*B.J.Pol.S.* **43**, 425–449 *Copyright © Cambridge University Press, 2012*
doi:10.1017/S0007123412000312
First published online 19 July 2012

# *We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data*

SKYLER J. CRANMER AND JEFF GILL*

Missing values are a frequent problem in empirical political science research. Surprisingly, the match between the measurement of the missing values and the correcting algorithms applied is seldom studied. While multiple imputation is a vast improvement over the deletion of cases with missing values, it is often unsuitable for imputing highly non-granular discrete data. We develop a simple technique for imputing missing values in such situations, which is a variant of hot deck imputation, drawing from the conditional distribution of the variable with missing values to preserve the discrete measure of the variable. This method is tested against existing techniques using Monte Carlo analysis and then applied to real data on democratization and modernization theory. Software for our imputation technique is provided in a free, easy-to-use package for the R statistical environment.

Missing values are unfortunately common in political science data. These missing values occur for a variety of reasons in all sub-fields of the discipline: respondents refuse to answer some questions in surveys or experiments, data is not available that captures particular characteristics of some countries, states or other observational units, and so forth. Although many datasets used by political scientists have some amount of missing data,[1] only in the last decade have political science researchers become seriously concerned with the deleterious effects of so-called 'item non-response'. Except for rare cases, the common technique of deleting from the dataset all cases with any missing elements biases subsequent statistical analysis. Even state-of-the-art techniques for filling in, or *imputing*, missing data can fail when a discrete variable with a small number of categories has missing values. In this case, common imputation techniques, which rely on continuous distributional assumptions, can produce nonsensical imputations, biased results and artificially smaller standard errors. The result can be entirely different *substantive* conclusions. While this is an important issue in all empirical sub-fields of political science, it appears to occur more often in comparative politics, and in any area using survey research.

Missing data also harms research in political science in another way that mostly affects the study of comparative politics. The literature is replete with examples of authors who limit the scope of their analysis because data from a specific year or a given country are

* Department of Political Science, University of North Carolina; and Department of Political Science, Washington University (email: jgill@wustl.edu), respectively. The authors wish to thank Micah Altman, James Fowler, Katie Gan, Adam Glynn, Justin Grimmer, Dominik Hangartner, Michael Kellerman, Gary King, Ryan Moore and Randolph Siverson for valuable comments. Replication data is available at http://www.unc.edu/~skylerc/.

[1] The term 'missing data' can mean either missing values (e.g. item non-response in a survey) or missing observations such as refusal to take an entire survey. Throughout this work, we use the term exclusively to mean the first case.

not available. Representative examples are: Taagepera and Shugart's seminal work on votes and seats in Europe, which eliminates Denmark 1913–32 and Germany 1920–33,[2] Mair and van Biezen's study of membership decline in European parties, which drops a large number of smaller parties,[3] Palmer and Whitten's comparative evaluation of economic voting, which drops specific elections,[4] and Reiter's search for evidence that peace encourages the spread of democracy, which restricts the years of the analysis from 1960 to 1992.[5] This problem forces the analyst to restrict the scope of their research question and may change the impact of their results. The fault is often attributed to poor governmental record-keeping and, therefore, the problem affects the study of particular parts of the world in greater measure. So, this is essentially a different form of observation deletion, whereby the wholesale removal of components of the data causes restrictions to the type of analysis that can be conducted.

To help address these problems we present a non-parametric alternative to multiple imputation, which we call *multiple hot deck imputation*. This tool is designed specifically to work well in situations where (traditional) parametric multiple imputation falls short. Multiple hot deck imputation is a simple concept that maintains the integrity of the data by using draws of *actual* values from the variable with the missing values to impute missing items. Therefore, this technique maintains the discrete nature of discrete data and produces more accurate imputations than parametric multiple imputation in a majority of social science applications where the data are discrete. Since many political science applications rely on highly discrete measures, multiple hot deck imputation will typically provide the researcher with more accuracy in imputations than parametric multiple imputation, while requiring none of parametric multiple imputation's standard assumptions. We have also developed free and easy-to-use companion software to facilitate the application of multiple hot deck imputation. This is implemented in an R package that is discussed in Appendix B.

## THE MISSING DATA PROBLEM

We begin by describing the missing data problem and the shortcomings of parametric multiple imputation with discrete data in detail.[6]

### *Motivating Example*

Two primary dangers exist in the analysis of incomplete data: ignoring missing values with casewise deletion, and applying the wrong method for imputation. Both of these lead

[2] Rein Taagepera and Matthew Soberg Shugart, *Seats and Votes: The Effects and Determinants of Electoral Systems* (New Haven, Conn.: Yale University Press, 1989).

[3] Peter Mair and Ingrid van Biezen, 'Party Membership in Twenty European Democracies, 1980–2000', *Party Politics*, 7 (2001), 5–21.

[4] Harvey D. Palmer and Guy D. Whitten, 'The Electoral Impact of Unexpected Inflation and Economic Growth', *British Journal of Political Science*, 29 (1999), 623–39.

[5] Dan Reiter, 'Does Peace Nurture Democracy?' *Journal of Politics*, 63 (2001), 935–48.

[6] Recent book-length works on these topics include: Anastasios A. Tsiatis, *Semiparametric Theory and Missing Data* (New York: Springer, 2010); Craig K. Enders, *Applied Missing Data Analysis* (New York: The Guilford Press, 2010); Ming T. Tan, Guo-Liang Tian and Kai Wang Ng, *Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation* (New York: Chapman & Hall/CRC, 2009); Geert Molenberghs and Michael G. Kenward, *Missing Data in Clinical Studies* (New York: Wiley, 2007); Patrick E. McKnight, Katherine M. McKnight, Souraya Sidani and Aurelio Jose Figueredo, *Missing Data: A Gentle Approach* (New York: The Guilford Press, 2007).

to biased coefficient estimates and misleading results. A study of the Special Migration Statistics from the 1991 British Census by Rees and Duke-Williams illustrates both of these dangers clearly.[7] The authors are concerned that the deliberate embargo of some values in the data to protect the privacy of respondents harms the accurate estimate of patterns of migration flows. These researchers were able to fill in much, but not all, of the missing data with a method called *logical data patching*. This approach uses observed structure in the data to *deterministically* fill in missing values. As a simple example, suppose that the response to a survey question asking whether or not the respondent was college educated is missing, but the respondent answered 'solicitor' to the occupation question. The analyst can then with high confidence *patch in* 'yes' for the missing value. Rees and Duke-Williams focus on migration between local government districts within Great Britain, and are missing the destinations for some individual cases.[8] Yet they were able to deterministically obtain many of these values by using sums of migrations in the margins of tables, since columns and rows must sum to the marginals. The rest of the data was filled in with a technique that lowers the level of measurement for variables with missing values and, therefore, will not appeal broadly to empirical political scientists. After these methods are applied, they find that ignoring the plentiful missing information about the economic position of the family head would have seriously biased coefficient estimates in subsequent model fits.

Logical data patching is an effective means of filling missing data since it is done deterministically andconsequently introduces no uncertainty in the form of variance. The two problems with this approach are that it consumes experts' time, and that it usually cannot fill in all the missing data. In the case of the migration study, the economic position of the family head is a nominal-measured variable (Self-employed, Other employed, Unemployed, Retired, Students (inactive), Other inactive, Students (active)), and most methods of imputing such missing data assume interval or near-interval measured data. Even highly sophisticated implementations of missing data do not handle such cases well, including elaborate dummy-variable set-ups and approximation methods. Thus, there is a need for some method focused specifically on missing categorical data.

### *Bias from Missing Values*

Until recently, the most common method of addressing missing data in empirical models has been to delete every case with even a single missing value: a technique called *casewise deletion* (also called list-wise deletion or – optimistically – complete-case analysis). With casewise deletion, rows of the data matrix corresponding to individual cases are removed from the dataset, which reduces the effective sample size and lowers the degrees of freedom for any subsequent modelling exercise. For example, if some survey respondents did not answer all of the items on a questionnaire, then those respondents would be effectively deleted, leaving a dataset of only respondents who answered every single question. Unfortunately, this approach is the default with many statistical packages, and there is typically not even a warning message. The obvious *convenience* of this approach, however, is that the subsequent process of statistical analysis is then unhampered by the missing values.

[7] Phil H. Rees and Oliver Duke-Williams, 'Methods for Estimating Missing Data on Migrants in the 1991 British Census', *International Journal of Population Geography*, 3 (1997), 323–68.

[8] Rees and Duke-Williams, 'Methods for Estimating Missing Data on Migrants in the 1991 British Census'.

Generally, casewise deletion induces bias with political data and cannot, therefore, be recommended when analysing data with missing values. The extent of the bias induced by casewise deletion depends on what proportion of the data are missing, how far the pattern of missing values is from being completely random throughout the data matrix, and how different the missing data are from the observed data.[9] In the special circumstance where missing values occur *completely* randomly, casewise deletion is unbiased in expectation regardless of what proportion of the data is missing, although small sample issues may emerge in columns of the data matrix. This bias from casewise deletion in real applications ranges from mild to catastrophic.[10] Replications of published work in political science have shown that the bias induced by casewise deletion can be severe enough to produce faulty inferences.[11] It is difficult to imagine a situation in which casewise deletion is an ideal treatment of missing values, because it is unbiased only for a narrow definition of data, and always produces greater inefficiency since some fraction of the data are removed. Therefore, a consensus has developed in the statistical, econometric and methodological literatures that missing values should be filled in (imputed) using some principled method if possible.

