Diagnostics for Multivariate Imputations*

Kobi Abayomi[†], Andrew Gelman[‡], Marc Levy[§]

June 29, 2004

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, 2000), 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% missingness in its components.

1 Introduction

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 2000, and Raghunathan, 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:

*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 Environmental Engineering, Columbia University. kobi.abayomi@columbia.edu

[‡]Department of Statistics, Columbia University. gelman@stat.columbia.edu

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

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 an imputation performed for a multivariate dataset that is used in constructing 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 imputation diagostic 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 Form 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 is conceptually similar to the growing cohort of environmentally concerned indexes (such as the environmental wellbeing index (EWI), and the human development index (HDI)) in its endeavor to condense dissimilar social and physical metrics into broadly defined yet conceptually cohesive numbers for national level comparisons [Prescott-Allen 2001, UNDP 2002].

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." The ESI variables in particular, and environmental data in general, are set with a high number of missing items across any reasonable coverage. Environmental data are often dissimilarly reported across regions or nations—rendering the data quality very poor, missing,

or so incomparable that it must be set to missing. The remediation of missingness, especially within the indexing community, tends to be unsophisticated: case-wise deletion and column averaging are commonly used methods. 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.

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)

A variable is *missing completely at random* (MCAR) if the probability of missingness is the same for 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. The more general condition, *missing at random* (MAR), is 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]

Under fairly general conditions, if data are missing at random, they can be imputed using regression-type models fit to observed data [Rubin 1976]. In practice, these are the sort of models that are generally fit, even though realistically we might expect the missing-at-random assumption to be wrong: in many cases, the level of an unobserved variable is a predictor of its exclusion.

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 standard in practice, however, to impute using available data and 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. For a simple example, suppose we had data on the heights and weights of 450 of the 500 boys in a high school and fit a simple missing-at-random model to impute the heights of the remaining 50 students. Further suppose that the observed and imputed heights appear to follow a normal distribution but with a truncated right tail. Given the general knowledge that heights follow a normal distribution, it would be natural to suspect the model underlying the imputations, and it would be appropriate to examine the data collection—perhaps, for example, the basketball and volleyball teams were traveling on the day that the measurements were taken. See Figure 2 for an illustration.

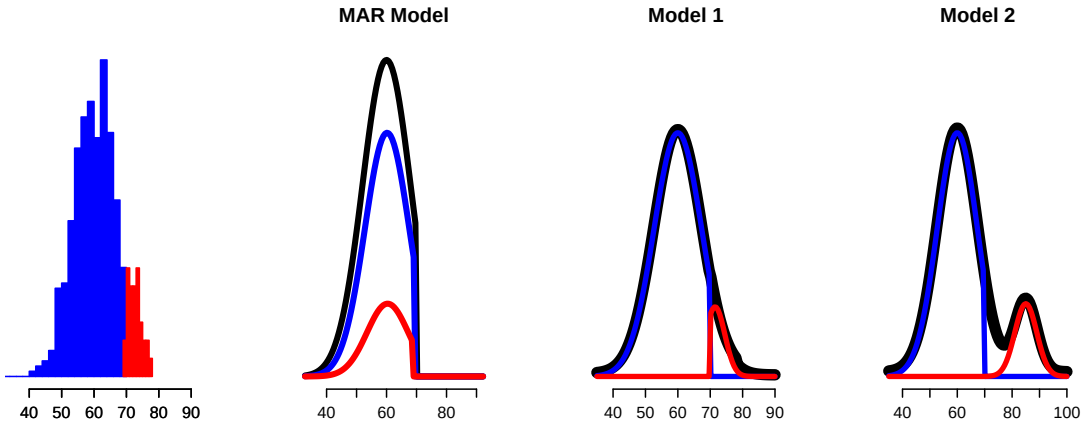

Figure 2: Histograms and imputations of truncated observations of heights of students: The tallest students are missing, not at random, because of a volleyball game. The histogram (at left) has the missing observations in red. Imputations (in red, second graph from left) generated under the MAR assumption would, here incorrectly, follow the distributions of the observed values in blue. The completed dataset is in black. Model 1 would be flagged because the imputed data markedly differ from the observed. The imputations, in red, are improperly generated from the truncated values only. Model 2 would be flagged as well because the completed data distribution is bimodal.

For another setting, sometimes selection models are used for sensitivity analysis. The constructed completed dataset can then 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 data set) 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 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 data set 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.

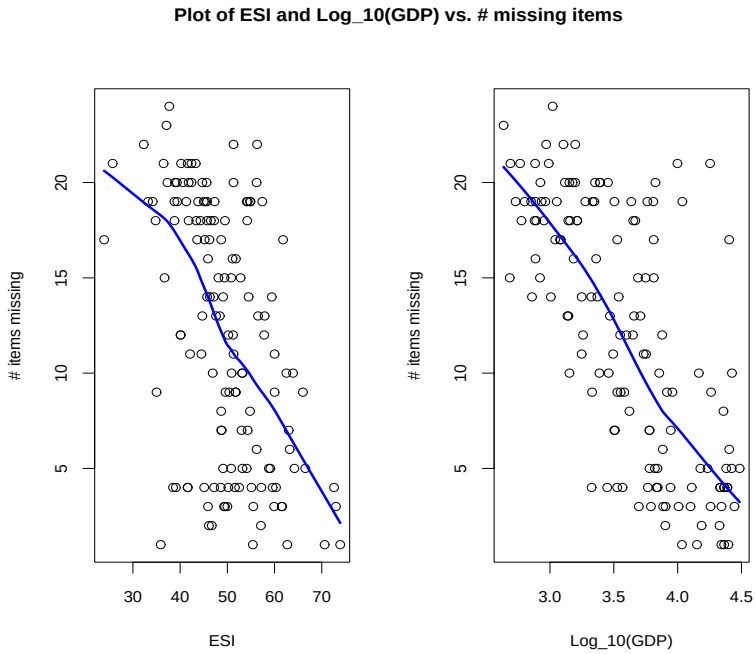

Figure 3: For each country, percent of variables missing, plotted vs. ESI and GDP, with fitted lowess lines [Cleveland 1979]. Countries with higher environmental sustainability indexes and higher incomes tend to have fewer missing items. The graphs clearly demonstrate that the variables are not missing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

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 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.

