Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2008)
57, *Part* 3, *pp.* 273–291

Diagnostics for multivariate imputations

Kobi Abayomi, Andrew Gelman and Marc Levy

Columbia University, New York, USA

[Received July 2004. Final revision November 2007]

Summary. We consider three sorts of diagnostics for random imputations: displays of the completed data, which are intended to reveal unusual patterns that might suggest problems with the imputations, comparisons of the distributions of observed and imputed data values and checks of the fit of observed data to the model that is used to create the imputations. We formulate these methods in terms of sequential regression multivariate imputation, which is an iterative procedure in which the missing values of each variable are randomly imputed conditionally 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, which is a linear aggregation of 64 environmental variables on 142 countries.

Keywords: Environmental statistics; Missing values; Multiple imputation; Multivariate statistics; Sustainability

1. Introduction

When considering models to impute missing data, the hypothesis of missingness at random (MAR) can, inherently, never be tested from observed data. However, any specific imputation model, whether the mechanism is MAR or not, will be fitted 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. In this paper, we consider imputation methods from the perspective of multiple imputation, where imputed values are drawn from a predictive model for the variable that is missing given all other variables. We shall focus on diagnostics for a single imputed data set and, in principle, our approach could be applied when single (rather than multiple) imputation is employed, provided that the imputed values are obtained from a predictive modelling approach.

Additionally, the completed data sets can be checked for plausibility, though this is not a formal hypothesis test since 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—it is a means of diagnosing possible problems with the imputation model.

1.1. Missingness

Multiple imputation has become popular in the 30 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 *et al.* (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, generally,

Address for correspondence: Kobi Abayomi, Department of Statistics, Columbia University, New York, NY 10027, USA.
E-mail: kobi.abayomi@columbia.edu

 0035–9254/08/57273

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.

(a) Imputations can be checked by 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 whether they make sense in the context of the problem being studied.
(b) Imputations are typically generated by using models (such as regressions or multivariate distributions) that are fitted to observed data. The fit of these models can be checked.

Diagnostic techniques do exist: we can characterize them as *external*—comparisons with outside knowledge—or *internal*—specific to the observations and modelling. This paper 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 data set that was constructed by using a multiple-imputation procedure. The completed data were used to construct an index of environmental sustainability. We believe that 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 environmental sustainability index

The environmental sustainability index (ESI) was created as a measure of overall progress towards environmental sustainability and was designed to permit systematic and quantitative comparison between nations (World Economic Forum, 2002). The ESI is a scaled linear combination of 64 variables of environmental concern. Environmental measures (such as oxide emissions and concentration) are included along with political indicators (such as civil liberty and level of corruption) that are relevant to environmental sustainability (World Economic Forum, 2002).

The ESI, like other indices of environmental concern (such as the environmental wellbeing index and the human development index), condenses dissimilar social and physical metrics into cohesive summaries for national level comparisons (Prescott-Allen, 2001; United Nations Development Program, 2002). Our goal for the 2002 ESI was to capture the most recent version of available data to obtain the best snapshot for 2001. Our approach was to use the most recent year that is available for each variable at each country.

The breadth of the ESI—64 dissimilar variables from varied sources—presented aggregation and processing challenges beyond missingness. Some variables are composites of information from several sources: pollutant yield divided by land area conditioned on population density, for variables in the 'environmental systems' and 'environmental stresses' indicators—for example. Others may be imprecise across observations: mortality and disease variables in the 'vulnerability' indicator, for instance. See Annex 1 and Annex 2 of the 2002 ESI report for elucidation (World Economic Forum, 2002).

The ESI can be partially disaggregated across measurably similar groups of variables (components):

(a) *environmental systems* (13 variables)—measurements on the state of natural stocks such as air, soil and water;

(b) *environmental stresses* (15 variables)—measurements on the stress on ecosystems such as pollution and deforestation;
(c) *vulnerability* (five variables)—measurements on basic needs such as health, nutrition and mortality;
(d) *capacity* (18 variables)—measurements of social and economic variables such as corruption and liberty, energy consumption and rate of schooling;
(e) *stewardship* (13 variables)—measurements of global co-operation such as treaty participation and compliance.

1.3. The environmental sustainability index and missingness

As noted in World Economic Forum (2002),

> 'missing data are an endemic problem for anyone working with environmental indicators'.

Environmental data are often dissimilarly reported across regions or nations—rendering their 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. Imputation allows estimation of a complete-data statistic, by fitting a complete-data model to the observed data.

A variable is *missing completely at random* 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 MAR, where the probability that 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* gross domestic product, and this gross domestic product 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 *missingness not at random* (MNAR) (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 by using regression-type models that are fitted to the available data under the MAR assumption, with the understanding that these imputations, although imperfect, may be useful, especially if the fraction of missingness in the data set is small.

In principle, it is impossible to test the assumption of MAR 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 thus 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 tonnes of biochemical oxygen demand emissions per cubic kilometre of water) (World Economic Forum, 2002). Most of the observed data are of the order of 10^{-1}–10^{1}. The exception is for Kuwait, which, as a net importer of freshwater, is at 10^{9}. Under the general MAR assumption, one imputation draw is at the right-hand tail of the observed distribution. The imputation model is sensitive to this outlier; 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 Fig. 1. 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 Fig. 1, the completed data (histogram) in Fig. 1(a) are bimodal. Observed data are shown in blue, imputations in red and completed data in black. The histogram (Fig. 1(b)) has the imputed data, from one draw, at the right-hand tail of the distribution. The observed outlier is rightmost and blue. Imputations that are generated under this model are incorrect. The model would be flagged because the imputed data markedly differ from the observed data. A *post hoc* plot of the completed data illustrates the problem: the influential outlier in the imputation model (blue at the upper left-hand side of Fig. 1(c)) is Kuwait. Available observed data for cases where BODWAT is imputed may be similar to those for Kuwait; the imputation model at this variable has, incorrectly, low precision. This example illustrates where a diagnostic method can highlight problems in an automated imputation procedure: here, as is common in default imputation models, the normal distribution imputes values near the arithmetic mean. The extreme outlier exaggerates this effect. The imputation algorithm cannot know that Kuwait would be a problem; the *post hoc* diagnostic flags the problem with the imputation model.