There are three basic mechanisms by which data can be missing:[12] it can be missing completely at random (MCAR), missing at random (MAR) or non-ignorable (NI). Each missing- data generating mechanism is described by a simplification of $p(\mathbf{M}|\mathbf{D})$, where $\mathbf{D}$ is the data matrix (including all explanatory variables as well as the outcome variable) and $\mathbf{M}$ is a dichotomous indicator matrix of the same dimensions as $\mathbf{D}$, with values of 1 if a datum is missing and values of 0 if it is present.[13] Unfortunately, it is not possible to test directly for any of these mechanisms; one of them must be assumed. So researchers with this problem need to pay attention to the *substantive nature* of how the data became missing. For example, it is well-known that survey respondents are less likely to answer the personal income question if their income is very high or very low. This not only means that casewise deletion gives an inaccurate view of the subjects' income, it also means that characteristics of interest related to these two groups will be censored from the data as well.

The best-case scenario for casewise deletion is when observations are missing completely at random (MCAR). MCAR means that there is no underlying associative process that causes the absence of data: the occurrence of missing values is uniformly random throughout the data matrix. When the missing values are orthogonal (mathematically unrelated) to any values of the data, missing or observed, the missing data are MCAR: $p(\mathbf{M}|\mathbf{D}) = p(\mathbf{M})$.

[9] Roderick J. A. Little and Donald B. Rubin, *Statistical Analysis with Missing Data*, 2nd edn (New York: Wiley, 2002), p. 42.

[10] Paul D. Allison, *Missing Data* (Thousand Oaks, Calif.: Sage, 2001); Little and Rubin, *Statistical Analysis with Missing Data*; Roderick J. A. Little, 'Regression with Missing X's: A Review', *Journal of the American Statistical Association*, 87 (1992), 1227–37; Roderick J. A. Little, 'Approximately Calibrated Small Sample Inference about Means from Bivariate Normal Data with Missing Values', *Computational Statistics & Data Analysis*, 7 (1988), 161–78; Donald B. Rubin, 'Inference and Missing Data (with Discussion)', *Biometrika*, 63 (1976), 581–92; Gary King, James Honaker, Anne Joseph and Kenneth Scheve, 'Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation', *American Political Science Review*, 95 (2001), 49–69.

[11] King, Honaker, Joseph and Scheve, 'Analyzing Incomplete Political Science Data'; James Honaker and Gary King, 'What to Do about Missing Values in Time-Series Cross-Section Data', *American Journal of Political Science*, 54 (2010), 561–81.

[12] Rubin, 'Inference and Missing Data'; King, Honaker, Joseph and Scheve, 'Analyzing Incomplete Political Science Data'; Little and Rubin, *Statistical Analysis with Missing Data*.

[13] Little and Rubin, *Statistical Analysis with Missing Data*, p. 12.

TABLE 1 *Assumptions about Missing Data*

| Assumption | Definition |
|---|---|
| MCAR | $p(\mathbf{M}\|\mathbf{D})=p(\mathbf{M})$ |
| MAR | $p(\mathbf{M}\|\mathbf{D})=p(\mathbf{M}\|\mathbf{D}_R)$ |
| NI | $p(\mathbf{M}\|\mathbf{D})=p(\mathbf{M}\|\mathbf{D})$ |

A simple way to think of MCAR missing values is that both the observed and missing data, independently, have the properties of a random sample of the population. More specifically, the data for which we observe responses, $\mathbf{D}_R$, and the data for which we do not have responses, $\mathbf{D}_{NR}$, have the same underlying distribution. Though unbiased, casewise deletion in this circumstance may not necessarily be ideal since information is lost by discarding incomplete observations and, even if no bias ensues, precision is lost and standard errors from regression-style models will be artificially large.

Unfortunately, data where missing values are completely random are not particularly common in political science: survey respondents are more sensitive to some questions than others, certain types of regimes are more likely to refuse to provide economic or political data, some information is more difficult to recall, and so on. This is why the term 'independent variables' is a poor description for the social sciences. It is difficult to think of a situation in political science, other than a computer malfunction, that would result in missing values being entirely unrelated to *any* attribute or political phenomena, observed or unobserved.

More commonly, missing data is assumed to be *missing at random* (MAR). If missing values are MAR, then their occurrence is related to the observed data $\mathbf{D}_R$, but not to any unobserved data $\mathbf{D}_{NR}$. The MAR assumption is expressed formally as $p(\mathbf{M}|\mathbf{D}) = p(\mathbf{M}|\mathbf{D}_R)$. MAR is a fairly intuitive concept: it simply means that missing values in one variable, say income, can be related to other variables (education, occupation, neighbourhood and so on) but those other variables must be recorded in the dataset.

When missing values are MAR, casewise deletion will *always* result in bias. This is because the missing data, $\mathbf{D}_{NR}$, are not a random subsample of all the data $\mathbf{D}$. The bias induced by casewise deletion under MAR is similar to sampling bias in survey research: the sample of values that are recorded, $\mathbf{D}_R$, is biased away from subjects with characteristics that make them less likely to respond and, consequentially, the recorded data $\mathbf{D}_R$ are not representative of the full sample, $\mathbf{D}$. Multiple imputation techniques, conversely, are ideal for dealing with continuously measured MAR data since they produce efficient and unbiased estimators with relatively little effort.

Finally, missing data can be *non-ignorable* (NI). Non-ignorable missing data occur when missing values are related to unknown and/or unobserved parameters. Consider again the example of missing values in an income question on a survey, but now we do *not* observe the variables of education, occupation and neighbourhood that can help fill in the missing income data. In such a situation, the missing data are NI. Formally, when NI holds, the expression for the missing data mechanism, $p(\mathbf{M}|\mathbf{D})$, cannot be simplified and specific information is fully missing from the dataset. With NI missing data, both casewise deletion and imputation tools produce biased statistical results since the missing data is purposefully different from the observed data. Non-ignorable missing data also biases subsequent modelling with multiple imputation since there is no observed information from which to build an imputation process for filling in missing values.

https://doi.org/10.1017/S0007123412000312 Published online by Cambridge University Press

One difficulty with non-ignorable missing values is that the condition of non-ignorability is, by definition, impossible to test for and analysts will never know with certainty whether missing values in their dataset are non-ignorable.[14] Whether missing values are treated as MCAR, MAR, or NI comes strictly from an *assumption* researchers make based on their understanding of the data generation and data missing process. The required assumption about the structure of missing values has fundamental implications for the modelling process: assuming MCAR or MAR allows the researcher to handle the missing data in a principled manner, but, if this assumption is incorrect, then subsequent inferences are likely to be misguided. However, recognizing NI missing data means that no standard statistical method exists for repairing the problem. An obvious course of action is to gather further data in an attempt to create conditional information such that the missing values are now MAR, but this is often not possible. If the researcher has a well-developed understanding of the response mechanism, Bayesian model-based imputation through a Markov chain Monte Carlo process may be helpful because assumptions are made about relationships in the data and information may be provided through the prior distribution.[15]

The three patterns of missing data can be further illustrated by looking at the development of a likelihood function for the data and an arbitrary parameter vector $\theta$ to be estimated. Segment the data matrix, $\mathbf{X}$, by rows into two constituent parts: $\mathbf{X} = [\mathbf{X}_R, \mathbf{X}_{NR}]$ for responses and non-responses, and write the likelihood:

$$L(\boldsymbol{\theta}|\mathbf{X}) = f(\mathbf{X}_R, \mathbf{X}_{NR}|\boldsymbol{\theta}) = f(\mathbf{X}_R|\boldsymbol{\theta})f(\mathbf{X}_{NR}|\mathbf{X}_R, \boldsymbol{\theta}) \tag{1}$$

The second equality comes from the standard definition of conditional probability. If we use the standard logarithmic transformation of the likelihood function, then the product above becomes a sum:

$$\ell(\boldsymbol{\theta}|\mathbf{X}) = \ell(\boldsymbol{\theta}|\mathbf{X}_R, \mathbf{X}_{NR}) = \log f(\mathbf{X}_R|\boldsymbol{\theta}) + \log f(\mathbf{X}_{NR}|\mathbf{X}_R, \boldsymbol{\theta}). \tag{2}$$

We first notice that casewise deletion is biased whenever $\log f(\mathbf{X}_{NR}|\mathbf{X}_R, \boldsymbol{\theta})$ is non-zero since then $\ell(\boldsymbol{\theta}|\mathbf{X}) \neq \ell(\boldsymbol{\theta}|\mathbf{X}_R)$. If the data are MCAR, then there is no conditional relationship between $\mathbf{X}_{NR}$ and $\mathbf{X}_R$ or other missing data. Therefore, the second term is not material and casewise deletion is acceptable. However, if there is a conditional relationship between missing data and observed data (MAR) then $\log f(\mathbf{X}_{NR}|\mathbf{X}_R, \boldsymbol{\theta})$ is clearly non-zero, but a tool that informs the likelihood using this conditionality can overcome the problem. This is what imputation tools do: impute missing values conditionally based on their relationship with observed values. In the case of NI data, the second term in Equation 1 is actually $\log f(\mathbf{X}_{NR}|\mathbf{X}_{ONR}, \mathbf{X}_R, \boldsymbol{\theta})$, where $\mathbf{X}_{ONR}$ is 'other non-responses'. In this case, the term here is non-zero and not estimable with the resources at hand. Therefore, assumptions about how to treat the missing data are actually assumptions about the relationship between $\mathbf{X}_{NR}$ and other data elements, observed or missing.

### *The Origins of Hot Deck Imputation*

Because of the many problems associated with casewise deletion, academic and professional survey researchers began looking for a better way to analyse incomplete data in the 1960s.

[14] King, Honaker, Joseph and Scheve, 'Analyzing Incomplete Political Science Data'.

[15] Andrew Gelman and Jennifer Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models* (New York: Cambridge University Press, 2007), pp. 540–1.