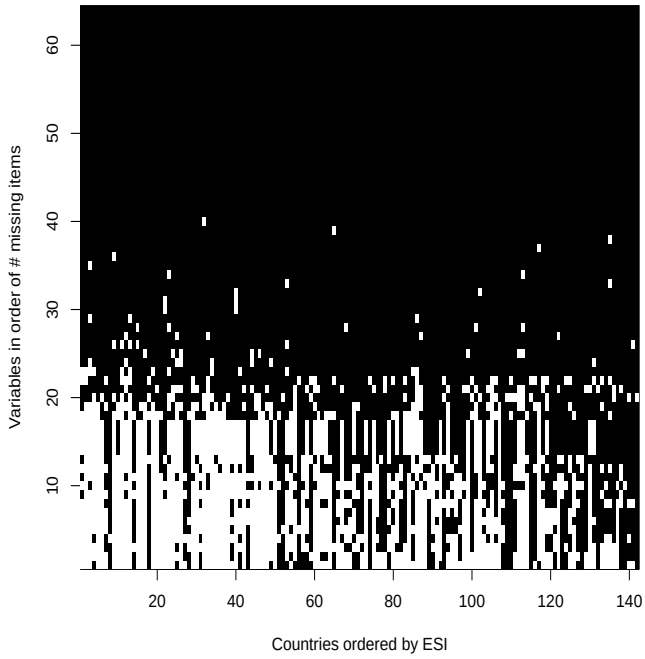

Figure 4: The pattern of missingness—missing values in white. Countries are listed in rank order of ESI, variables are listed in order of number of missing items in the ESI. (Kuwait is the first country on the abscissa, Finland is the last. WATCAP (water capacity) is bottom on the ordinate, GMS.SS (suspended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

[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.

2 Methods

2.1 Multiple Imputation: The SRMI procedure

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. For simplicity we shall work with just a single randomly-chosen imputation in our example.

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 et al. [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.

In the end, we imputed approximately 19% of the data for the ESI using this procedure. Several variables were considered unfit for imputation before the imputation procedure, some after, and these were removed from the definition of the ESI. We imputed a total of 10 complete data sets 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 violation of the missingness assumptions. In both cases the flagging procedures compare the imputed values to the observed. In the sense that the completed data set is model generated, these are tests of the imputation mechanism. A raised flag indicates a 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, like the SRMI procedure, are particularly amenable to multivariate model checking. It is a misconception that the non ignorable missingness assumption (NMAR) 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 data set, which suggests the existence of hypotheses about characterizations of a complete data set. The point is that imputation modelers usually have a notion about what this complete data set 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 can say this and we can say it several times: differences in distribution between the imputed and the observed *do not necessarily* indicate violations of the missingness assumptions or problems with the imputation model. 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 significant differences. 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 observed 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.

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 in the distributions of missing and observed values, given 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.

For the ESI example, we use a maximum of five comparisons for each variable: 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 Fixing: 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 (Y_j) regressed on all others ($\mathbf{Y}_{-j}$). For each imputation, 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. For each predicted value $\hat{y}_{ij}$, we compute its residual r_{ij} as the difference from the observed value in the completed data; for the imputed data this is output of the SRMI model y_{ij}.

Differences in residual plots—between the imputed and observed—indicate where our complete data model (here, the best stepwise regression) is insufficient. Under the model, the pattern of residuals versus expected values should be random: we generate the imputations from a series of linked, linear regressions. We can treat a non-random pattern in the residuals as a calibration of the best available linear model, and the penultimate step in the imputation procedure.

2.3.2 Fixing the imputations

The aim here is to refine the complete data model, and thus bolster the second MI assumption. 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 and the SRMI output. We then update the imputations only, using the lowess curve as the proper residual function. To restate: We correct (or calibrate) the imputed values by supposing a function from the predicted (of the observed) to the residuals (of the observed) and forcing the residuals of the imputed to match that function.

At each variable we can judge the propriety of our calibration by comparing the distributions of the observed with the distributions of the imputations—corrected and uncorrected. We believe that refining the imputations via the residual plots is justifiable where differences have been flagged by alternate tests, and where the differences in distribution are likely artifacts of the complete data model. This refinement is based upon observed or available data. It is plausible that the differences between observed and imputed data are true and a success of the imputation model. In that case meeting the MAR assumption is not an issue but the fit of the multivariate model still is - and the conditional residual plots should have flat expectations (at zero). If the residuals do not have zero expectation conditional on fitted values, we suspect a failure of the imputation model and

the recalibration is an approach to a fix. Alternately, where violation of the MAR assumption is *the* issue, we should not expect much from the residual recalibration, given the battery of tests. 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.

3 Application

We shall illustrate the proposed methods with the data and imputations for the Environmental Sustainability Index. Our first step is to display the observed and imputed data for all imputed variables, versus the overall index, as shown in Figure 6. We shall first discuss these scatterplots and then demonstrate the methods for each group of variables in the ESI.

There are plausible explanations for the differences in scatterplot patterns observed versus the ESI, beyond violations of complete random missingness. 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 dissent from environmental theory. An example is the BODWAT variable—a measure of industrial pollutants in available fresh water. An outlier in the observed distribution (Kuwait) attenuated the effect of an improper complete data model at this variable. The imputed values of BODWAT were characteristically, and unjustifiably, greater than the observed values.

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 the subsections, via (what we believe could be) semi-automatic processes, is 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 four variables as extremely problematic ($p < .001$): NO2 concentration (NO2), radioactive waste (NUKE), child death rate from respiratory diseases (DISRES), and total marine fish catch (FSHCAT).