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 Fig. 1) is correct; the model may fit the observed but not the missing data. In these cases, the diagnostic in Fig. 1 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 data set 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 *et al.*, 2004).

The graphical displays that have just been described are *external* (in the sense of the observed data set) diagnostics of an imputation procedure. There is no *internal* test of MAR (or, for that matter, of whatever MNAR model that might be used). However, internal tests can be performed of the imputation model itself, in the context of the observed data that are 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 to recalibrate if residuals do not have mean value 0 conditional on available predictors. Our general procedure is to use external tests to flag possible problems which then must be checked by using subject matter knowledge. Internal tests can be performed more automatically, by analogy with regression diagnostics.

These examples illustrate where and how external tests motivate inspection of the multivariate model that is used to generate the imputed data. Remember that the goal is not data modelling, but generation of a (completed) data statistic. In both of the cases illustrated, 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. For imputation purposes, the multivariate model, even when implicit, can and should be checked by using comparisons of observed and imputed distributions; under a default assumption modelling idiosyncracies are distinguishable. Indeed, *a fortiori*, using the completed data set to check

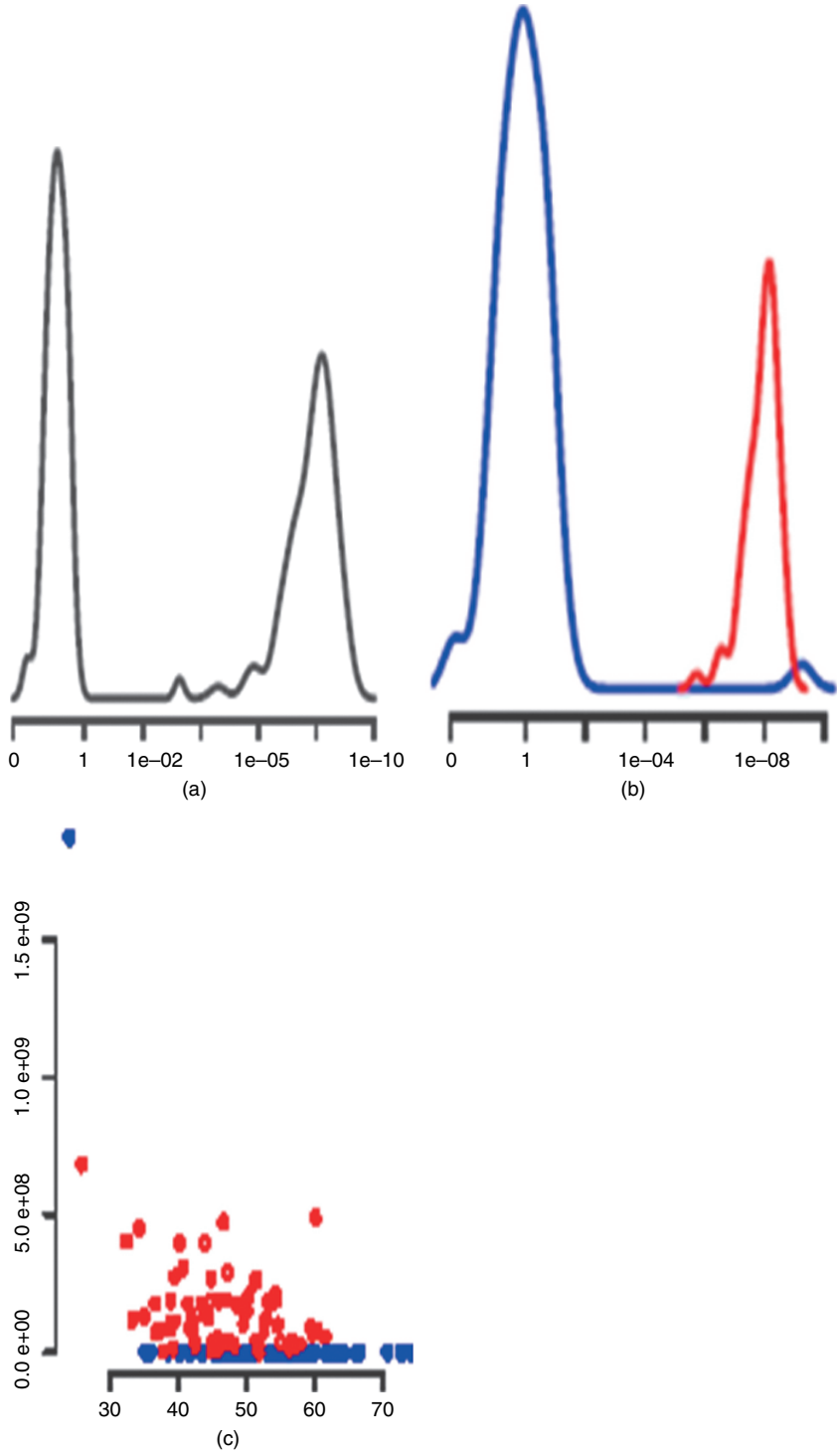

Fig. 1. (a) Completed and (b), (c) observed data for BODWAT (the axes have been transformed for illustration), with imputations based on a fitted normal distribution

