NO2 is not possible—we need both concentrations to return a full measure of air quality. We treat the KS test as an indicator, but not a determinant, of a potential problem. The difference in the distributions between observed and imputed values of NO2 appears to be driven by overprediction at moderate to moderately high levels. Again, this may or may not be problematic—it is possible that higher polluters have not reported appropriately and that we are imputing them correctly. At a glance, the imputed values of NO2 look more different from the observed values—with respect to SO2. One or two cases appear to drive the upward trend in NO2 imputations (Iran). Our supposition may be correct: the residual values for the imputations of NO2 have a greater magnitude

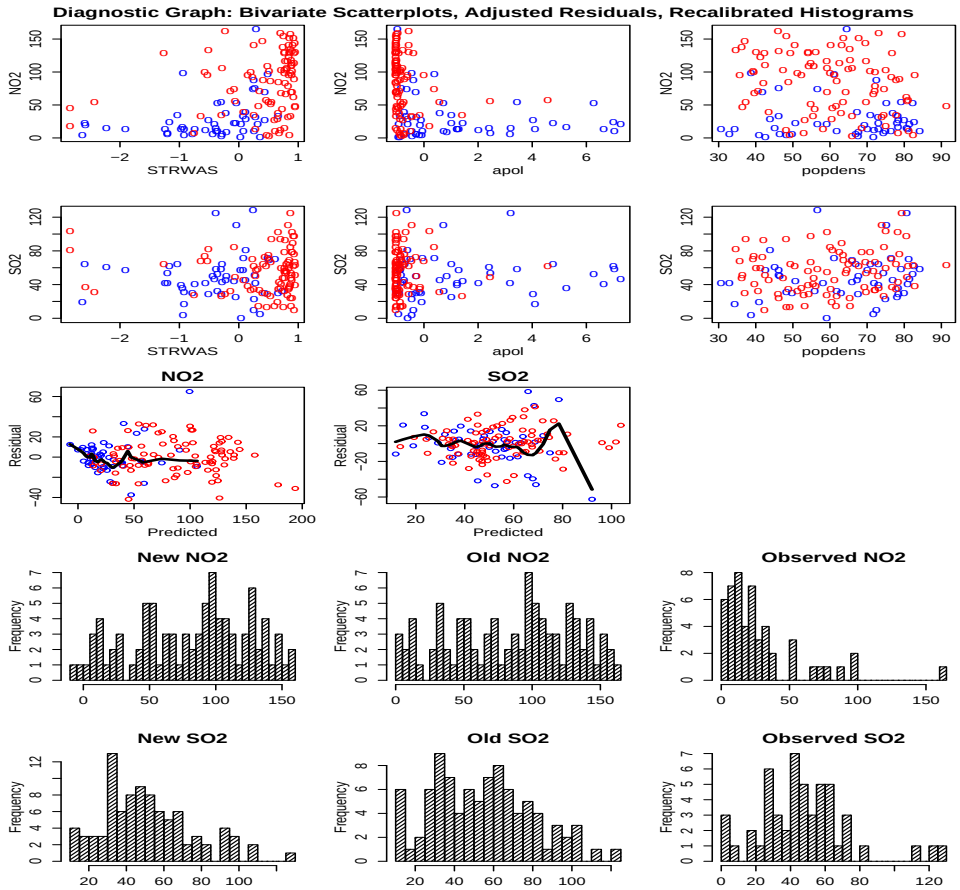

Figure 8 Environmental systems. NO2 is flagged as significantly different by KS test. Bivariate scatterplots highlight distributional differences. SYSAIR is a composite of air quality measurements used in the ESI. APOL is a composite of air quality measurements not used in the ESI. POPDENS is a measure of population density. The residual plots plot the predicted values from the best stepwise regressions against the difference between the (randomly selected) imputation and this predicted value. Histograms of the updated imputations are on the final rows.

and predicted range than the observed values. The values for SO2, in contrast, are more similar.

We adjusted the imputations for both variables by fixing the residuals of the imputations to match the lowess curve through the residuals of the observed. The adjustment affects the univariate histogram of SO2 more dramatically than NO2: the distribution of the imputed values matches the observed more closely. SO2 was not flagged as significantly different-the recalibration may not be appropriate.

As noted at the beginning of Section 3, we apply similar checking procedures on the remainder of the ESI data.

4 Discussion

Missingness in the ESI arises from the dearth of environmental metrics and is attenuated by the breadth of the ESI's coverage. The ESI has a high number of missing items because it broadly defined.