Within the ESI component groupings, we select the above variables. At each group we choose one variable that did not significantly differ with one that did, for purpose of illustration.

The KS tests indicated significant differences for all the variables in this component, save SO2 concentration and percent of birds and mammals threatened by human activity. This may, or may

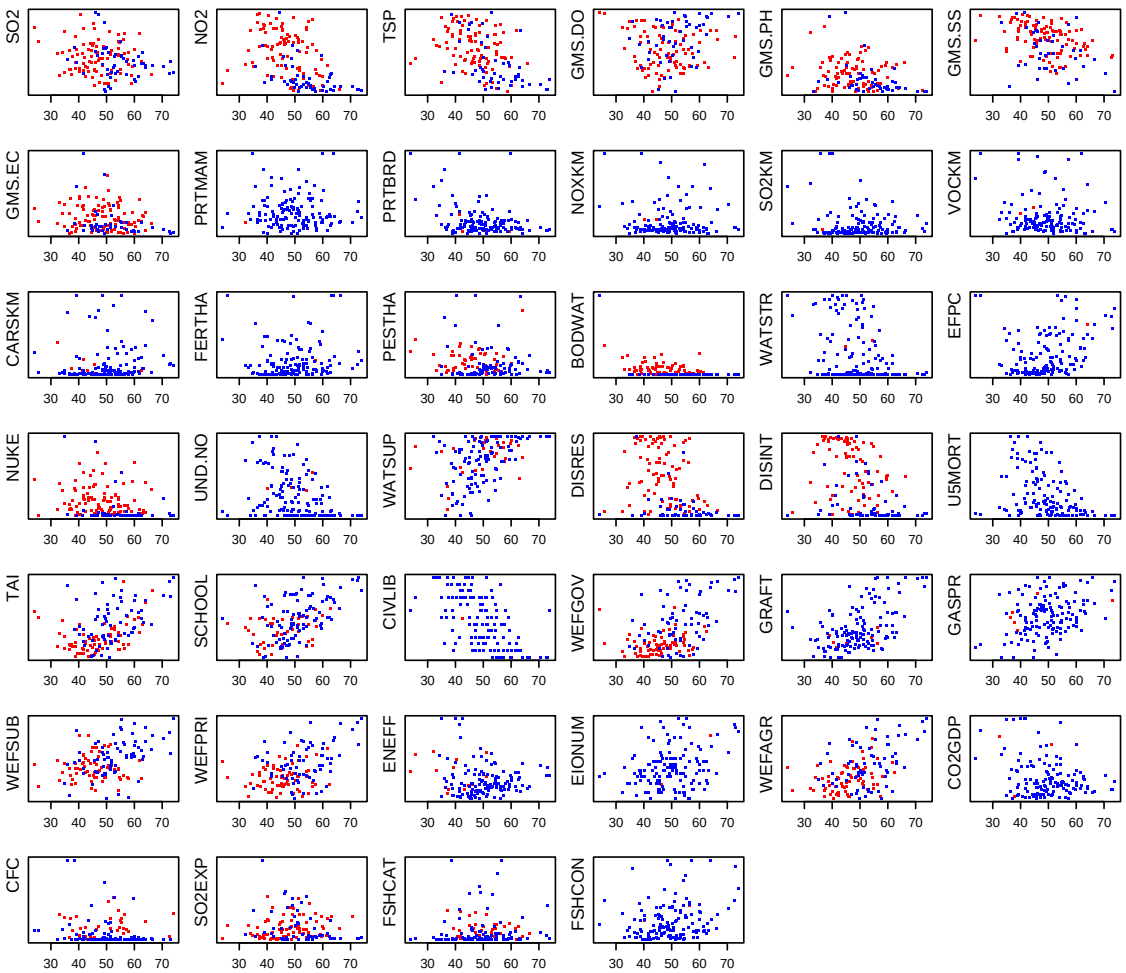

Figure 5: For each variable, the observed and imputed values of the versus the Environmental Sustainability Index. Imputed values, everywhere, are in red. Observed values are blue. At a glance there is evidence for non random patterns of missingness in many variables, as discussed in detail in the text.

not, indicate a problem with the validity of the indicator. We can state, on the one hand, that in absence of the knowledge of the missingness mechanism, perhaps the observed difference in distribution is a correct function of differences in the predictors. On the other hand, here especially, that statement may not be defensible: first, we expect that some countries may misreport or restrict—intentionally or not—air and water concentration data; secondly, we believe that anomalies in distribution, in a few cases, are caused by just a few influential cases. Extreme outliers in the distributions of WATCAP and WATINC (internal water capacity and per capita inflow) are

idiosyncratic: Kuwait, for example, imports most of its water.