An early remedy developed roughly in the 1970s by market researchers and census takers was called 'hot decking' as a literal reference to taking draws from a deck of computer punch-cards.[16] In hot deck imputation, if a respondent has a missing value, a set of 'similar' respondents are pulled by hand from the data and one of them is randomly selected to provide a fill-in value for the missing datum. So there are two steps to the process: human selection of a sub-sample of cases, and uniform random selection from within this sub-sample. An obvious challenge is that it is not always clear what is meant by 'similar', although practitioners of the time became quite accomplished at performing this task over regular iterations. Hot deck imputation methods subsequently grew from simple random sampling to more complicated algorithms in attempts to find respondents as similar as possible to those with missing responses.

Hot deck imputation preserves the integrity of the data to the extent that only observed values of the variable being imputed are candidates for imputation. This, in turn, means that it is impossible for hot deck imputation to impute values outside the observed range of the data, a reasonable limit on how extremes of values can be imputed. Also, within the observed range of the data, only values that have occurred for other cases can be imputed, thus ensuring that nonsensical values cannot be imputed. The technique works best for discrete data because, while it is technically impossible that two respondents will have the exact same value of a continuous variable in the absence of rounding, it is quite likely that several respondents will have the same value for discrete variables (such as gender).

The major problem with traditional hot decking is that it does not reflect uncertainty in the imputed values: since only one value is imputed for each missing datum, all imputations are subsequently treated as factual responses rather than subjective probabilistic imputations. So there is no direct way to account for the uncertainty resulting from the process of sub-sample specification and then case selection from this sub-sample. Therefore, the subsequent model fit produces smaller standard errors and tighter confidence bounds than is appropriate.[17] In addition, the magnitude of this underestimation increases as the amount of missing data increases and as the quality of the hot deck matches decreases. We address this problem in the section ' Multiple Hot Deck Imputation'.

### *Parametric Multiple Imputation and Its Shortcomings*

A crucial improvement in the empirical analysis of incomplete data was precipitated by Rubin when he developed *multiple imputation*.[18] Rubin and his collaborators further refined the technique over the next two decades,[19] and it is now the dominant strategy for handling missing data in the social sciences and elsewhere for those not casewise-deleting.

[16] See, for instance, works by: John C. Bailar III and Barbara A. Bailar, 'Comparison of the Biases of the "Hot Deck" Imputation Procedure with an "Equal Weights" Imputation Procedure', *Symposium on Incomplete Data: Panel on Incomplete Data of the Committee on National Statistics, National Research Council*, 1997), 422–47; Brenda. G. Cox, 'The Weighted Sequential Hot Deck Imputation Procedure', *Proceedings of the Section on Survey Research Methods, American Statistical Association* (1980), 721–6; Richard C. Rockwell, 'An Investigation of Imputation and Differential Quality of Data in the 1970 Census', *Journal of the American Statistical Association*, 70 (1975), 39–42.

[17] Little and Rubin, *Statistical Analysis with Missing Data*; Donald B. Rubin, *Multiple Imputation for Nonresponse in Surveys* (New York: Wiley, 2004).

[18] Rubin, 'Inference and Missing Data'.

[19] Rubin, 'Inference and Missing Data'; Donald B. Rubin, 'Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys', *Journal of the American Statistical Association*, 72 (1977), 538–43; Donald B. Rubin, 'Multiple Imputations in Sample Surveys: A Phenomenological Bayesian

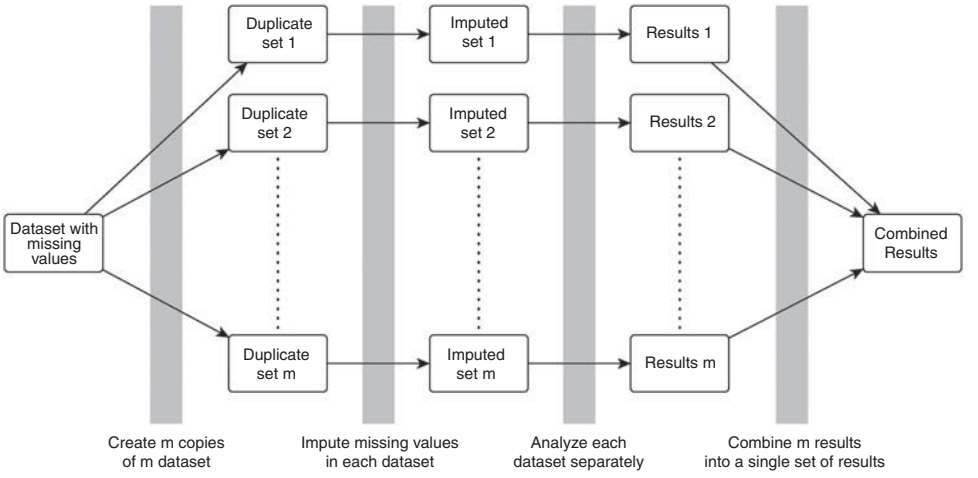


*Fig. 1. Multiple imputation*
*Note*: This figure illustrates the method of multiple imputation by which a dataset is copied several times, imputations are drawn for the missing values in each of the datasets, a statistic of interest is computed for each of the datasets, and, finally, those statistics are combined into a single result.

With multiple imputation each missing datum is replaced by several ($M \geq 2$) imputed values from a conditional distribution using other present values for that case. Thus, several imputed datasets are created with the appropriate missing datum being imputed by one of the $M$ imputed values to produce $M$ complete datasets. Then the complete-data statistic, $\theta$ (for example, a regression coefficient), is estimated $M$ times and the $\hat{\theta}_m$, $m = 1, \ldots M$ statistics are averaged to create a single estimate of $\bar{\theta}_M$.[20] A graphical overview of the process is shown in Figure 1.

While parametric multiple imputation remains the state-of-the-art technique for filling in missing observations, it can be problematic when used on discrete data. Multiple imputation, as commonly implemented, assumes a continuous metric for the missing data and, therefore, produces imputations that are continuous. The multiple imputation process then transforms categorical definitions into continuous values.[21] In this way a strictly binary variable for gender might be imputed with a value of 0.6 with parametric multiple imputation.

*(F'note continued)*
Approach to Nonresponse', *Proceedings of the Survey Research Methods Section of the American Statistical Association* (1978), 20–34; Donald B. Rubin and Nathaniel Schenker, 'Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse', *Journal of the American Statistical Association*, 81 (1986), 366–74; Donald B. Rubin, 'Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations', *Journal of Business and Economic Statistics*, 4 (1986), 87–94; Rubin, *Multiple Imputation for Nonresponse in Surveys*; Donald B. Rubin, J. L. Schafer and Nathaniel Schenker, 'Imputation Strategies for Missing Values in Post-Enumeration Surveys', *Survey Methodology*, 14 (1988), 209–21; Donald B. Rubin, 'Multiple Imputation after 18+ Years', *Journal of the American Statistical Association*, 91 (1996), 473–89.

[20] The combined $\bar{\theta}_M$ is in fact an average, but the treatment of the variability of this estimate is slightly more complicated than an average since it needs to account for within imputation variation and between imputation variation. The subject of multiple estimate combination will be discussed in some detail below. See Little and Rubin, *Statistical Analysis with Missing Data*, for a more detailed treatment.

[21] Jae Kwang Kim, 'Finite Sample Properties of Multiple Imputation Estimators', *Annals of Statistics*, 32 (2004), 766–83; Jae Kwang Kim and Wayne Fuller, 'Fractional Hot Deck Imputation', *Biometrika*,

Such problems intensify as the granularity of the discrete data declines. Alternatively, while binary response models will not fail if a nonsensical value is imputed into an explanatory variable, it may produce impossible or nonsensical results. A more basic problem is that such imputations fail to respect the meaning of the measure they are imputing; even if no bias is introduced, they degrade the substantive meaning of a variable of interest.

The most common solution to the categorical measurement problem with parametric multiple imputation has been to round continuous imputations to their nearest discrete value.[22] This approach allows for the use of multiple imputation and its attractive theoretical properties, while allowing the user to use complete data techniques appropriate for their discrete data. King and his coauthors have implemented this approach in their easy-to-use multiple imputation program called 'Amelia',[23] which has become a popular implementation of multiple imputation in political science.[24] Since models produced from parametric multiple imputation produce unbiased and consistent coefficient vectors, the acceptance of multiple imputation in the field has been a major improvement over previous approaches and an informal consensus has emerged holding that the process of rounding continuous imputations is not problematic.

Problems with the rounding approach, though subtle, are non-trivial. Chief among the pitfalls of this approach is that it produces biased results in nearly every case. Consider the imputation of missing values in a binary variable using parametric multiple imputation and rounding. If an imputation of 0.3 is made, it is rounded to 0. The imputation model however, did not predict that the value should be zero, it predicted that the value should be 0.3. Therefore, we *know* that parametric multiple imputation produces unbiased imputations for MAR data and we *still* deliberately bias the results by rounding with the entirely arbitrary cutpoint of 0.5. As we continue to draw imputations closer to zero than one and round them to zero, we are biasing our imputations further and further from what ought to be the centre of their distribution. The parametric multiple imputation with rounding technique will *only* be unbiased in the unlikely situation of imputations having a unimodal symmetric distribution centred *exactly* on 0.5. In all other cases, using parametric multiple imputation with rounding on discrete data will necessarily result in some degree of bias. Furthermore, coefficients from parametric multiple imputation and rounding on discrete data will have inappropriate standard errors because the distance between the imputation and the nearest integer is lost. Further still, in multinomial models when no latent variable may be assumed, an imputed and rounded continuous value can lead to substantially different results because of rounding error.

## MULTIPLE HOT DECK IMPUTATION

The intuition behind our multiple hot deck imputation technique is simple. The technique used is a variation of hot deck imputation combined with the repeated imputation and

*(F'note continued)*
91 (2004), 559–78; Wayne Fuller and Jae Kwang Kim, 'Hot Deck Imputation for the Response Model', *Statistics Canada*, 31 (2005), 139–49.