We already know that countries with more missingness have performed worse on the observable measurements; we don't know if the level of performance on unobserved measurement is dictating the missingness—several of the tests are suggestively affirmative. We can at least state that the distributions of the imputed and observed values differ, and we should state the there is evidence that the differences are attributable to the level of measurement—in violation of the least restrictive of the missingness assumptions. It is possible that many of the data are not missing at random.

The model used here for the imputations is far from perfect. In fact, *the point of this paper is to develop semi-automatic diagnostics in recognition of the fact that missing values are typically imputed using semi-automatic procedures.*

In our examples, we flagged some problems and then reviewed the imputations that highlighted obvious potential flaws. We began with numerical diagnostics—the Kolmogorov Smirnov tests—to flag problems, and we attended to the flags by using semi-automatic graphical techniques.

We recommend that these methods be applied *en suite*, perhaps as an included suffix to a standard MI package such as MICE [Van Buuren and Oudshoom 2001].[5] With a specified, available, imputation model, we would expect the refinement procedure to perform "better"—in the sense that discord between the imputed and observed observations will be more clearly characterized.

We have used post hoc methods to compare and adjust imputation models, in a sense investigating meta-parameterizations of missingness mechanisms. By flagging sets of imputations that look particularly troublesome, using observed values and related external values, we have shown—at least—where we should lower our confidence in our imputed values. Further, we have investigated where we can improve upon our imputation model by revisiting the observed and exploiting the difference in patterns of the observed and missing data with respect to the imputation model.

Finally, the ESI is an attractive case for the development of MI diagnostics. Environmentalism in general, and sustainability in particular, have much to do with what is unknown about the systematization of individually well-understood concepts. The ESI is a case where we can intelligently diagnose and correct problematic imputed values: we have at our disposal rich internal, as well as external, information and require only a framework from which to procedurally investigate and correct our modeling.

[5] Think of a graph array—Figure 8—for each of the components, as a complementary, necessary diagnostic output to a completed dataset for any imputation software.

A Appendix

A.1 Computation of the ESI

The environmental sustainability index [World Economic Form 2002] is defined as ESI $= 100 * \Phi\left(\frac{1}{|k|}\sum_k \frac{1}{|J_k|}\sum_{j\in J_k}\left(\frac{Y_j-\overline{Y}_j}{\mathrm{var}(y_j)^{1/2}}\right)\right)$. Here $\mathbf{Y} = (\mathbf{Y}_{J1},\ldots,\mathbf{Y}_{Jk}) = (Y_1,\ldots,Y_{68})$ where the J's are groups of similar information, and Φ is the inverse standard normal distribution function.

A.2 Missingness assumptions

Extending the above:$\mathbf{Y} = (\mathbf{Y}_{J1},\ldots,\mathbf{Y}_{Jk}) = (Y_1,\ldots,Y_{68}) = (\mathbf{Y}_m,\mathbf{Y}_o)$ we can say that the pattern of missingness is completely random (MCAR) if it is distributed independently of the dataset, or $f(M|\mathbf{Y},\phi) = f(M|\phi)\forall\mathbf{Y},\phi$, where M is an indicator matrix of the same dimension as $\mathbf{Y}$

A weaker condition, missing at random (MAR), exists if the pattern of missingness is dependent only upon the observed values, i.e.: $f(M|\mathbf{Y},\phi) = f(M|\mathbf{Y}_o,\phi)\forall\mathbf{Y}_m,\phi$. Here, M is a random variable characterizing the missingness process, usually $\mathbf{M}$, and ϕ are possible unknown parameters.

We say that the pattern of missingness is not at random (NMAR) if both conditions are unmet, that is, $\exists\mathbf{Y},\phi,s.t.,f(M|\mathbf{Y},\phi) = f(M|\mathbf{Y}_m,\phi)$.

A.3 SRMI procedure

Commonly, $G(\mathbf{Y},\theta)$ is supposed $|k|$-variate joint normal, and the missing data are imputed as draws from the joint posterior (as in MCMC imputation). Van Buuren [2001] and Raghunathan [2001] investigated that a G can be replaced with a set of conditional distributions $G = \prod_{J_k\in K} G_{J_k}$, in many cases. Sequential Regression Multiple Imputation (SRMI) proceeds by partitioning the dataset: $\mathbf{Y} = (\mathbf{Y}_{J1},\ldots,\mathbf{Y}_{Jk}) = (Y_1,\ldots,Y_{68}) = (\mathbf{Y}_m,\mathbf{Y}_o) = (Y_1,\ldots,Y_{|k|-r},Y_{|k|-r+1},\ldots,Y_{|k|})$. in order of missingness, where r is the number of variables with missing values. Then $\mathbf{X} = (Y_1,\ldots,Y_{|k|-r})$; and $\mathbf{Y}^* = (Y_{|k|-r+1},\ldots,Y_{|k|})$. $\mathbf{Y}^*$ is regressed, iteratively, on $\mathbf{X}$. The steps, in this application, are