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

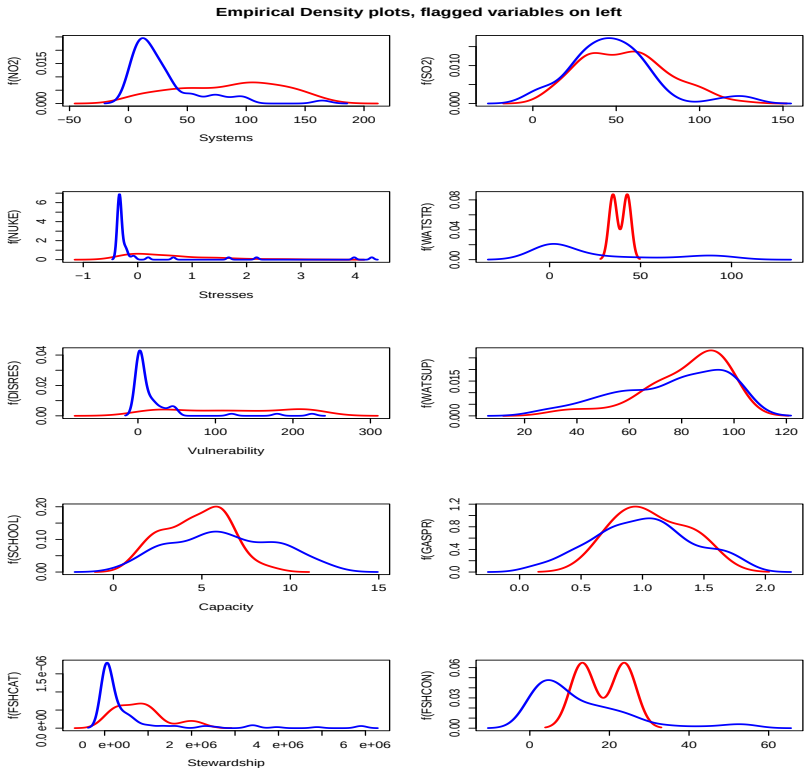

Figure 7: The left graphs the variable whose imputed values (red) differ significantly from observed values (blue). For comparison, variables which we did not flag by the KS tests are shown on the right side. Possible flaws in imputations may appear, in the graphs, even when not indicated by the KS tests.

3.1 Environmental Systems

We discuss each set of variables (see Figure 6) in turn.

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 comparable across nations in the sense that the true values are easily observable and calculable. However, this component had the highest number of missing items. 36 percent of the data were missing.

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.

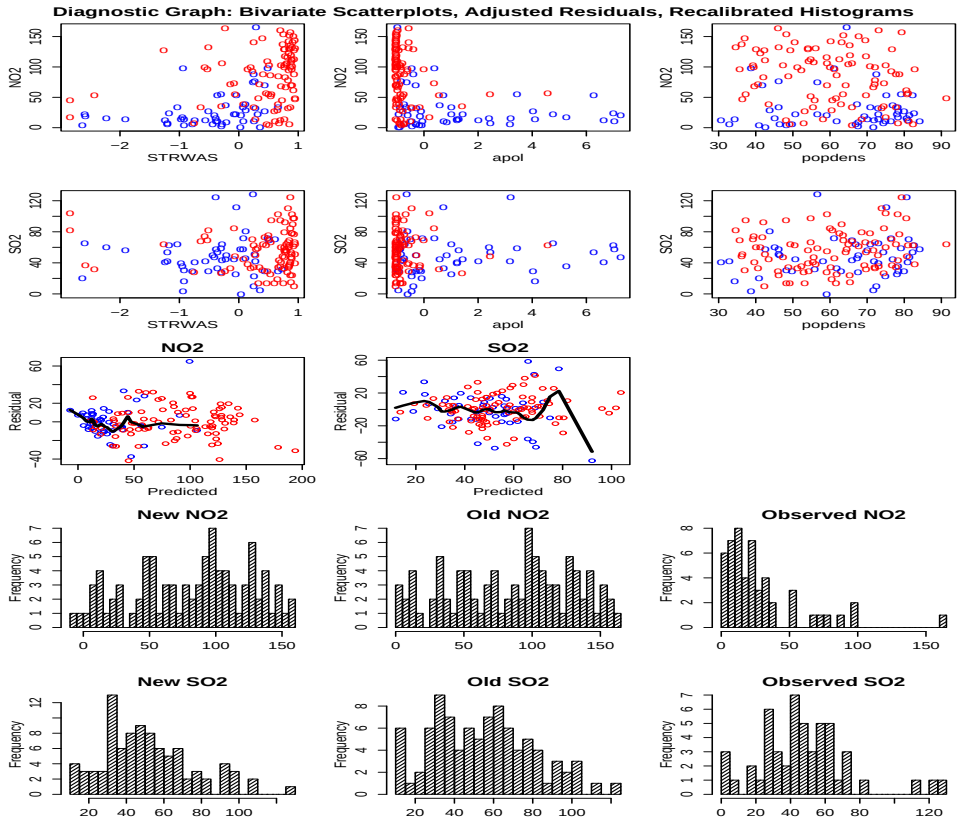

Figure 8: Environmental Systems—NO2 is flagged as significantly different by KS test. Bivariate scatterplots highlight distributional differences. SYSAIR is a composite of air quality measurements used in the ESI. APOL is a composite of air quality measurements not used in the ESI. POPDENS is a measure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

The difference in the distributions between observed and imputed for NO2 appears to be driven by over prediction 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. We can see that 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 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 loess 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.

3.2 Environmental Stresses

The environmental stresses, in contrast to those in environmental systems, are measurements on rates of degradation—possible predictors of future environmental conditions. The number of missing items is much lower than in environmental systems, again in contrast: only 11 percent of the data are missing from this component.

The imputations are significantly different by KS test in only three places: PESTHA—pesticide use per hectare, BODWAT—industrial organic pollutants per available fresh water, NUKE—radioactive waste. Here, the KS results may or may not be indicative of a problem with the imputations. Similarity in the distributions of observed and imputed does not imply validity; though we take dissimilarity to be a flag for closer inspection, case wise verification of imputed values may be appropriate where dissimilarity is non significant; where the number of flagged variables is lesser, it may be appropriate to make observation level (not variable level) adjustments.

We looked at the imputations for NUKE more closely and took those for WATSTR (percentage of country under severe water stress) for comparison. The imputed distribution of NUKE significantly differs from the observed—the imputed distribution is only slightly peaked below zero.

The picture here can be compared with NO2 in environmental systems. See Figure (9). The observed values for NUKE are generally very low except for a few extreme high values. The distribution of the imputed values is less dramatic with a similar, though less pronounced shape.

The high number of missing items in NUKE is immediately apparent from the bivariate plots, as is the starkly different imputed distribution. The similarity in imputed and observed distributions of WATSTR comforting but not remarkable: just a few values are imputed.