[22] King, Honaker, Joseph and Scheve, 'Analyzing Incomplete Political Science Data'; Honaker and King, 'What to Do about Missing Values in Time-Series Cross-Section Data'; Joseph L. Schafer, *Analysis of Incomplete Multivariate Data* (New York: Chapman & Hall/CRC, 1997).

[23] King, Honaker, Joseph and Scheve, 'Analyzing Incomplete Political Science Data'; Honaker and King, 'What to Do about Missing Values in Time-Series Cross-Section Data'.

[24] The articles describing the Amelia procedure have received over 330 ISI citations as of this writing.

https://doi.org/10.1017/S0007123412000312 Published online by Cambridge University Press

estimation method typical of parametric multiple imputation.[25] The advantages of this over previous approaches are substantial. First, imputing multiple datasets then running multiple analyses and combining coefficients and covariance matrices overcomes a major problem with traditional hot decking: that the uncertainty associated with imputation is not recorded. Secondly, by implementing a hot deck approach, the discrete properties of discrete data are maintained without rounding and its associated bias and shrinkage of standard errors – an improvement over previous multiple imputation techniques. Thirdly, multiple hot decking is a truly nonparametric approach to imputation. Traditional multiple imputation requires assumptions of normality (or $t$ approximations to normality for small samples) and is based on parametric models of the missing data; multiple hot decking avoids these assumptions. Fourthly, while multiple hot decking will work for continuous data, it works best where multiple imputation works worst: with discrete data. Fifthly and lastly, the assumptions and algorithm of multiple hot deck imputation are intuitive and easy to understand.

In hot deck imputation, when there are missing values for the $i$th case, observed values in the entire dataset $(\mathbf{y}, \mathbf{x})$ are used to fill-in the missing values. We call the respondent (or observation) providing the value the *donor* and the respondent (or observation) with the missing value, who will receive the donor's value, the *recipient*. Versions of this approach have been shown to be unbiased and efficient.[26] One advantage of this class of methods over parametric multiple imputation is that the imputed values are draws from the actual observed data, rather than draws from some constructed distribution. Therefore, when discrete variables are imputed with a hot deck method, their discrete properties are maintained.

What distinguishes multiple hot decking from other forms of hot decking is that, in multiple hot deck imputation, several values are used for a single missing observation.[27] This means that the values of several donors will be used for a single recipient. This *multiple* aspect of multiple hot deck imputation is what allows it to take imputation variance into account (an area where traditional hot decking fails). Multiple hot deck imputation harnesses a strength of hot deck imputation in that imputations are drawn from the conditional distribution of the actual data, and a strength of multiple imputation: accounting for imputation variance. It is even possible to derive the asymptotic distribution and finite-imputation variance of the hot deck estimator.[28]

The steps of the multiple hot deck imputation algorithm are quite simple. The first step of the algorithm creates $M \geq 2$ copies of the dataset $\mathbf{D}$, each with both missing and observed values. These are denoted $\mathbf{D}_m$ for $m = 1, \ldots, M$. The second step of the algorithm searches down each column of the dataset sequentially looking for missing values. When a missing value is found, a vector of *affinity scores* is computed, which measures how close the other cases/rows are to the one with missing data. This vector will be as long as the number of rows in the dataset less any rows with missing values for the

[25] Little and Rubin, *Statistical Analysis with Missing Data*; Marie Reilly, 'Data Analysis Using Hot Deck Multiple Imputation', *The Statistician*, 42 (1993), 307–13; Kim, 'Finite Sample Properties of Multiple Imputation Estimators'; Kim and Fuller, 'Fractional Hot Deck Imputation'; Fuller and Kim, 'Hot Deck Imputation for the Response Model'.

[26] Graham Kalton and Leslie Kish, 'Some Efficient Random Imputation Methods', *Communications in Statistics – Theory and Methods*, 13 (1984), 1919–39; Robert E. Fay, 'Alternative Paradigms for the Analysis of Imputed Survey Data', *Journal of the American Statistical Association*, 91 (1996), 490–8.

[27] Reilly, 'Data Analysis Using Hot Deck Multiple Imputation'.

[28] Reilly, 'Data Analysis Using Hot Deck Multiple Imputation'.

TABLE 2 *The Multiple Hot Deck Imputation Algorithm*

1. Create several copies of the dataset.
2. Search down columns of the data sequentially looking for missing observations.
   a) When a missing value is found, compute a vector of affinity scores, for that missing value.
   b) Create the cell of best donors and draw randomly from it to produce a vector of imputations.
   c) Impute one of these values into the appropriate cell of each duplicate dataset.
3. Repeat Step 2 until no missing observations remain.
4. Estimate the statistic of interest for each dataset.
5. Combine the estimates of the statistic into a single estimate.

same variable because non-observed values cannot be eligible donors (the recipient observation included). The affinity score is used to create the best possible imputation cell for that missing value, from which imputations are randomly drawn. The procedure draws a vector of $M$ donor values and imputes these values into the appropriate cells in the $M$ datasets. This procedure is repeated until all missing values in the $M$ datasets have been imputed. Finally, the statistical model of interest is fitted for each dataset independently and the separate analyses are combined to produce a single estimate using the combination rules standard of parametric multiple imputation. These steps are outlined in Table 2 and described in detail in the following sections.

### *Affinity Scoring*

The selection of donor values is critically important for the validity of the values to be imputed. Donor values should be drawn from respondents (observations) as similar as possible to the respondent with a missing value.[29] In order to measure the extent to which a respondent with a missing value is similar to other respondents, we create a set of affinity scores bounded by 0 and 1, whose components are denoted $\alpha_{i,j}$ and measure the degree of similarity recipient $i$ has to each potential donor $j$. Affinity is defined in terms of the degree to which each potential donor matches the recipient's values across all variables other than the one being imputed.

For each respondent we have the vector $(y_i, x_i)$ where $y_i$ indicates the outcome variable and $x_i$ is a $k$-length vector of purely discrete explanatory variables (an unrealistic assumption that we will relax later), either of which may contain missing values. If the $i$th case under consideration has $q_i$ missing values in $x_i$, then a potential donor vector, $x_j, j \neq i$, will have between 0 and $k - q_i$ exact matches with $i$.

Now define $z_{i,j}$ as the number of variables for which the potential donor $j$ and the recipient $i$ have different values. Thus, $k - q_i - z_{i,j}$ is the number of variables on which $j$ and $i$ are perfectly matched. This value, scaled by the highest number of possible matches $(k - q_i)$ is then the affinity score:

$$\alpha_{i,j} = \frac{k - q_i - z_{i,j}}{k - q_i}. \tag{3}$$

Notice that the affinity score as defined in Equation 3 has the desirable properties that $\alpha_{i,j} = 1$ for $i \in \mathbf{D}_R$ (data with responses) and $\alpha_{i,j} = 0$ for $i \in \mathbf{D}_{NR}$ (data missing responses).

[29] For linguistic parsimony, we generally use the term 'respondent' below, but these methods are immediately applicable to datasets where the rows reflect any other type of observation.

https://doi.org/10.1017/S0007123412000312 Published online by Cambridge University Press

TABLE 3 *Affinity Scoring*

| Matches | Non-matches ($z_{i,j}$) | $\alpha_{i,j}$ |
|---|---|---|
| $k-q_i$ | 0 | $(k-q_i)/(k-q_i)$ |
| $k-q_i-1$ | 1 | $(k-q_i-2)/(k-q_i)$ |
| $k-q_i-2$ | 2 | $(k-q_i-3)/(k-q_i)$ |
| ⋮ | ⋮ | ⋮ |
| 2 | $k-q_i-2$ | $2/(k-q_i)$ |
| 1 | $k-q_i-1$ | $1/(k-q_i)$ |
| 0 | $k-q_i$ | $0/(k-q_i)$ |

*Note*: As the number of matches declines, the affinity score moves towards zero.

TABLE 4 *Affinity Scoring Example*

| | Data matrix (**D**) | | | | |
|---|---|---|---|---|---|
| Observations | $\mathbf{x}_1$ | $\mathbf{x}_2$ | $\mathbf{x}_3$ | | $\boldsymbol{\alpha}_\mathbf{j}$ |
| 1 | 1 | 1 | 0 | | 0.5 |
| 2 | 1 | 1 | 1 | | 1 |
| 3 | 0 | 1 | 1 | | 1 |
| 4 | 1 | 1 | 0 | | 0.5 |
| 5 | 1 | 1 | 1 | ⇒ | 1 |
| 6 | NA | 1 | 1 | | – |
| 7 | 0 | 0 | 1 | | 0.5 |
| 8 | 1 | 0 | 0 | | 0 |
| 9 | 0 | 0 | 0 | | 0 |
| 10 | 1 | 1 | 1 | | 1 |

*Note*: A vector of affinity score is computed for observation $i$=6, which has only one missing value denoted by R language notation for missing values: NA. Three observations with perfect matches on the $X$ variables get a sore of 1, three that match on half get 0.5 and those with no matches get 0. The affinity score for $i$=6 (denoted '–' ) is not a missing element in the α vector, but rather not an element of that vector at all.

Now write the vector of all $\alpha_{i,j}$ as $\boldsymbol{\alpha}_i$. Table 3 shows how the value of $\alpha_{i,j}$ decreases as the number of matches decrease. Cases where the recipient and the donor are *both* missing values in the same covariate are deducted from $k$ and $q_i$ prior to the calculation of Equation 3.

As a simple example, consider the illustrative data in Table 4. For the single missing value of $x_1$ at observation 6, a vector of affinity scores is computed using the formula in Equation 3. Observations that match observation 6 perfectly on the other $x$ variables get affinity scores of 1. For observations that match the recipient observation (observation 6) on one variable but not the other, $\alpha_{i,j} = (3–2)/(3–1) = 0.5$. Lastly, those observations that do not match the recipient on the other $x$ values get an affinity score of 0.