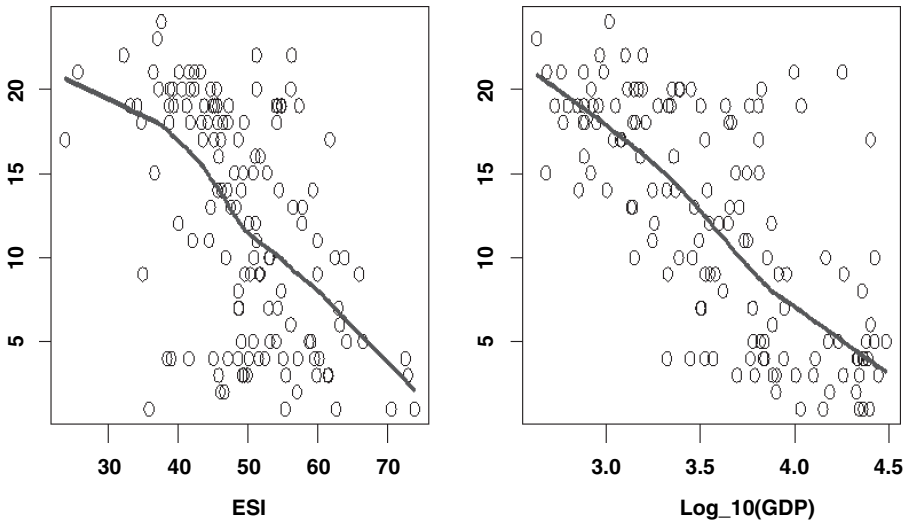

Fig. 2. For each country, the percentage of variables missing is plotted against (a) ESI and (b) gross domestic product, with fitted LOWESS curves: countries with higher ESIs and higher incomes tend to have fewer missing items; the graphs clearly demonstrate that the variables are not missing completely at random

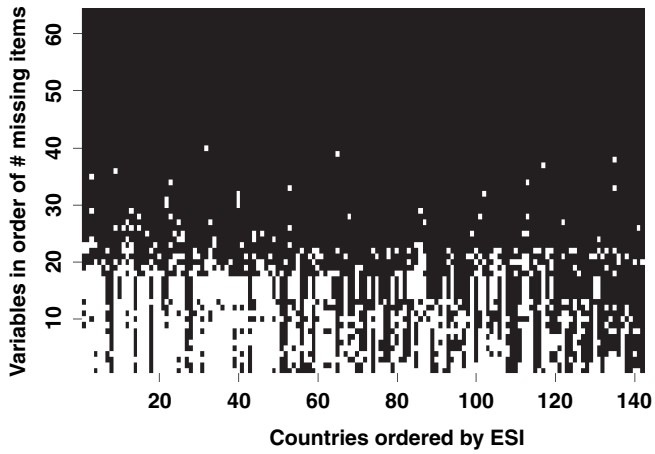

Fig. 3. Pattern of missingness (missing values are in white): countries are listed in rank order of ESI (Kuwait is the first country on the abscissa; Finland is the last); variables are listed in order of number of missing items in the ESI; on the bottom, with 101 missing values, is the 'Global environmental monitoring system' suspended solids variable

the imputation model should flag, at least, where the modelling may be inappropriate—if not explicitly where the missingness assumptions are not met.

1.4. Missingness in the environmental sustainability index

As is shown in Fig. 2, the countries with low ESIs and low incomes tend, unsurprisingly, to have more missing items in the ESI. (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.) (ESI and *per capita* gross domestic product are positively correlated, but this correlation is only 0.4.) Fig. 3 displays the overall pattern of

missing data: every country is missing some data, and a total of 19% of the data will be imputed. In this take we ignored possible temporal dependence in the missingness structure. Observed values were observed at the most recent year that was available; missing values were completely missing all the way back in time. It should be acknowledged that the use of the last value is a crude approach to imputation; in retrospect, a more sophisticated procedure could have been used. Constructing the ESI by using only available cases would have severely restricted its scope; yet it was important to have a reasonability check for the imputations. As such, we sought an automatic method to screen the imputations and to identify potential problems. This motivated the suite of tools that are developed in this paper.

2. Methods

2.1. Multiple imputation using sequential regressions

We begin with a data set—a data matrix with missing values—and suppose that the user has already decided on a multiple-imputation procedure, has fitted it to the data and has constructed a set of imputations. We then have several imputed *completed data sets*. Our diagnostics can be applied independently to each completed data set. 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. Strictly speaking, our approach is applicable to any random predictive imputation model. Most likely—in practice—these methods could be used in multiple imputation where many completed data sets are generated (see Rubin (1996)) and then diagnosed. Although any one imputation will yield multitudes—problems that should be illustrated by imputation diagnostics—a further research question is how to examine multiple imputations without being overwhelmed by the graphical displays.

We shall assume that 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.* (1999), Van Buuren and Oudshoom (2000) and Raghunathan *et al.* (2001) defined imputations by 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 data set (see Gelman and Raghunathan (2001)). Sequential regression multiple imputation (SRMI) proceeds by partitioning and ordering the data set by number of missing items, and 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 to proceed iteratively until convergence in model parameters is achieved.

We used the software of Raghunathan *et al.* (2001), in the end imputing approximately 19% of the data for the ESI. (Of the 64 variables in the ESI, 24 were not included in the imputation process at all, for reasons that were entirely based on the ESI context and having nothing to do with the statistical analysis. We are using the method that is described in this paper to evaluate the imputations for the remaining 40 variables.) 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 that is used to generate the imputations or can indicate a more serious violation of the missingness assumptions. In both cases the flagging compares the imputed values with the observed values. In the sense that the completed data set 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 shall judge the propriety of the imputed values by comparison with the observed values. 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 modelling 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 modellers usually have a notion about what this complete data set looks like and can use these notions to frame their flagging procedures. 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, to flag potential problems and to fix or refine our imputation model.

We emphasize that differences in distribution between the imputed and the observed data *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 data set 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 data by using the Kolmogorov–Smirnov (KS) test for each variable, raising the flag when we find statistically significant differences. (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 data.) 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 by using this assumption, then any differences in distributions are necessarily explainable by other variables in the data set. Nonetheless, as discussed in the hypothetical examples of Appendix A, 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.

Classical statistical significance provides a convenient cut-off rule that seems to work well in our example. More generally, a procedure for deciding which discrepancies to examine further 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% level of significance will be appropriate, but we present this rule here as a starting point, which is worthy of further examination. To the extent that we can examine the distributions visually, this is not necessarily a crucial issue in practice. However, 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 scatter plots

Bivariate scatter plots 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 allows 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.

Fig. 4 is an example of these type of comparisons—we plot the completed data against the ESI. The ESI includes *external* knowledge—data that are not included in the imputation procedure—and *internal* data. Each completed variable can be plotted against external or internal data separately, as well.