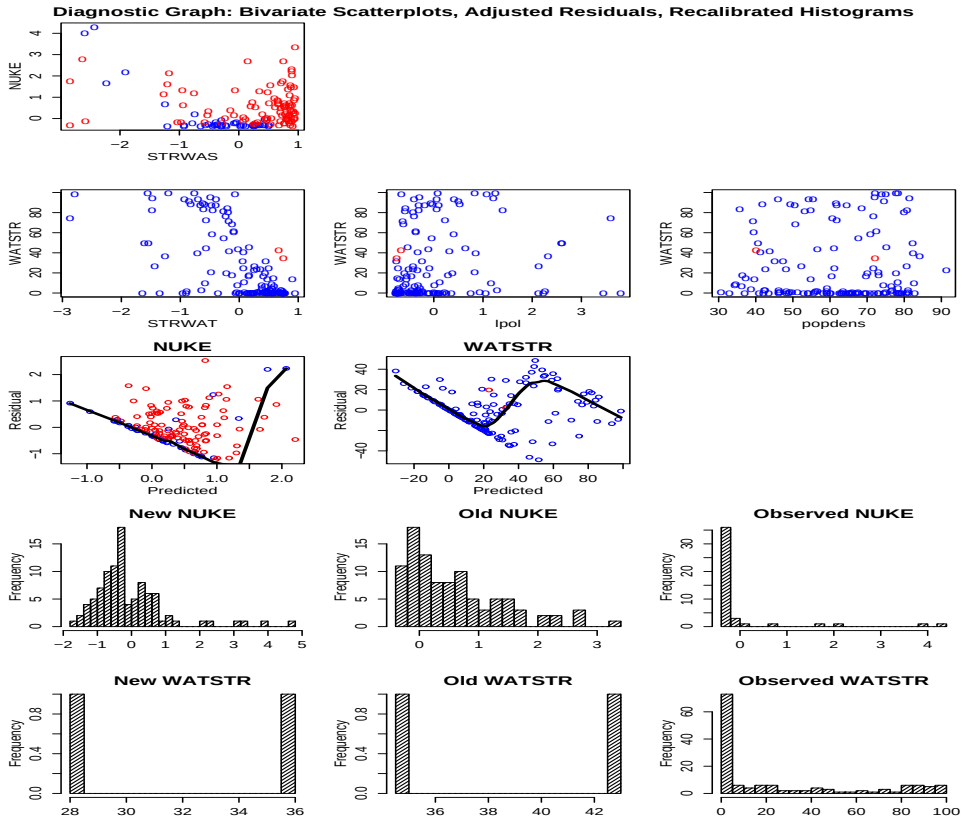

Figure 9: Environmental Stresses—NUKE is flagged as significantly different by KS test. Bivariate scatterplots highlight distributional differences. STRWAS is a composite of waste disposal measurements used in the ESI. STRWAT is a composite of water stress measurements used in the ESI. LPOL is a measure of land use policy. POPDENS is a measure of population density. The residual plots plot the predicted values from the best stepwise regressions against the difference between a single random imputation and its predicted value. Histograms of the updated imputations are on the final rows.

The residual plot illustrates that the imputed values for NUKE are significantly greater than the observed values.

The histograms illustrate that the fix works to improve the similarity of the imputed distribution for NUKE, and has little effect upon the imputed distribution for WATSTR.

3.3 Vulnerability

The vulnerability data are measures on the social stocks of environmental sustainability. The number of missing items is quite high, 28 percent, though proportionately lesser than in environmental systems. We imputed values for all five of the variables here.

Just two of the imputed values are similar by KS test: WATSUP-percent of population with access to improved drinking water and U5MORT-under five mortality rate. We apply our fix to DISRES—child death rate from respiratory diseases and WATSUP, for comparison.

The distribution of the imputed values is not as skewed as that of the observed. The density plot for DISRES imputed radically differs from the observed—especially in comparison to the density plots for WATSUP. The imputed values span a greater range than the observed with a less pronounced mode.

The illustration is more vivid in the bivariate scatterplots. The imputed values of DISRES are almost uniformly greater than the observed values on all comparisons. The imputed values for WATSUP match the observed much more closely.