When variables are continuous (or discrete with many categories such that their structure begins to resemble that of continuous variables), the probability of a matched value between a potential donor and a recipient goes to 0 due to the measurement. It is wrong to say that any non-matched values should be treated equally as non-matches in

https://doi.org/10.1017/S0007123412000312 Published online by Cambridge University Press

the discrete sense. Potential donors, $j$, for the $h$th variable in $x_{j[h]}$ that are 'close' to the corresponding $x_{\ell[h]}$ value of the recipient should have greater affinity. This means that there needs to be a definition of 'close' here. If a value $x_{j[h]}$ is closer to $x_{i[h]}$ than some other case's value $x_{\ell[h]}$, then in terms of the variable $h$, $j$ has greater affinity to $i$ than $\ell$. At the same time, if $x_{j[h]}$ and $x_{\ell[h]}$ are very close to one another, but $x_{j[h]}$ is somewhat closer to $x_{i[h]}$, it does make sense to say that the two values are so disparate that they should carry very different affinity values. To allow our affinity scoring technique to address these circumstances, we extend the technique in the following way: for a continuous variable in $x$, we count $x_{j[h]}$ as matching $x_{i[h]}$ if $x_{i[h]}$ and $x_{j[h]}$ are in the same concentric standard deviation from the mean of variable $h$. Values outside of this range are penalized in affinity score by distance. These decisions are the defaults in our R package (hot.deck), but they may be changed by users.

Another option for dealing with continuous variables is to use a metric that measures the distance of observation $i$ and observation $j$ over several $x$ variables. An example of such a technique would be the vector norm $\| x \| = (\mathbf{x}'\mathbf{V}\mathbf{x})^{1/2}$, with positive definite weight matrix $\mathbf{V}$. One may then define $\|x_i - x_j\|$ to be the distance between the vectors $x_i$ and $x_j$, where different standard choices for the matrix $\mathbf{V}$ are possible. For example, by putting the inverse of the of the sample variances of x (in each cell) on the main diagonal of $\mathbf{V}$ or the Mahalanobis distance metric (the inverse of the sample variance-covariance matrix of x (in each cell)). The drawback of such approaches is that, while they provide information about the distance between values of x, whether a given distance would count as a match or not when computing the affinity scores would still depend on arbitrary cutoffs.

Our nonparametric affinity score is based on empirical description, and many other variants exist in various fields. The general affinity score is only one type of similarity measure, and the full class has many uses. Much of the work in this area is done in biology where the concern focuses on similarity and dissimilarity of species types (taxonomy). There is a set of classic alternatives measures in this area including:

- *Czekanowski's Quantitative Index* ($CZ_{ik} = 2\sum_{j=1}^{n} \min(x_{ij}, x_{kj}) / \sum_{j=1}^{n}(x_{ij} + x_{ik})$),
- *Morisita's Index* ($M_{ik} = 2\sum_{j=1}^{n}(x_{ij}x_{kj}) / (\sum_{j=1}^{n} x_{ij}^2 + \sum_{j=1}^{n} x_{kj}^2)$),
- *Canberra Metric* ($C_{ik} = \sum_{j=1}^{n}[(x_{ij} - x_{kj}) / (x_{ij} + x_{kj})] / n$),

matrix-based scores, and more.[30] The tradeoffs between these alternatives are usually application specific. This is also an ongoing research area in information theory where research is centred on comparing text or electronic messages, as well as image restoration. In recent political science methodological research there has been substantial work on causal inference. In this area propensity scores are a common tool for matching cases in observational data as a means of asserting pseudo-experimental control over the assignment of treatment to cases.[31] The propensity score is the probability that a case receives the treatment effect, relative to the control, given this case's covariate levels. Thus, propensity scores are analogous to our affinity scores since they are also a single number summary for each case in the data, bounded by [0:1], and reflect unit-wise similarity. They differ in that propensity scores are an $n$-length vector relative to a single criteria (treatment versus control),

[30] J. C. Gower, 'A General Coefficient of Similarity and Some of its Properties', *Biometrics*, 27 (1971), 857–71.

[31] Paul R. Rosenbaum and Donald B. Rubin, 'The Central Role of the Propensity Score in Observational Studies for Causal Effects', *Biometrika*, 70 (1983), 41–55.

whereas affinity scores are an $n \times n$ lower-triangular matrix of cross-relations between all cases in the data. Propensity scores also differ substantially from our nonparametric affinity score in that they are model based where a statistical specification (usually a logit/probit model) is applied to produce the assignment probabilities for the cases. A critical advantage to our approach, over such strategies, is that a regression model specification, with all of the concerns about covariate and functional form choices, is not necessary.

### *Imputation Using Affinity Scores*

From the vector of affinity scores, our multiple hot deck imputation method is quite intuitive and possesses attractive theoretical properties. Specifically, within the best imputation cell (the set of donors with the highest affinity scores), the realization of missing values is independent of the data. Unless the missing data is non-ignorable, it *must* be the case that missing values within this best cell are missing completely at random (MCAR). Therefore, we can draw randomly from the best cell to fill in missing values without introducing bias. Also, we expect random draws from the best cell to be high quality imputations; because all donors in the cell have the highest possible affinity with the recipient, not only must the occurrence of missing values be completely random within this cell, but so must be the values of the variable itself. In Appendix A, we describe a hybrid hot deck-parametric algorithm to handle datasets that have a mixture of discrete and continuous variables with missing values.

Consider a dataset $\mathbf{D}$, with row vectors $(y_i, x_i)$ and missing values in some or all of its variables, where $\mathbf{D}_R$ continues to denote observed responses and $\mathbf{D}_{NR}$ continues to denote non-responses. For some variable $h$, the value of $x_{i[h]}$ is missing. Intuitively, we could not produce reliable imputations for the missing $x_{i[h]}$ by randomly drawing from the vector $x_{[h]}$ in $\mathbf{D}_R$ since realizations of $x_{[h]}$ may be related to the design of the data. But it is possible to subset the observed data (and thus the $x_{[h]}$ in $\mathbf{D}_R$) in such a way that the realization of $x_{[h]}$ is unrelated to the design.

Each respondent with a missing $x_{i[h]}$ can be considered a member of an *imputation cell*. Imputation cells (sometimes called estimation cells) are subsets of the sample to which respondents are assigned based on characteristics. For instance, if the characteristics of interest are gender and voter turnout, we would have four imputation cells: male voters, male non-voters, female voters, and female non-voters. Assume, for purposes of exposition, that there are an arbitrary number of respondents in $\mathbf{D}_R$ who match recipient $i$ perfectly on all $x_{i[-h]}$ variables.

If the imputation cell, $\mathbf{C}$, contains the set of best possible donors (those with the highest affinity scores), random draws from the observed $x_{j[h]}$ in the cell will be unbiased and have the least amount of imputation variance theoretically supportable. Formally, the observations composing best imputation cell, $\mathbf{C}$, are those for which $\alpha_{i,j} = max\ \boldsymbol{\alpha}_i$ for all $j$ in the cell $\mathbf{C}$. In some cases this set will consist of only perfect matches and in others it will simply consist of the best matches. Returning to the example from Table 4 for intuition, the best imputation cell would consist of only those observations with the highest (in this case, 1) affinity score. So, observations 2, 3, 5 and 10 would be chosen for the imputation cell. Their values on $x_1$ are respectively: $\{1, 0, 1, 1\}$.

By subsetting respondents appropriately into imputation cells, we are able to treat the responses in the best imputation cell $\mathbf{C}$ as the realization of independent and identically distributed (iid) random variables having mean $\mu_C$ and variance $\sigma_C^2$. Now,

$$x_{i[h]} \sim f(\mu_C, \sigma_C^2), \text{ for all } i \text{ in } \mathbf{C}. \tag{4}$$

https://doi.org/10.1017/S0007123412000312 Published online by Cambridge University Press

So, all values of $x_{i[h]}$ in the cell will be drawn from the same distribution, whatever that distribution turns out to be.

More importantly for our purposes, dividing the respondents into cells independent of the sampling and response mechanisms implies that the distribution in Equation 4 holds for both observed and unobserved values alike.[32] Formally,

$$x_{i[h]} \mid (\mathbf{D}, \mathbf{D}_R, \mathbf{D}_{NR}) \sim f(\mu_C, \sigma_C^2), \text{ for all } i \text{ in } \mathbf{C}. \tag{5}$$

This shows that the distribution of $x_{i[h]}$ in the imputation cell is the same regardless of whether it is conditioned on all the data $\mathbf{D}$, the recorded data $\mathbf{D}_R$, or the missing data $\mathbf{D}_{NR}$. Even though the values of $x_{i[h]}$ may be related to the data in the population of interest and the sample, the division into imputation cells removes that dependence within the cell.[33] Therefore, the missing values *must* be missing completely at random (MCAR) within the imputation cell.

Since the missing values within the best imputation cell are now MCAR and there exist realisations of $x_{[h]}$ that are independent of the design, we can draw randomly from the (non-missing) values of $x_{[h]}$ to impute the missing value $x_{i[h]}$. We still need *multiple* imputations in order to capture the imputation variance. The next step is to draw with replacement $M \geq 2$ values of $x_{[h]}$, then assign the $m$th draw to the $m$th duplicate dataset as $x_{i[h]}$ for $\mathbf{D}_m$, $m = 1, \ldots, M$. This process is repeated for every missing value in every column of the dataset. The result is $M$ imputed datasets each with no missing values.

### *Estimation after Multiple Hot Deck Imputation*

The final step, as outlined in the original work on multiple imputation by Rubin *et al.*, is to run $M$ copies of the model specification, taking the mean of the coefficient estimates and the weighted mean of the standard errors, where the weighting accounts for between and within variation of the imputations.[34] Let $\hat{\theta}_m$, $m = 1, \ldots, M$ be coefficient estimates computed individually from the $M$ imputed datasets, and $\sum_m$, $m = 1, \ldots, M$ be the associated variances for $\hat{\theta}_m$. A single estimate of $\theta$,

$$\bar{\theta}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m, \tag{6}$$

is created from the mean of $\hat{\theta}_m$ over the $m$ values.