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

2.3.1. Residual plots

The SRMI software of Raghunathan *et al.* (2002) does not allow inspection of the imputation model—this is a disadvantage with respect to checking the validity of the second multiple-imputation assumption. We constructed a proxy for the iterated SRMI models, however, by selecting the best stepwise model for each variable in the completed data (Y_j) regressed on all others ($\mathbf{Y}_{-j}$). We generated predicted values ($\hat{y}_{ij}$)—i the ith observation; j the jth variable—for the 40 variables in the analysis, and we consider these analogues 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 and 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 data and then updating our guess for each imputation.

We fit a LOWESS curve (Cleveland, 1979) to each of the scatter plots of residual differences between an available stepwise model (one that 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 *versus* the residual differences between the observed and the predicted data. 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. See Appendix A.4.

We applied our method of residual refinement to a sample environmental data set (Johnson and Wichern, 1998) under complete (MCAR), random (MAR) and non-random (MNAR) missingness mechanisms. See Appendix A.

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,

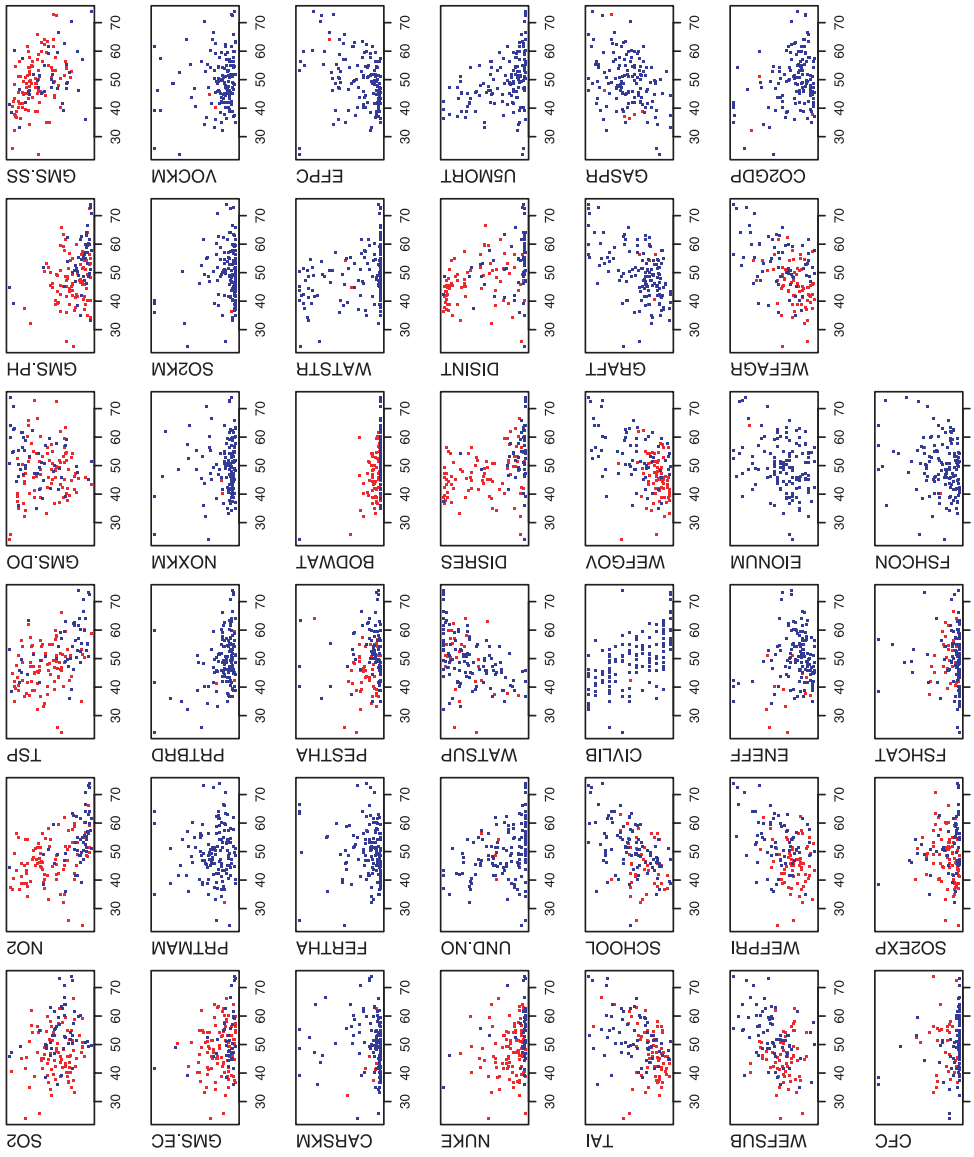

Fig. 4. For each variable, the observed and imputed values for the 142 countries are plotted against the ESI (imputed values, everywhere, are in red; observed values are in blue): at a glance there is evidence for non-random patterns of missingness in many variables, as discussed in detail in the text

when the assumption is false, differences between observed and imputed data are not correctable by the residual calibration.

It can be difficult to fix the imputation with the method proposed because fixing is done on the basis of 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 ESI. 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; Fig. 5. A second step is to display the observed and imputed data for all imputed variables, *versus* the overall index, as shown in Fig. 4. We discuss these plots in particular for a group of variables (a 'component') in the ESI. As practitioners, we would investigate all the data similarly.

The 2002 ESI data are available from `http://sedac.ciesin.columbia.edu/es/esi/ESI2002_dat.xls`.

3.1. A quick look at all the variables

There are plausible explanations for the differences in scatter plot patterns that we see when plotting the data from each variable against the composite ESI score in Fig. 4. Taking the environmental systems group as an example, we may expect that some countries with lower values, in gross domestic product for instance, will have higher emissions—a finding that does not contradict environmental theory. (An example is the BODWAT-variable; see Fig. 1.)

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 which are 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 differ 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 $p < 0.001$): nitrogen dioxide 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 in each ESI component grouping (Fig. 5):

(a) *environmental systems*—NO2 (flagged variable)—urban nitrogen dioxide concentration; SO2—urban sulphur dioxide concentration;
(b) *environmental stresses*—NUKE (flagged variable)—radioactive waste; WATSTR—percentage of the country's territory under severe water stress;
(c) *vulnerability*—DISRES (flagged variable)—child death rate from respiratory diseases; WATSUP—percentage of population with access to improved drinking water supply;
(d) *capacity*—SCHOOL (flagged variable)—mean years of schooling (age 15 years and above); GASPR—ratio of gasoline price to international average;
(e) *stewardship*—FSHCAT (flagged variable)—total marine fish catch; FSHCON—seafood consumption *per capita*.