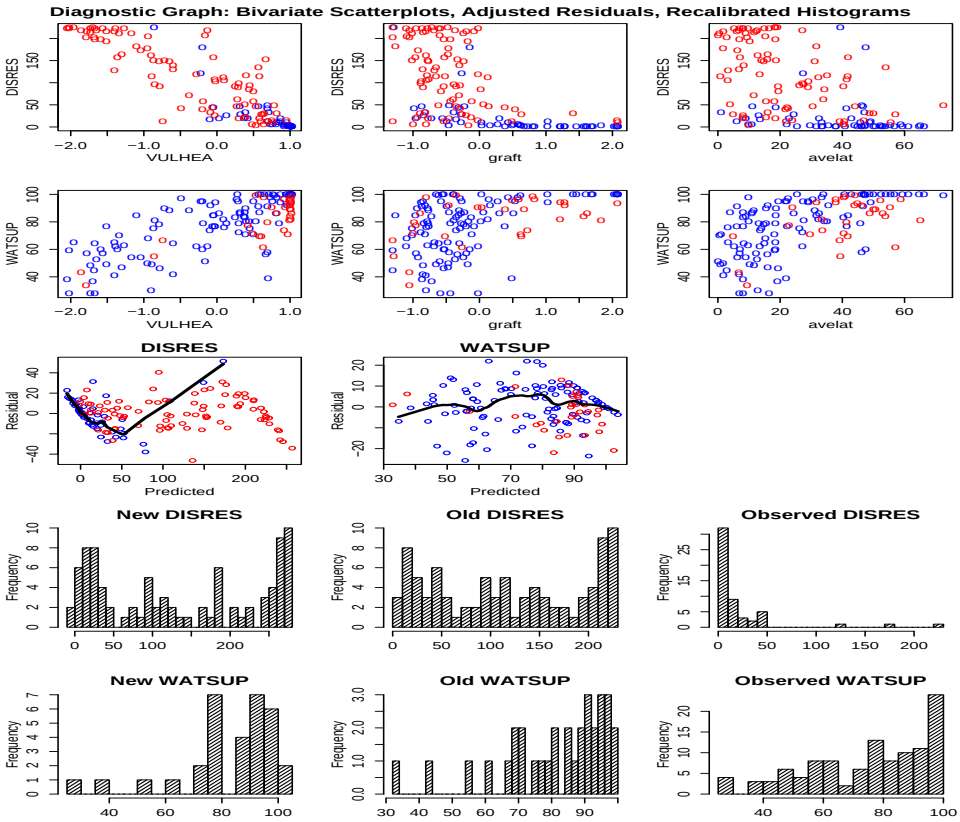

Figure 10: Vulnerability—DISRES is flagged as significantly different by KS test. Bivariate scatterplots highlight distributional differences. VULHEA is a composite of health vulnerability measurements used in the ESI. GRAFT is a World Bank composite index of corruption used in the ESI. AVELAT is the average latitude of the country. VULSUS is a composite measure of vulnerability to sustainability used in the ESI. The residual plots plot the predicted values from the best stepwise regressions against the difference between a single random imputation and its predicted value. Histograms of the updated imputations are on the final rows.

The residual plot for DISRES illustrates that the imputed values are over predictions within the range of the observed values. Beyond the range of the observed, the pattern is less clear, however, the lowess curve is drawn to a high observed value.

The refined imputations match the observed values more closely, especially at the mode, within the range of the observed values. Both the refined and original imputations are over represented beyond the range of the observed—a possible problem we noted from the residual plots.

3.4 Capacity

Capacity measurements are on possible predictors of future stocks of social welfare. The capacity measure had the lowest proportion of missing items—only 15 percent of the data are missing.

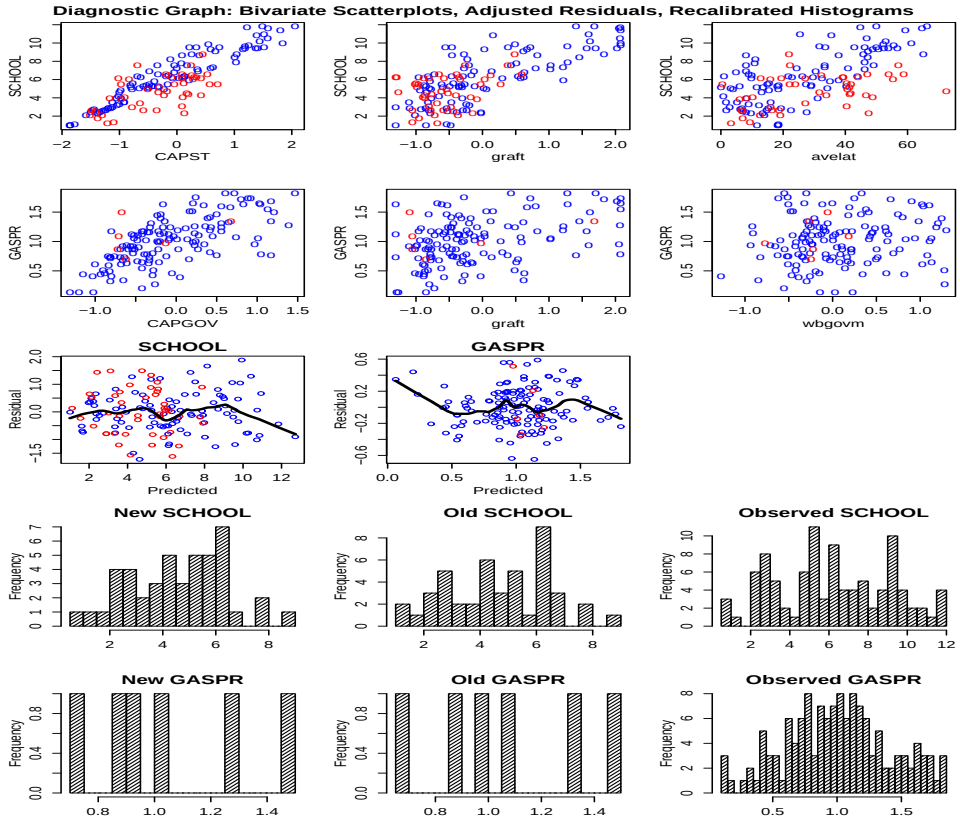

Figure 11: Capacity—SCHOOL is flagged as significantly different by KS test. Bivariate scatterplots highlight distributional differences. CAPST is a composite of science and technology measurements used in the ESI. GRAFT is a World Bank composite index of corruption used in the ESI. AVELAT is the average latitude of the country. CAPGOV is a composite measure of environmental governance used in the ESI. The residual plots plot the predicted values from the best stepwise regressions against the difference between a single random imputation and its predicted value. Histograms of the updated imputations are on the final rows.

Five of the ten imputed variables differ by KS test from the observed values. SCHOOL—mean years of schooling, and GASPR—ratio of gasoline price to international average are selected for fixing and comparison.

The imputed distribution of SCHOOL is more skewed than that of the observed. While the imputed values of SCHOOL are more tightly distributed than the observed the general shape of the frequency curves are quite similar. The density curves for the GASPR variable, imputed and observed, are nearly identical.

The imputations for SCHOOL do not appear to differ greatly on any of the comparisons. The number of imputations for GASPR is low. The similarity between the imputations and the observed cannot be disputed by inspection of the bivariate scatters.

The predicted values for SCHOOL are even about the residual lowess line, though at the lower range. Again, GASPR was imputed only sparsely. The refined imputed distribution for SCHOOL does not appear visually different from the imputed. The refinement for GASPR is uninformative.