Calculating the variability of the estimate of $\theta$ is slightly more involved than the mean used to produce $\bar{\theta}_M$ since the total variance is composed of the variation of the coefficient estimates *within* each imputed dataset and the variation of the coefficient estimates *between* the imputed datasets. The within imputation variance is the mean of individual coefficient variances across models:

$$W_M = \frac{1}{M} \sum_{m=1}^{M} \Sigma_m, \tag{7}$$

[32] Kim, 'Finite Sample Properties of Multiple Imputation Estimators'; Kim and Fuller, 'Fractional Hot Deck Imputation'; Fuller and Kim, 'Hot Deck Imputation for the Response Model'.

[33] Kim, 'Finite Sample Properties of Multiple Imputation Estimators'.

[34] Little and Rubin, *Statistical Analysis with Missing Data*; Rubin, 'Multiple Imputations in Sample Surveys'; Rubin, *Multiple Imputation for Nonresponse in Surveys*; Rubin, 'Multiple Imputation after 18+ Years'.

https://doi.org/10.1017/S0007123412000312 Published online by Cambridge University Press

The between imputation variance is the variance of the $M$ coefficient estimates:

$$B_M = \frac{1}{M-1}\sum_{m=1}^{M}(\widehat{\theta}_m - \bar{\theta}_M)^2. \tag{8}$$

The total variance for $\bar{\theta}_M$ is an ANOVA-style weighted sum:

$$T_M = W_M + \left(1 + \frac{1}{M}\right)B_M. \tag{9}$$

where $[1+1/M]$ is the adjustment for finite $M$, and the new degrees of freedom are:[35]

$$df_{M1} = (M-1)\left[1 + \frac{1}{M+1}\ \frac{W_M}{B_M}\right]. \tag{10}$$

These quantities are produced automatically for the user by statistical software implementing multiple imputation.[36]

### *Monte Carlo Experiments*

Researchers in political science with incomplete data generally either casewise delete or use software for parametric multiple imputation. The two most common parametric MI solutions are the R packages 'Amelia'[37] and 'mice'.[38] Other approaches exist, but are not often considered by researchers in the discipline, including specification of the missing data into a Bayesian model as (nuisance) parameters to be estimated, or application of EM.[39] Both of these are model-specific approaches, meaning that they must be uniquely specified for each application, and are, therefore, not directly comparable to imputation tools.

To evaluate the performance of multiple hot deck imputation relative to the most commonly used missing data techniques, we provide a series of Monte Carlo experiments. For all of the experiments, we randomly generated datasets of 500 observations, each with five binary variables with fairly high correlations between them produced by a multivariate normal, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with:

$$\boldsymbol{\mu} = [0,0,0,0,0],\ \ \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.8 & 0.8 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 & 0.8 & 0.8 \\ 0.8 & 0.8 & 1 & 0.8 & 0.8 \\ 0.8 & 0.8 & 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 0.8 & 0.8 & 1 \end{bmatrix} \tag{11}$$

[35] Little and Rubin, *Statistical Analysis with Missing Data*.

[36] Our software formats its output so that the output can be used seamlessly with the R package Zelig; Koske Imai, Gary King and Olivia Lau, 'Zelig: Everyone's Statistical Software', *Comprehensive R Archive Network* (2006). This has the advantage of allowing the user to run, in a single line of code, a great variety of models on the multiple imputed datasets and have the combination handled automatically.

[37] King, Honaker, Joseph and Scheve, 'Analyzing Incomplete Political Science Data'; Honaker and King, 'What to Do about Missing Values in Time-Series Cross-Section Data'.

[38] Stef van Buuren, Jaap P. L. Brand, C. G. M. Groothuis-Oudshoorn and Donald B. Rubin, 'Fully Conditional Specification in Multivariate Imputation', *Journal of Statistical Computation and Simulation*, 76 (2006), 1049–64; Stef van Buuren, 'Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification', *Statistical Methods in Medical Research*, 16 (2007), 219–42.

[39] A. P. Dempster, N. M. Laird and D. B. Rubin, 'Maximum Likelihood from Incomplete Data via the EM Algorithm', *Journal of the Royal Statistical Society, Series B*, 39 (1977), 493–510.

Binary values were created by rounding the draws to 0 or 1 depending on whether they were below or above 0, respectively. This data setup maximally advantages parametric multiple imputation and maximally handicaps our multiple hot deck imputation technique, since it is based originally on a continuous parametric specification. We introduce missing values into two of the five variables where the values are missing at random (MAR), meaning that the occurrence of missing values is related to the observed variables.[40] We designed the introduction of MAR missing values to cause the greatest possible bias. This was done at different levels: the probability that a value was missing from the two affected variables, conditional on the observed values of the other variables all being zero, was applied at 20 per cent, 50 per cent and 80 per cent, respectively. These experiments are labelled MAR-1, MAR-2, and MAR-3, and are run for 10,000 Monte Carlo iterations each, and the results are summarized in Table 5.

The left portion of Table 5 provides the means and standard deviations of the variables with missing values for: the true data (before missing values were introduced), casewise deleted data, multiply hot decked results, and multiply imputed results (as implemented by Amelia). For each of the three MAR studies, the mean of the true data was 0.5 from the data setup. Casewise deletion consistently results in bias, and it is a bias that increases in intensity as the number of missing values increases: a mean of 0.530 for MAR-1 to a mean of 0.656 for MAR-3). Parametric multiple imputation displays a smaller bias than casewise deletion, and this is a bias that also increases as the amount of missing data increases. Notice that the means from the multiply hot decked data display no bias across these three studies, although there is a slight increase in the standard deviation of the means as the proportion of missing values increases, which indicates increasing uncertainty (a desirable result as the volume of missing values goes up). The right portion of Table 5 shows the percentage of imputed values that were imputed correctly for both approaches using the nave criteria. Both multiple hot decking and parametric multiple imputation perform remarkably well in this regard as more missing values are added: hot deck imputation getting about 83 per cent of the missing values correct and parametric multiple imputation getting about 75 per cent correct.

Figure 2 presents the results of logistic regressions conducted at each iteration of the Monte Carlo studies; for each regression, one of the variables with missing values was the outcome variable and two fully observed variables served as explanatory variables. The figure displays the distribution of point estimates over the 10,000 iterations of the Monte Carlo studies for $\beta_0$ (upper-left cell), $\beta_1$ (lower-left cell) and $\beta_2$ (lower-right cell). Within each cell are three sub-plots, from left to right, for the MAR-1, MAR-2 and MAR-3 separate studies. Within each study, looking from left to right, the first boxplot displays the true values of the point estimates, the second boxplot displays the point estimates under casewise deletion, the third boxplot displays those from multiple hot deck imputation, and the fourth boxplot for parametric multiple imputation.

Figure 2 shows again that casewise deletion results in consistently biased point estimates and that this bias becomes more severe as progressively more missing values are added to the data. As with the mean summary in Table 5, parametric multiple imputation typically displays less bias than casewise deletion, but also a bias that gets worse as the volume of missing values increases. Interestingly, for the estimation of $\beta_1$, parametric

[40] We also ran experiments where the missing values were MCAR, but, as we would expect theoretically, no method was biased under those conditions.

https://doi.org/10.1017/S0007123412000312 Published online by Cambridge University Press

TABLE 5 *Means and Percentages Correctly Imputed for Different Imputation Techniques in the Monte Carlo Experiments*

| | Means | | | Percentage correctly imputed | | |
|---|---|---|---|---|---|---|
| | MAR-1 | MAR-2 | MAR-3 | MAR-1 | MAR-2 | MAR-3 |
| True | 0.500 (0.023) | 0.500 (0.022) | 0.500 (0.022) | | | |
| Casewise deletion | 0.530 (0.023) | 0.585 (0.024) | 0.656 (0.025) | | | |
| Multiple hot deck | 0.500 (0.023) | 0.500 (0.024) | 0.500 (0.028) | 83.682 (3.667) | 83.387 (2.969) | 83.120 (3.666) |
| Parametric MI | 0.507 (0.023) | 0.518 (0.023) | 0.530 (0.025) | 75.253 (3.306) | 74.805 (2.460) | 74.218 (2.560) |

*Note*: Standard deviations are in parentheses.