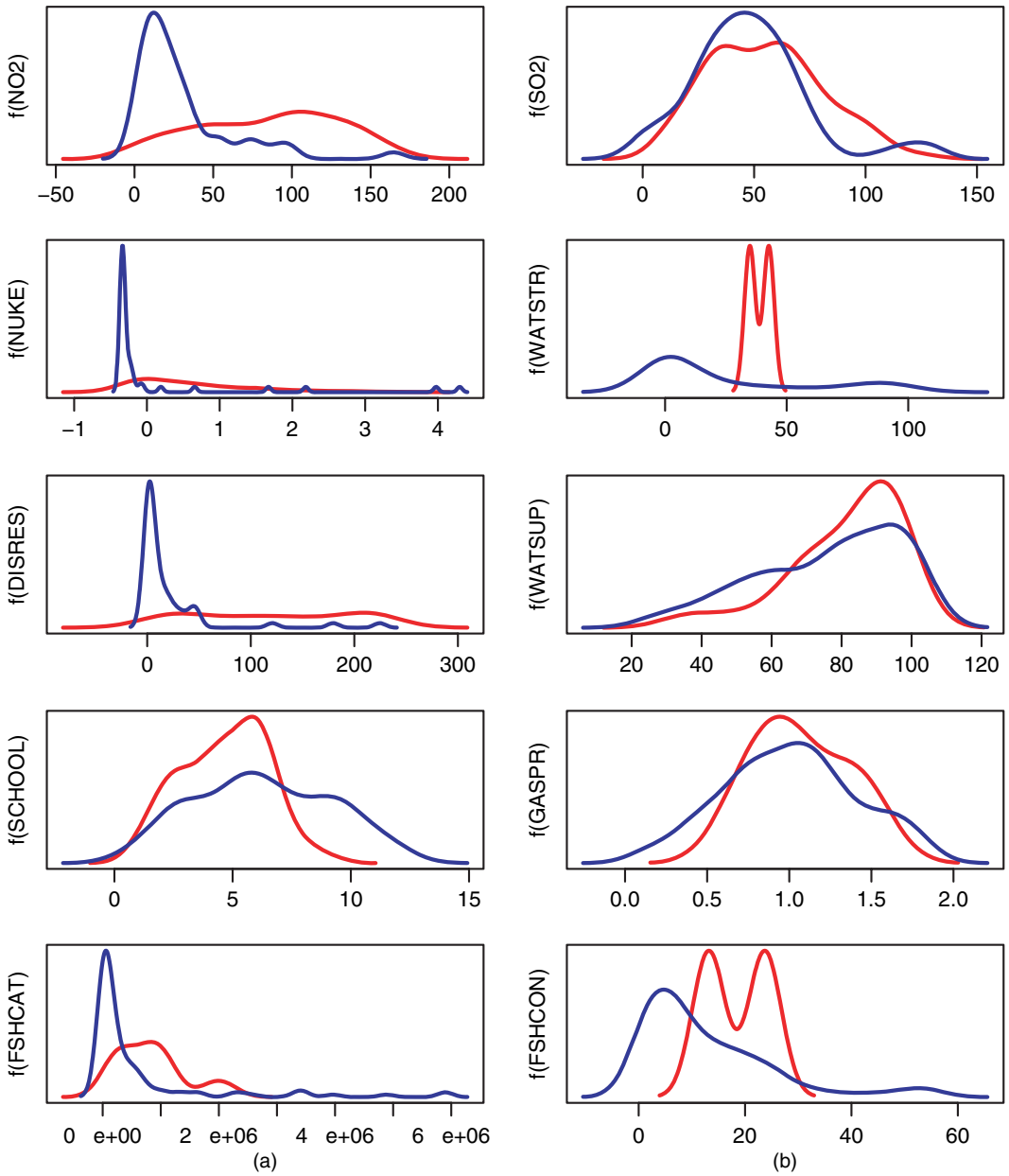

Fig. 5. (a) For each component of the ESI, a variable whose imputed values (red) differ significantly from observed values (blue) and (b) for comparison, a variable from each component which we did not flag: possible flaws in imputations may appear in the graphs even when not indicated by the KS tests; 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 are necessary

In comparison, for each grouping, we also chose one variable that did not significantly differ.

Fig. 4 provides a snapshot of the differences between the observed and imputed values for the entire data—in some cases the differences are striking. Differences in the distributions are either functions of differences in the predictors—or functions of the latent missingness mechanism. In the latter case, we may expect that some countries misreport or restrict—intentionally or not—data (e.g. air and water particulate concentrations). In the former case, 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 demand 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. Fig. 6 is an example of the sort of requisite post-imputation diagnostic plots that 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 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 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 data. The adjustment affects the univariate histogram of SO2 more dramatically than that of NO2: the distribution of the imputed values matches the observed values 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. In this fashion, we reduce the problem of checking the full multivariate imputation or completed data statistic to a series of decisions which may be influenced by the practitioner's knowledge of each variable.

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 is broadly defined.