The refinement procedure had little effect upon the shape of distribution of SCHOOL. The range of the new imputed values is slightly wider. The difference, by histogram, for GASPR is undetectable.

3.5 Stewardship

The data in the stewardship component measure international cooperation and are generally more widely available and comparable. The proportion of missing items is the least of all the components, 11 percent. FSHCAT—total marine fish catch and FSHCON—seafood consumption are selected to review.

Four of the six imputed distributions differ by KS test. The imputed distribution (imputation seven, as before) is much less skewed than that of the observed. The imputed values for FSHCAT have a shorter range and are less tightly distributed about the modal value. The approximated density curve for the FSHCON imputations is influenced by the low number of values and inconclusive.

The difference in distribution for FSHCAT is marked with respect to GDP. The imputed values are generally for countries at the lower end of the GDP range—evidence for violation of the MCAR assumption. The scatterplots for FSHCON are nondescript.

The imputed values of FSHCAT appear to be over predictions when compared with the observed values. The loess curve for FSHCON is irregular. The distribution of the refined imputations should be truncated, negative values are invalid. The observed distribution has a much greater range than the imputed: imputations here severely differ from the observed; the refinement worsens the difference.

The histogram for FSHCAT indicates that the refined imputations match the observed values more closely in shape, though the difference in range is exacerbated. FSHCON differs little after refinement—there are too few values for detection.

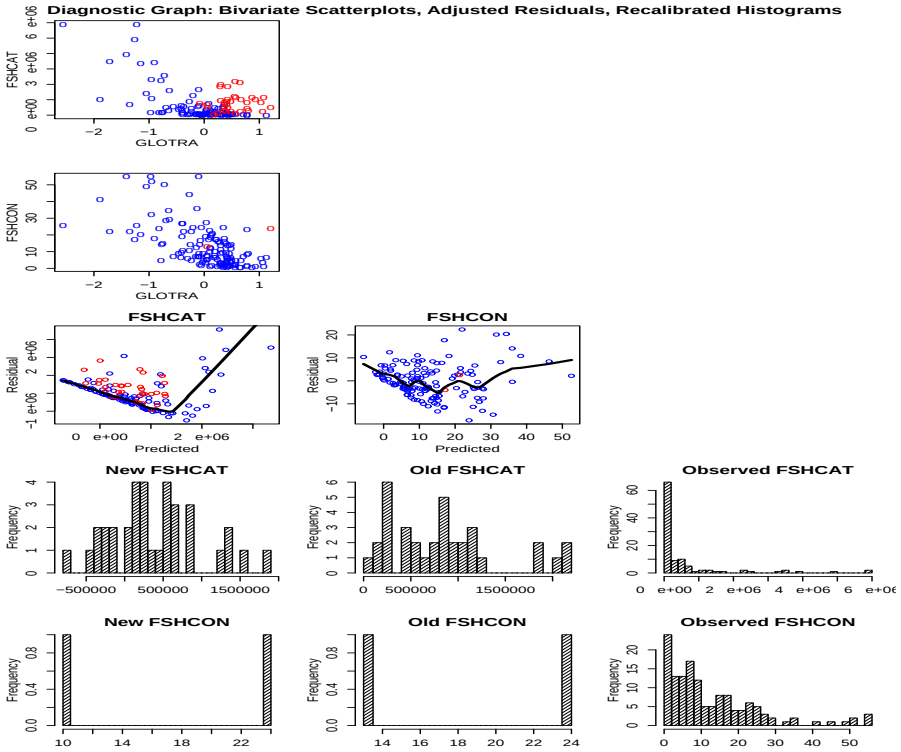

Figure 12: Stewardship—FSHCAT is flagged as significantly different by KS test. Bivariate scatterplots highlight distributional differences. GLOTRA is a composite of global trade measurements used in the ESI. The residual plots plot the predicted values from the best stepwise regressions against the difference between a single random imputation and its predicted value. Histograms of the updated imputations are on the final rows.

4 Discussion

Missingness in the ESI arises from the dearth of environmental metrics and 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 which highlighted obvious flaws. We began with numerical diagnostics—the Kolmogorov Smirnov tests—to flag problems, and we attend 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 2001].[2] 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 investigate 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.

A Appendix

A.1 Computation of the ESI

The environmental sustainability index [World Economic Form 2002] is defined as

$$\text{ESI} = 100 * \Phi \left(\frac{1}{|k|} \sum_{k} \frac{1}{|J_k|} \sum_{j \in J_k} \left(\frac{Y_j - \overline{Y}_j}{\text{var}(y_j)^{1/2}} \right) \right).$$

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.

[2] Think of the graph arrays—Figures 8 through 11—for each of the components, as a complementary, necessary diagnostic output to a completed dataset for any imputation software.

A.2 Missingness Assumptions

Extending the equation 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

$$f(M|\mathbf{Y}, \phi) = f(M|\mathbf{Y}_o, \phi) \forall \mathbf{Y}_m, \phi$$

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 $\exists \mathbf{Y}, \phi, s.t., f(M|\mathbf{Y}, \phi) = f(M|\mathbf{Y}_m, \phi)$.

Not of the variables in the ESI are complete across observations-the total rate of missingness is 22 percent. Constructing the index on available cases only is a severe restriction. Further, weighting the index using randomization inference on the observed data is inappropriate given an apparent violation of the MAR assumption.

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:

$$\mathbf{Y} = (\mathbf{Y}_{J1}, \ldots, \mathbf{Y}_{Jk}) = (Y_1, \ldots, Y_{68}) = (\mathbf{Y}_m, \mathbf{Y}_o) = \mathbf{Y} = (Y_1, \ldots, Y_{|k|-r}, Y_{|k|-r+1}, \ldots, Y_{|k|})$$

$$\mathbf{X} = (Y_1, \ldots, Y_{|k|-r}); \mathbf{Y}^* = (Y_{|k|-r+1}, \ldots, Y_{|k|})$$

in order of missingness, where r is the number of variables with missing values, and $\mathbf{Y}^*$ 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.[3].