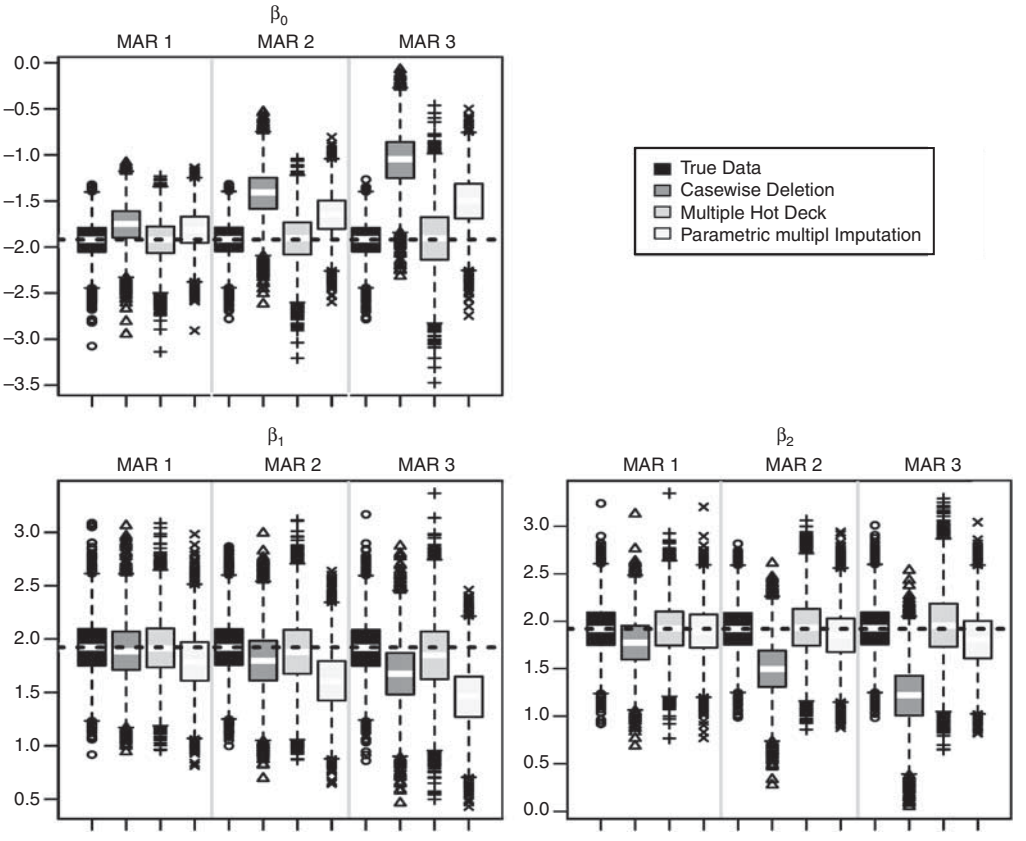


*Fig. 2. Comparison of logistic regression results from Monte Carlo experiments*

multiple imputation produces a more severe bias than casewise deletion. The bias of the point estimates produced by parametric multiple imputation is due to the rounding discussed in the section 'Parametric Multiple Imputation and Its Shortcomings'. Here, multiple hot deck imputation performs quite well: its median point estimates are consistently closer to the median of the true point estimates than the other techniques. The interquartile range of the multiply hot decked point estimates increases as more missing values are added to the data, which usefully describes increased uncertainty as there is less data on which to condition.

## EXAMPLE: MODERNIZATION THEORY

We saw that parametric multiple imputation can produce biased results in simulation studies, and, through consideration of this problem in an applied political science context, we demonstrate the advantages of using the multiple hot deck method with highly discrete data.

Transitions to and from democracy are both normatively and scientifically important. A voluminous literature on such transitions, springing from a seminal study by Lipset,[41]

[41] Seymour M. Lipset, 'Some Social Requisites of Democracy: Economic Development and Political Legitimacy', *American Political Science Review*, 53 (1959), 69–105.

https://doi.org/10.1017/S0007123412000312 Published online by Cambridge University Press

has focused on the tendency of states with higher average incomes to be more democratic. The theory that higher average wealth produces and maintains democracies is referred to as *modernization theory*. The central claim of modernization theory had been robustly supported by large-*N* empirical studies[42] until a line of research produced principally by Przeworski *et al*. called part of it into question.[43] Specifically, Przeworski *et al*. claim that transitions to and from democracy are not correlated with income, but the maintenance of democracy is.[44] This claim has sparked theoretical controversy as well as a debate about the use of transition models and modernization model specification generally.[45] While engaging the theoretical debate or the methodological discussion related to modernization theory is beyond the scope of our pedagogical example, we use this debate as an opportunity to demonstrate our technique while examining the relationship between economic performance and transitions to and from democracy. We build our example around a re-analysis of the data examined by Przeworski *et al*.[46]

Using Przeworski *et al*.'s data, which include annual measures on 135 countries between the years 1950 and 1990, we consider a simple logistic switching model for several reasons.[47] First, the model has a simple form and specification, making it useful for demonstration purposes. We keep the specification simple so as to focus on the missing data problem and not reduce the transparency of the application through more elaborate specification. Secondly, many of the variables of interest are discrete in nature; further supporting the example's pedagogical aims. Lastly, Przeworski *et al*.'s data have no missing values.[48] This last point is critical because, if values were actually unknown, we would not be able to determine which method is better here. At most, we would be able to show divergent results, but such a study is necessarily inconclusive. Since the data are

[42] Phillips Cutright, 'National Political Development: Its Measurement and Social Correlates', in Nelson W. Polsby, Robert A. Dentler and Paul A. Smith, eds, *Politics and Social Life: An Introduction to Political Behavior* (Boston, Mass.: Houghton Mifflin, 1963), 569–81; Karl W. Deutsch, 'Social Mobilization and Political Development', *American Political Science Review*, 55 (1961), 493–510; Robert A. Dahl, *Polyarchy: Participation and Opposition* (New Haven, Conn.: Yale University Press, 1971); Ross E. Burkhart and Michael S. Lewis-Beck, 'The Economic Development Thesis', *American Political Science Review*, 88 (1994), 903–10; John B. Londregan and Keith T. Poole, 'Does High Income Promote Democracy?' *World Politics*, 49 (1996) 1–30.

[43] Adam Przeworski, *Democracy and the Market: Political and Economic Reforms in Eastern Europe* (New York: Cambridge University Press, 1991); Adam Przeworski, *Democracy and the Market: Political and Economic Reforms in Eastern Europe* (New York: Cambridge University Press, 1991); Adam Przeworski and Fernando Limongi, 'Political Regimes and Economic Growth', *Journal of Economic Perspectives*, 7 (1993), 51–69; Adam Przeworski, Michael E. Alvarez, Jose A. Cheibub and Fernando Limongi, 'What Makes Democracies Endure?' *Journal of Democracy*, 7 (1996), 39–55; Adam Przeworski and Fernando Limongi, 'Modernization: Theories and Facts', *World Politics*, 49 (1997), 155–83; Adam Przeworski, Michael E. Alvarez, Jose A. Cheibub and Fernando Limongi, *Democracy and Development: Political Institutions and Well-Being in the World, 1950–1990* (New York: Cambridge University Press, 2000).

[44] Przeworski, Alvarez, Cheibub and Limongi, *Democracy and Development*.

[45] See, for example: Carles Boix, *Democracy and Redistribution* (New York: Cambridge University Press, 2002); Carles Boix and Susan Stokes, 'Endogenous Democratization', *World Politics*, 55 (2003), 517–49; David L. Epstein, Robert Bates, Jack Goldstone, Ida Kristensen and Sharyn O'Halloran, 'Democratic Transitions', *American Journal of Political Science*, 50 (2006), 551–69.

[46] Przeworski, Alvarez, Cheibub and Limongi, *Democracy and Development*.

[47] Przeworski, Alvarez, Cheibub and Limongi, *Democracy and Development*.

[48] Przeworski, Alvarez, Cheibub and Limongi, *Democracy and Development*.

completely observed we can compare casewise deleted, multiple hot decked, and multiple imputed results with the 'true' results.[49]

The outcome variable of *Autocracy* is coded 1 for autocracies and 0 for democracies, with transition years coded according to the regime that emerged that year. The explanatory variables we consider, based largely on Przeworski *et al.*'s models,[50] are lagged values of real gross domestic product (*GDP*) per capita (standardized to 1985 prices), the annual rate of *Growth* in per-capita income, an indicator variable coded 1 if the country is a former *British Colony*, an indicator coded 1 if the country is a primary *Commodity Exporter*, and an indicator coded 1 for countries that gained their independence within the previous five years (*New Country*).

The first of our two switching models examines transitions from democracy to autocracy. The model is specified as:

$$P(Autocracy_t = 1 \mid Autocracy_{t-1} = 0) = \text{logit}^{-1}(GDP + GDP\,Growth + British\,Colony + Commidity\,Exporter + New\,Country). \quad (12)$$

Modelling the probability of an autocratic regime at time $t$ ($Autocracy_t = 1$) given that the previous year's regime was a democracy ($Autocracy_{t-1} = 0$) with a logistic regression captures the probability of a democracy-to-autocracy switch conditional on the covariates. By the same logic, we use:

$$P(Autocracy_t = 0 \mid Autocracy_{t-1} = 1) = \text{logit}^{-1}(GDP + GDP\,Growth + British\,Colony + Commidity\,Exporter + New\,Country). \quad (13)$$

to model the switch from autocracy in the previous year ($Autocracy_{t-1} = 1$) to democracy in the current year ($Autocracy_t = 0$).

We introduced missing values into the explanatory variable, *Autocracy*, according to the MAR assumption: observations were deleted conditional on extreme values of the explanatory variables until slightly more than half (57 per cent) of the values were missing. We then recorded the true results, the casewise deleted results, those from our multiple hot deck imputation method, and those from parametric multiple imputation (as implemented in Amelia).

The results, displayed in Figure 3, are compelling. First, the casewise deleted results show a loss of statistical reliability for five coefficient estimates across the two models. Worse yet, the casewise deleted coefficient was of a different sign than the true point estimate for the new country indicator in the democracy-to-autocracy model and for British Colony in the autocracy-to-democracy model. The multiple hot decking method generally produces point estimates closer to the true point estimate than parametric multiple imputation. This is clearly visible for lagged GDP and New Country in the autocracy-to-democracy model as well as for lagged GDP growth, and the indicators for British colonies, commodity exporters and new countries in the autocracy-to-democracy model. It is striking that if one used any technique other than multiple hot deck imputation, one would make a substantively false inference about the effect of being a new country has on transitions to democracy. These results suggest that not only is the

[49] The true results are true to the extent that they are the results actually obtained by analysing the complete data. They are not true in the more traditional sense of being the true population parameters an empirical analysis attempts to estimate.