1. The first round of the SRMI algorithm begins by regressing Y_1, the variable with the least missing items, on $\mathbf{X}$.
2. Now Y_1 is entered into $\mathbf{X}$ and the algorithm regresses Y_2 on $(\mathbf{X},Y_1)$. The algorithm continues until $Y_{|k|}$ is completed by regressing it on $(\mathbf{X},Y_{|k|-1})$.
3. The next round continues in the same manner, with $(\mathbf{X},Y_1,\ldots,Y_{|k|})$ the new predictor set.
4. The algorithm cycled through the above steps until the imputed values converged.

We repeated the algorithm $m = 10$ times, averaged the imputed data sets, and calculated the ESI on the final averaged imputed data set.

Gelman and Raghunathan [2001] discuss why SRMI imputations might be useful, depsite that in general they do not correspond to a specific joint model.

The SAS implementation of the SRMI procedure allows bounds to be set for each variable—we set the allowable extrema by the observed distribution. We noticed that unconstrained imputed values tended to ranges far wider than the observed distributions. At each variable, this may or may not be a problem: if the missingness mechanism is, perhaps, not completely at random, difference in the imputed values may be a function of the observed values and possibly appropriate. We cannot say which mechanism is present and allowed for the truncation of extreme imputations.

A.4 Fixing imputations—refinement procedure

Let $\hat{G}$ be an estimate of G; $\hat{G}$ is the imputation model used to generate a complete dataset.[6]. Let $\hat{y}_j = \hat{G}(\mathbf{Y}_{-j})$ be the predicted values from the imputation model for each variable j. Take $H(\hat{y}_j, y_j) = \hat{G}(\mathbf{Y}_{-j}) - y_j = \tilde{y}_j$. H may, typically, be a non-parametric function: an estimate of the differences between predicted values of the observed and the actual observed. Then $\tilde{y}_j$ are the refined imputations when the arguments to the above are the imputed values. In this paper H is a lowess curve.

We correct (or calibrate) the imputed values by supposing a function (H) from the predicted (of the observed) to the residuals (of the observed) and forcing the residuals of the imputed to match that function (by subtracting or adding the difference at each residual).

A.5 Simulation study

Beginning with an example set of air quality data [Johnson and Wichern 1999] we investigated the behavior of our imputation refinement procedure under three simulated missingness mechanisms: MCAR, MAR, NMAR. Let z_{ij} be 1 indicating that observation $y_{i,j}$ is missing. distributed as following under each assumption: **MCAR**—$z_{i,j} \sim Bernoulli(p_j)$; **MAR**—$z_{i,j} \sim logit^{-1}(a1_j + (\hat{y}_{i,j} - b1)/c1)$; **NMAR**—$z_{i,j} \sim logit^{-1}(a2_j + (y_{i,j} - b2)/c2)$.

We set the p_j, $a1$ and $a2$ to decrease with j to generate a pattern of monotone missingness under each of the assumptions. Constants $b1$,$b2$,$c1$,$c2$ exist so that the number of missing items is relatively equivalent for each of the missingness mechanisms.

We found, in general, that the refined imputations replicated the shape and range of the observed distributions more closely for all missingness mechanisms. The improvement in similarity was less pronounced, though, for imputations under the NMAR assumption — and more so for the imputations on the MCAR assumptions.

[6] In this application $\hat{G}$ is set as the best stepwise regression of Y_j on $\mathbf{Y}_{-j}^{(k)}$