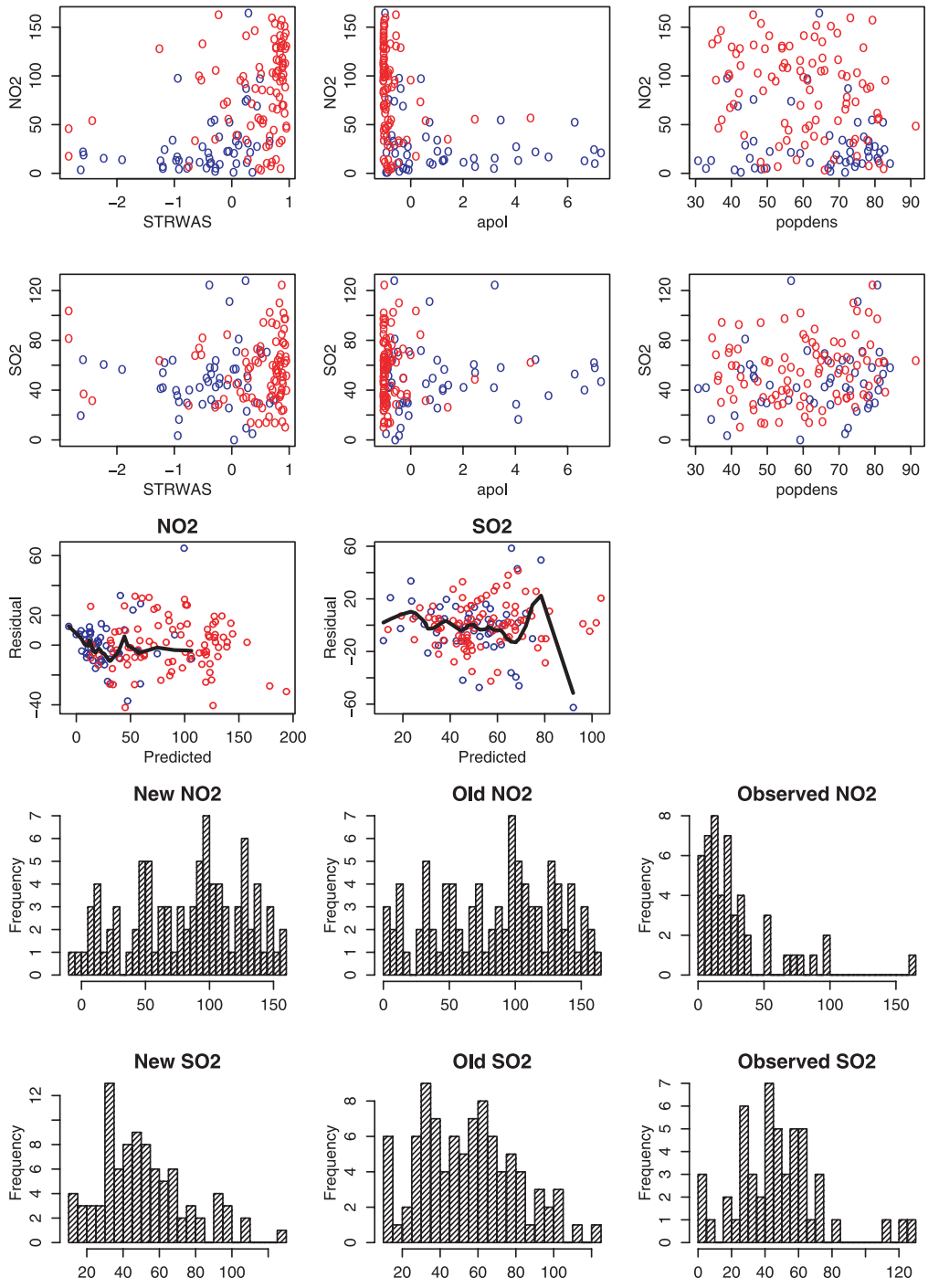

Fig. 6. Environmental systems: NO2 is flagged as significantly different by the KS test; the bivariate scatter plots highlight distributional differences; STRWAS is a composite of air quality measurements that is used in the ESI; popdens is a measure of population density; the residual plots show the 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

We already know that countries with more missingness have performed worse on the observable measurements; we do not know whether 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 that 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 that is 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 by 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 KS 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 addition to a standard multiple-imputation package such as MICE (Van Buuren and Oudshoom, 2000). (Think of a graph array—Fig. 6—for each of the components, as a complementary, necessary diagnostic output to a completed data set for any imputation software.) 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 metaparameterizations 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 on our imputation model by revisiting the observed data and exploiting the difference in patterns of the observed and missing data with respect to the imputation model.

Acknowledgements

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 ESI is the result of collaboration between 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.

Appendix A

A.1. Computation of the environmental sustainability index

The 64 variables in the ESI (World Economic Forum, 2002) are grouped into 21 subsets called 'indicators'—the average of these indicators is the final ESI. In notation the ESI is

$$\mathrm{ESI} = 100\,\Phi\left(\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\frac{1}{|J_k|}\sum_{j\in J_k}\frac{Y_j-\bar{Y}_j}{S_{y_j}}\right).$$

Here J_k is the index set for the variables in the kth indicator of the ESI: the ESI is averaged over the indicators; the ESI 'components' are a heuristic grouping, which is not used in calculating the index (see World Economic Forum (2002)); $\mathcal{K}$ is the index set for the indicators; $|\mathcal{K}|$ and $|J_k|$ are the number of indicators and number of variables in the kth indicator; $\bar{Y}_j$ is the sample mean for variable j—across

countries; S_{y_j} is the sample standard deviation for variable j; Φ is the inverse standard normal distribution function. See the list at the end of Section 1.2 and Table 3 in World Economic Forum (2002).

A.2. Sequential regression multiple-imputation procedure

Commonly, $G(\mathbf{Y}, \theta)$—the complete-data model—is supposed multivariate joint normal, and the missing data are imputed as draws from the joint posterior (as in Markov chain Monte Carlo imputation). Van Buuren *et al.* (1999) and Raghunathan *et al.* (2001) investigated that a G can be replaced with a product of conditional distributions $G = \Pi_{t \in \mathcal{T}} G_t$, where $\mathcal{T}$ is some index set on the conditionally independent distributions. SRMI proceeds by partitioning the data set $\mathbf{Y} = (Y_1, \ldots, Y_T)$ in order of missingness, T is the number of variables; at step $1 \leqslant r \leqslant T$, $\mathbf{X}$ is the set of $T - r$ complete(d) predictors and $\mathbf{Y}^* = (Y_{T-r+1}, \ldots, Y_T)$. $\mathbf{Y}^*$ is regressed, iteratively, on $\mathbf{X}$. The steps, in this application, are as follows.