[50] Przeworski, Alvarez, Cheibub and Limongi, *Democracy and Development*.

https://doi.org/10.1017/S0007123412000312 Published online by Cambridge University Press

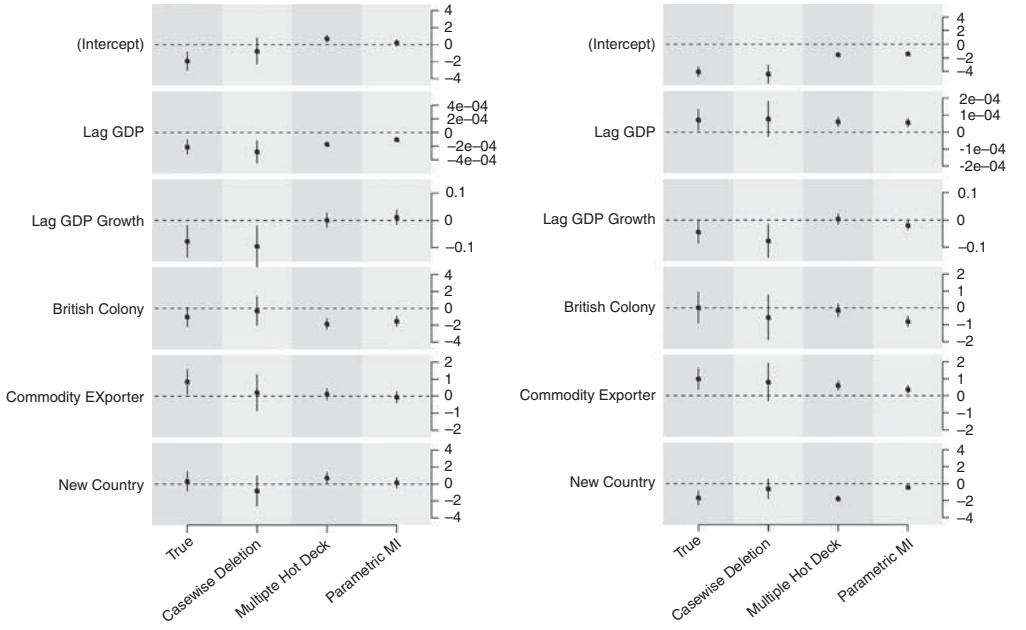


*Fig. 3. Switching model results under different treatments of missing values*
*Note*: The left cell of the figure examines transitions from democracy to autocracy and the right cell examines transitions from autocracy to democracy. Note that if the vertical 95 per cent confidence bands emanating from the point estimates include zero, the coefficient is not statistically reliable at the traditional 0.05 level.

use of any imputation technique over casewise deletion essential, but multiple hot deck imputation consistently produces coefficients closest to the true coefficient and, in some cases, is the only technique that produces a substantively correct inference.

## LIMITATIONS

Intuitively, the multiple hot deck is best suited to cases where observations are many and variables are relatively few. In such cases, there will be many potential donors. A potential shortcoming of the method is that the best imputation cell could contain too few donor observations. It is possible that the best cell could be composed of a single donor, thus rendering the 'multiple' part of multiple imputation irrelevant and the method will not produce or capture imputation variance. A partial remedy is to look to other, hopefully related, variables and carefully select a set of these to include in the imputations model. This works well when there are a large number of potential target variables, and essentially returns to old-style hot decking. Relatedly, if the overall sample size is very small, then typically there are also too few potential donors for missing values. This problem is noted for sample sizes under fifty and for cases where there are many covariates relative to the sample size. However, all imputation techniques are challenged by small sample sizes since it reduces the availability of information required to create conditional statements.

It is also worth emphasizing a limitation that our approach shares with all other multiple imputation methods: multiple imputation will only produce unbiased results in situations where the missing values are missing completely at random (MCAR) or missing

https://doi.org/10.1017/S0007123412000312 Published online by Cambridge University Press

at random (MAR) as discussed above in the section 'Bias from Missing Values'. If the missing values are non-ignorable (NI), it is, by definition, impossible that *any* imputation method not drawing on information from outside the dataset will produce unbiased estimates except by chance. Researchers must be mindful of this limitation and consider carefully whether assumptions of MCAR or MAR are justified for a dataset under analysis.

## CONCLUDING REMARKS

Missing data is an empirical problem that cannot be ignored in political science. Researchers *must* make some decision about how to handle the missing values in their data since regression models and other tools fail when they reach an unobserved cell. This makes the study of missing data different from other methodological issues since it cannot be avoided, and we have shown here that casewise deletion is not a benign default solution. Therefore, all data-analysts need a set of approaches to handle this challenge under different circumstances. In this work we have added another approach to that set that is focused on one particular type of data measurement.

We have demonstrated that multiple hot deck imputation is a better approach than parametric multiple imputation for discrete data. While using any imputation technique is an improvement over casewise deletion, applying parametric multiple imputation to discrete data can lead to a number of problems, chief among which are nonsensical imputations and biased results associated with rounding such imputations. Multiple hot deck imputation is an attractive alternative to parametric multiple imputation because the hot deck method uses draws from the actual variable with missing values, thus preserving the discrete properties of the data as their missing values are imputed, and requires none of the parametric assumptions upon which traditional multiple imputation methods rely.

Multiple hot deck imputation also possesses appealing theoretical properties. First, it combines the strengths of traditional hot deck imputation and parametric multiple imputation, maintenance of discrete data and the incorporation of uncertainty based on the imputation process respectively, to overcome both of their weaknesses: a failure to account for imputation uncertainty and biased performance with discrete data. Secondly, because the occurrence of missing values in the best imputation cell are completely random and the values of variables in the cell are independent of the data, the multiple hot deck method must produce unbiased imputations and it must give the best possible estimates of the values to be imputed.

The advantages of the multiple hot deck method are clear in our Monte Carlo experiments. Both casewise deletion and parametric multiple imputation displayed the biases we expected based on theory, casewise deletion having a more severe bias than parametric multiple imputation on average, but both biases increasing with the number of missing values. Conversely, multiple hot deck imputation provided unbiased estimates and displayed the increase in variance with more missing values than one desires based on fewer data on which to condition.

The empirical example, with data from Przeworski *et al.*,[51] shows the dangers of using poor missing data solutions. Political scientists are generally aware of the dangers of casewise deletion, even if some of them apply it nonetheless. What is less well understood is that different remedies principally used to deal with missing values have different properties. Here, we add a new tool to the empirical political scientist's toolbox that does better than standard multiple imputation for highly non-granular discrete variables.

[51] Przeworski, Alvarez, Cheibub and Limongi, *Democracy and Development*.

https://doi.org/10.1017/S0007123412000312 Published online by Cambridge University Press

## APPENDIX A

### IMPUTING MIXES OF DISCRETE AND CONTINUOUS DATA

When the assumptions are met, parametric multiple imputation continues to be the ideal method for imputing continuous data, whereas the multiple hot deck method is better for imputing discrete data. However, in many cases a dataset has missing values in a mixture of continuous and discrete variables. Here, we describe a looping algorithm that makes use of parametric multiple imputation for continuous variables and multiple hot decking for discrete variables. The process alternates techniques and repeats itself so that every time the loop re-starts, it is drawing on better information (somewhat analogously to Gibbs sampling).

TABLE 6 *The Mixed Missing Values Imputation Algorithm*

Loop {
1. Parametric multiple imputation on continuous missing values
2. Multiple hot decking on discrete missing values
3. Remove parametrically imputed values
4. Repeat
5. Remove multiply hot decked values
6. Repeat

} until all changes $< \delta$

The mixed missing values imputation algorithm proceeds as follows. On the entire dataset, it (step A) runs parametric multiple imputation to fill in the continuous missing values and combines imputations according to the averaging rules (our section 'Estimation After Multiple Hot Deck Imputation'). This leaves a dataset for which the missing continuous values have been filled in via parametric multiple imputation, but the missing discrete values remain missing. The algorithm then (step B) uses multiple hot deck imputation, conditioning on the data including the just-imputed continuous values, to fill in the missing discrete values and combines imputations using a modified Rubin combination. This is modified only to take medians rather than means in the interest of preserving the discrete nature of the imputations. Then, the algorithm (step C) removes the imputations made via parametric multiple imputation in the first step A and repeats step A (running parametric multiple imputation for the continuous missing values). Note that removing the first continuous imputations and running parametric multiple imputation again, after having run both parametric multiple imputation and multiple hot decking, means that the second round of parametric multiple imputation draws on better information than the first. Next, (step D) the multiple hot decked imputations are removed and multiple hot deck imputation is run again (repeated step B). Now, the discrete imputations are drawing on better information than they did in the first step B. This process is repeated until changes in the imputations between one round of the loop and the next are less than some (small) threshold $\delta$. By default, our software uses the threshold for IWLS recovered from the user's settings via glm.control()$epsilon. If after IWLS's maximum iterations (glm.control()$maxit), iterative changes remain greater than $\delta$, the program terminates and warns the user of unstable imputations.

## APPENDIX B SOFTWARE

To implement the multiple hot deck imputation procedure, we have produced an R package called hot.deck (version 1.1), which we make freely available on the comprehensive R archive network (CRAN). The 'hot.deck package' is designed with user-friendliness in mind. The syntax is simple and consistent with the language standards. Users implement the multiple hot decking procedure with the single argument:

**mids.out <- hot.deck(my.data, m = 5),**

where **mids.out** is the output list of multiple hot decked datasets, **my.data** is the dataset with missing values, and **m** specifies the number of multiple hot decked datasets to create (the default is 5).

https://doi.org/10.1017/S0007123412000312 Published online by Cambridge University Press

The software determines the measurement of the data and applies the mixed missing values algorithm described in Appendix A if necessary (by default, variables with more than ten unique realizations are treated as continuous). The output from the **hot.deck** function is a set of multiply hot decked datasets. This output is in a format such that it can be used seamlessly with standard regression models and omnibus solutions such as the 'Zelig' package.[52]

[52] Imai, King and Lau, 'Zelig'.

https://doi.org/10.1017/S0007123412000312 Published online by Cambridge University Press