Let $\hat{y}_j = \hat{G}(\mathbf{Y}_{-j})$ be the predicted values from the imputation model for each variable j,

Take

$$H(\hat{y}_j, y_j) = \hat{G}(\mathbf{Y}_{-j}) - y_j = \tilde{y}_j$$

Then $\tilde{y}_j$ are the refined imputations when the arguments to the above are the imputed values.

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:

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

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**—$z_{i,j} \sim logit^{-1}(a2_j + (y_{i,j} - b2)/c2)$.

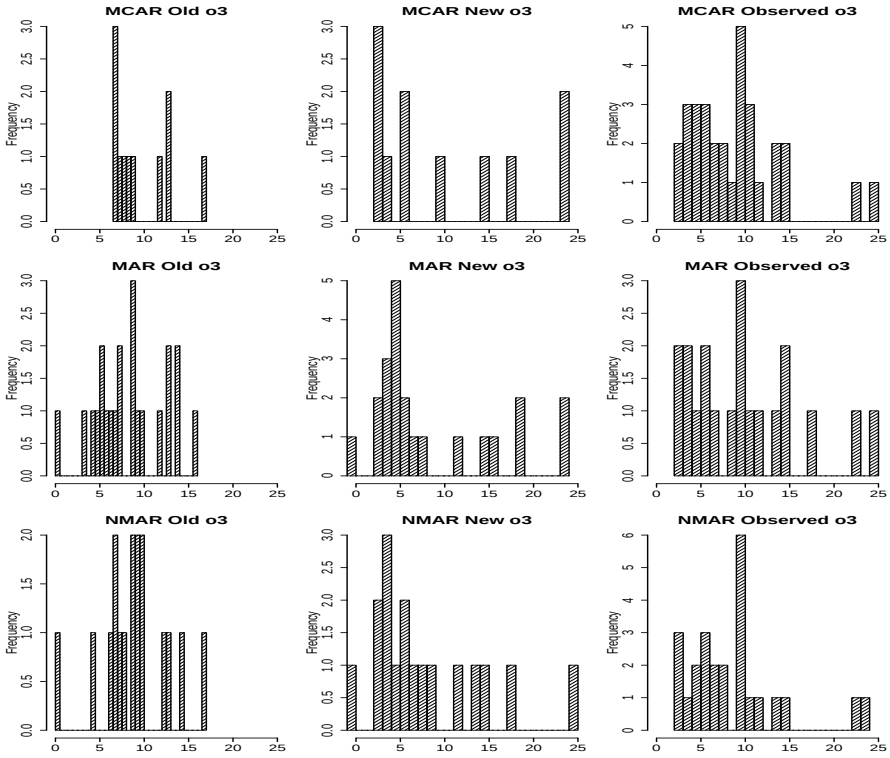

Figure 13: Simulated imputation refinement on air quality data. The first two graphs in each row are the distribution of O3 before and after recalibration. The last graph in each row are the observed data. The first two rows are imputations and recalibrations under MCAR and MAR models. The refinements 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.

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 assumptioon - and more so for the imputations on the MCAR assumptions.

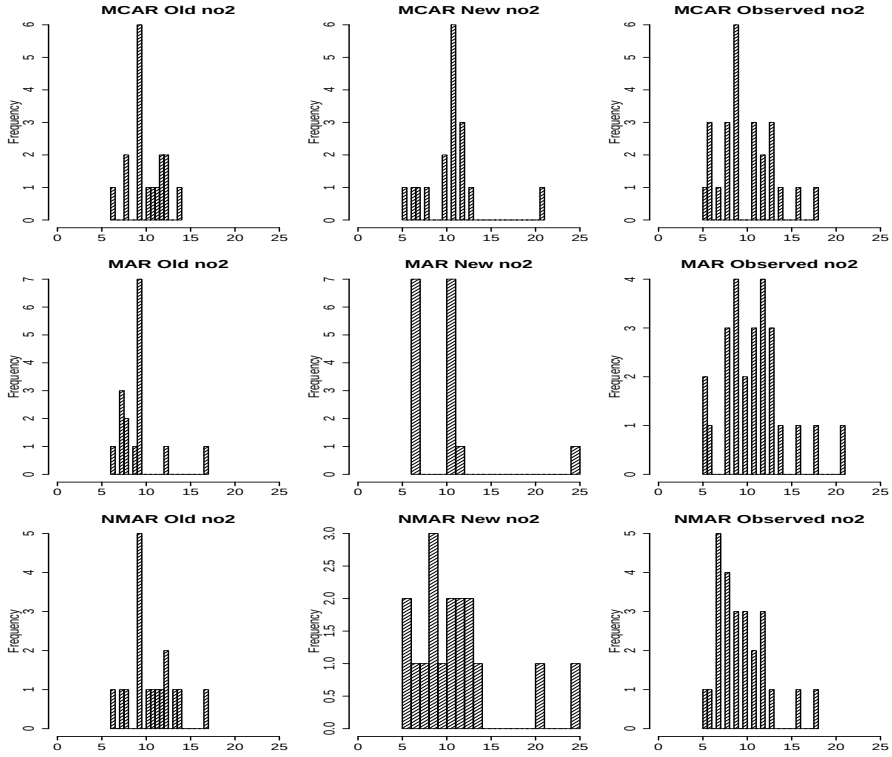

Figure 14: Simulated imputation refinement on air quality data. The first two graphs in each row are the distribution of NO2 before and after recalibration. The last graph in each row are the observed data. The refinements match the distribution of the observed better than the original imputations under 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. (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. *The 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 *to appear.*

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 (S software for missing-data imputation).
`web.inter.nl.net/users/S.van.Buuren/mi/`

World Economic Forum (2001). 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.

World Economic Forum (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. New York.