(a) The first round of the SRMI algorithm begins by regressing Y_1, the variable with the least missing items, on $\mathbf{X}$—the complete data. The missing values in Y_1 are imputed randomly from an approximate predictive distribution that is based on the fitted regression.
(b) Now Y_1 is entered into $\mathbf{X}$—the complete(d) data—and the algorithm regresses Y_2 on $(\mathbf{X}, Y_1)$. The algorithm continues until Y_T is completed by regressing it on $(\mathbf{X}, Y_{T-1})$.
(c) The subsequent rounds continue similarly, except $\mathbf{X} = \mathbf{Y}_{-r}$—the predictors for each Y_j are the remaining data.
(d) The algorithm cycles through the above step until the imputed values converge.

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 that were far wider than the observed distributions. At each variable, this may or may not be a problem: if the missingness mechanism is, perhaps, missingness not completely at random, the 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.3. Fixing imputations—refinement procedure

Let $\hat{G}$ be an estimate of G; $\hat{G}$ is the procedure that is used to generate random imputations which, when combined with the observed data, yield a complete data set. (In the example in this paper, $\hat{G}$ is set as the best stepwise regression of Y_j on $\mathbf{Y}_{-j}^{(k)}$. More generally $\hat{G}$ could be obtained from a multivariate model (e.g. Schafer (1997)), sequential regressions (e.g. Raghunathan *et al.* (2001)) or other methods.) Let $\hat{y}_j = \hat{G}(\mathbf{Y}_{-j})$ be the predicted values from the imputation model for each (vector) variable, where $\mathbf{Y}_{-j}$ are the data excluding the jth variable. $r_j(y_j) = y_j - \hat{G}(\mathbf{Y}_{-j})$ is the residual function, which in practice is estimated by binning or otherwise smoothing the residuals from the fitted model. In this paper we smoothed by using a LOWESS curve.

We correct (or calibrate) the imputed values by subtracting the estimated $\hat{r}_j$ from the data vector y_j.

A.4. Simulation study

Beginning with an example set of air quality data (Johnson and Wichern, 1998) we investigated the behaviour of our imputation refinement procedure under three simulated missingness mechanisms: MCAR, MAR and MNAR. Let z_{ij} be 1 if observation y_{ij} is missing and 0 otherwise, distributed as follows under each assumption: for MCAR—$\Pr(z_{ij} = 1) = p_j$; for MAR—$\Pr(z_{ij} = 1) = \text{logit}^{-1}\{a_{1j} + (\hat{y}_{ij} - b_1)/c_1\}$; for NMAR—$\Pr(z_{ij} = 1) = \text{logit}^{-1}\{a_{2j} + (y_{ij} - b_2)/c_2\}$.

We set the p_j, a_1 and a_2 to decrease with j to generate a pattern of monotone missingness under each of the assumptions. Constants b_1, b_2, c_1 and c_2 exist so 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 assumption of MNAR—and more so for the imputations on the assumptions of MCAR (Figs 7 and 8).

The data for the simulations are available from `http://esminfo.prenhall.com/math/johnsonwichern/data/Wichern_data.zip`.

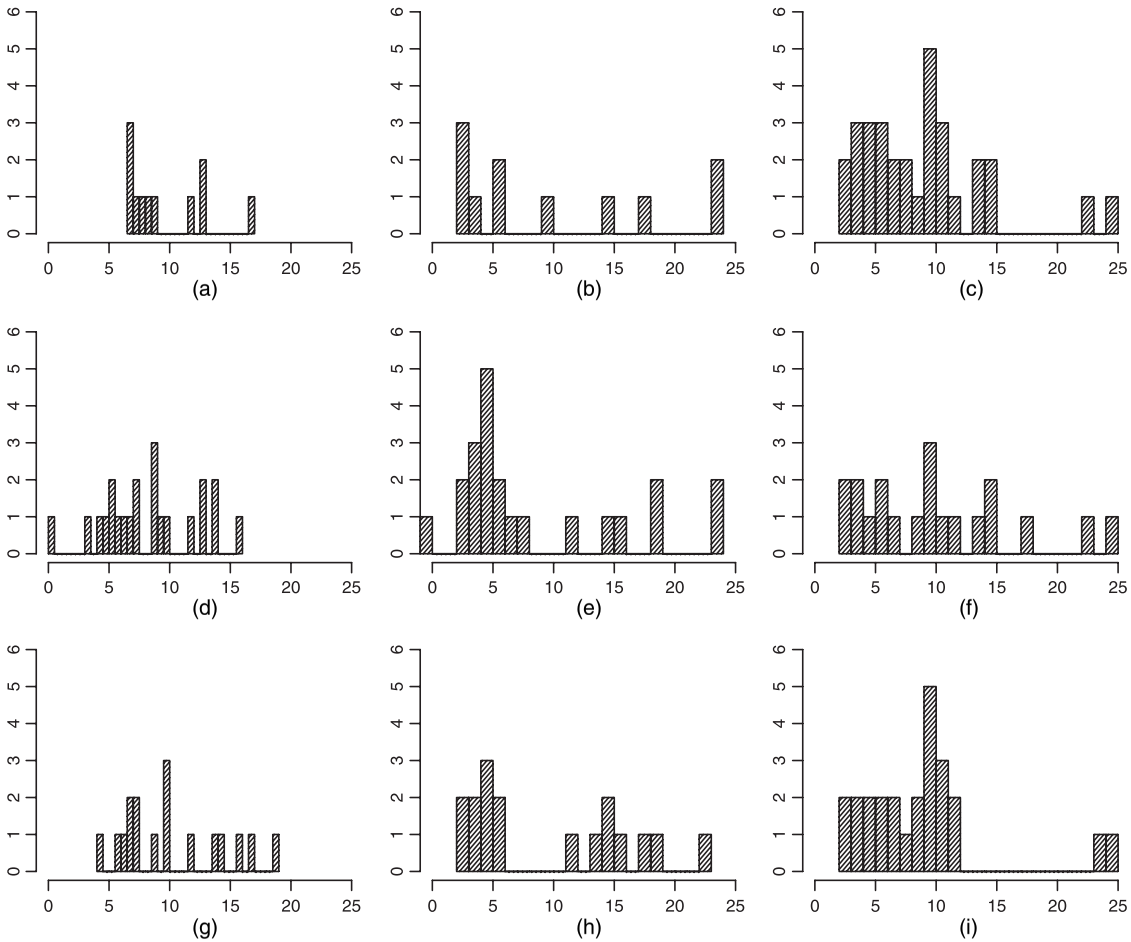

Fig. 7. Simulated imputation refinement on the air quality (ozone) data (the refinements more closely mimic the distribution of the observed values under the MCAR and MAR mechanisms; under MNAR the refinements perform less well—the imputed distribution has a wider range than the observed distribution): (a) MCAR, before calibration; (b) MCAR, after calibration; (c) MCAR, observed; (d) MAR, before calibration; (e) MAR, after calibration; (f) MAR, observed; (g) MNAR, before calibration; (h) MNAR, after calibration; (i) MNAR, observed