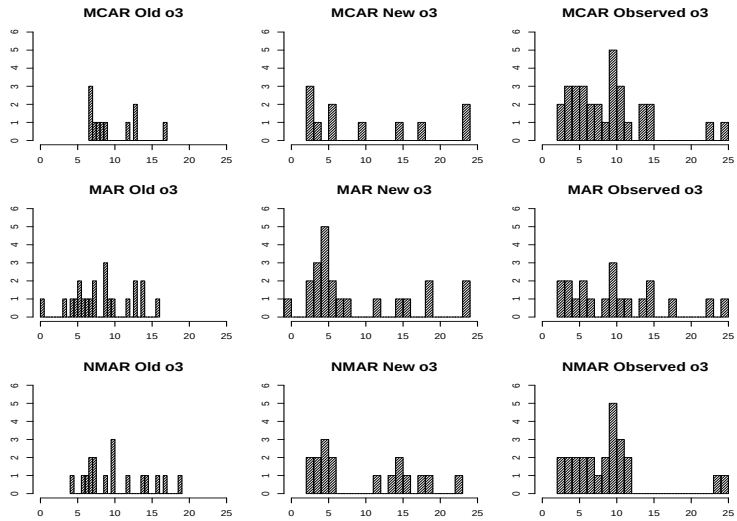

Figure 9 Simulated imputation refinement on air quality data. The first two graphs in each row are the distribution of O3 before and after recalibration. The last graph in each row are the observed data. The first two rows are imputations and recalibrations under MCAR and MAR models. The refinements more closely mimic the distribution of the observed under MCAR and MAR missingness mechanisms. Under NMAR the refinements perform less well - the imputed distribution has a wider range than the observed.

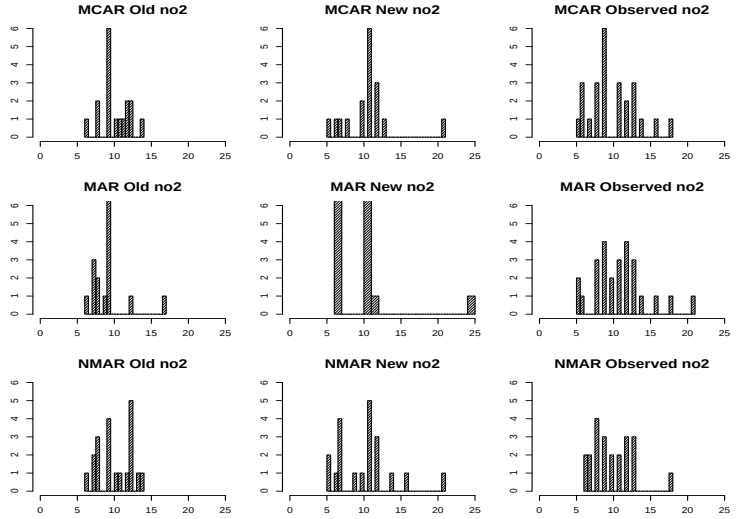

Figure 10 Simulated imputation refinement on air quality data. The first two graphs in each row are the distribution of NO2 before and after recalibration. The last graph in each row are the observed data. The refinements match the distribution of the observed better than the original imputations under MCAR missingness. The range of the refinements is greater than the observed under MAR; under NMAR the original imputations more closely match the observed data

B References

Cleveland, W. S. (1979). Locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829-836.

Diggle, P. J., and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics* **43**, 49-93.

Gelman, A., and Raghunathan, T. E. (2001). Using conditional distributions for missing-data imputation. Discussion of "Conditionally specified distributions" by Arnold et al. *Statistical Science*.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* **5**, 475-492.

Johnson, R. A., and Wichern, D. W. (1998). *Applied Multivariate Data Analysis*. Upper Saddle River, N.J.: Prentice Hall.

Little, R. J. A., and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, second edition. New York: Wiley.

Liu, C. (1995). Missing data imputation using the multivariate t distribution. *Journal of Multivariate Analysis* **48**, 198–206.

Raghunathan, T. E., Solenberger, P. W., and Van Hoewyk, J. (2002). IVEware. `http://www.isr.umich.edu/src/smp/ive/`

Raghunathan, T. E., Van Hoewyk, J., and Solenberger, P. W. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* **91**, 473–520.

Rubin, D. B. (1978). Multiple imputation in sample surveys—a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20–37.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. New York: Chapman and Hall.

Troxel, A., Ma, G., and Heitjan, D. F. (2004). An index of local sensitivity to non-ignorability. *Statistica Sinica*.

Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* **18**, 681–694.

Van Buuren, S., and Oudshoom, C. G. M. (2000). MICE: Multivariate imputation by chained equations. `web.inter.nl.net/users/S.van.Buuren/mi/`

World Economic Forum (2001, 2002). Environmental Sustainability Index. Global Leaders for Tomorrow Environment Task Force, World Economic Forum and Yale Center for Environmental Law and Policy and Yale Center for Environmental Law and Policy and Center for International Earth Science Information Network. Davos, Switzerland and New York.