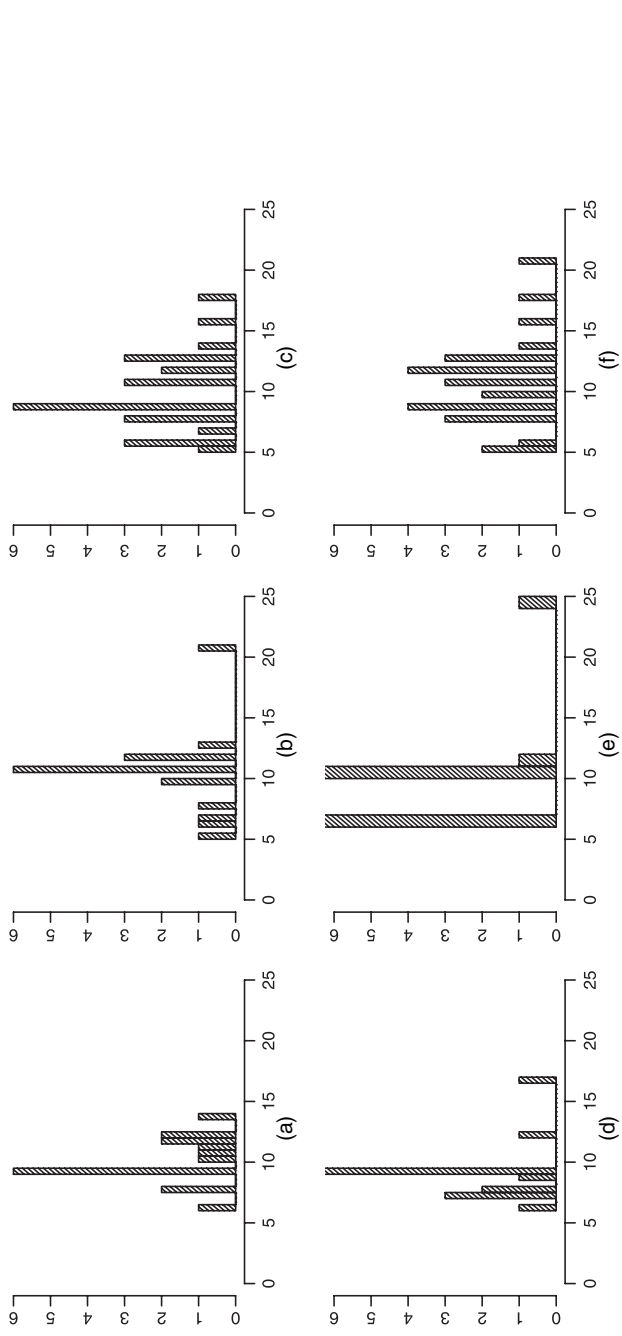

Fig. 8. Simulated imputation refinement on the air quality (nitrogen dioxide) data (the refinements match the distribution of the observed data better than the original imputations under MCAR; the range of refinements is greater than that of the observed data under MAR; under MNAR the original imputations more closely match the observed data): (a) MCAR, before recalibration; (b) MCAR, after calibration; (c) MCAR, observed; (d) MAR, before recalibration; (e) MAR, after recalibration; (f) MAR, observed; (g) MNAR, before recalibration; (h) MNAR, after recalibration; (i) MNAR, observed

References

Cleveland, W. S. (1979) Locally weighted regression and smoothing scatterplots. *J. Am. Statist. Ass.*, **74**, 829–836.

Gelman, A. and Raghunathan, T. E. (2001) Using conditional distributions for missing-data imputation: discussion of "Conditionally specified distributions" by Arnold et al. *Statist. Sci.*, **3**, 268–269.

Heckman, J. (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann. Econ. Socl Measmnt*, **5**, 475–492.

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

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, 2nd edn. New York: Wiley.

Liu, C. (1995) Missing data imputation using the multivariate *t* distribution. *J. Multiv. Anal.*, **48**, 198–206.

Prescott-Allen, R. (2001) *The Wellbeing of Nations*. Washington DC: Island.

Raghunathan, T. E., Solenberger, P. W. and Van Hoewyk, J. (2002) IVEware. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor. (Available from `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. *Surv. Methodol.*, **27**, 85–95.

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

Rubin, D. B. (1978) Multiple imputation in sample surveys—a phenomenological Bayesian approach to nonresponse. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 20–37.

Rubin, D. B. (1996) Multiple imputation after 18+ years (with discussion). *J. Am. Statist. Ass.*, **91**, 473–520.

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. *Statist. Sin.*, **14**, 1221–1237.

United Nations Development Program (2001) *Human Development Report*, Table A2.1. New York: Oxford University Press.

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

Van Buuren, S. and Oudshoom, C. G. M. (2000) MICE: multivariate imputation by chained equations. Netherlands Organization for Applied Scientific Research, Delft. (Available from `web.inter.nl.net/users/S.van.Buuren/mi/`.)

World Economic Forum (2002) *Environmental Sustainability Index*. New York: Global Leaders for Tomorrow Environment Task Force, World Economic Forum and Yale Center for Environmental Law and Policy and Center for International Earth Science Information Network. (Available from `sedac.ciesin.columbia.edu/es/esi/archive.html`.)