DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 675–700, http://dx.doi.org/10.2478/JOS-2014-0043

The Utility of Nonparametric Transformations for Imputation of Survey Data

Michael W. Robbins[1]

Missing values present a prevalent problem in the analysis of establishment survey data. Multivariate imputation algorithms (which are used to fill in missing observations) tend to have the common limitation that imputations for continuous variables are sampled from Gaussian distributions. This limitation is addressed here through the use of robust marginal transformations. Specifically, kernel-density and empirical distribution-type transformations are discussed and are shown to have favorable properties when used for imputation of complex survey data. Although such techniques have wide applicability (i.e., they may be easily applied in conjunction with a wide array of imputation techniques), the proposed methodology is applied here with an algorithm for imputation in the USDA's Agricultural Resource Management Survey. Data analysis and simulation results are used to illustrate the specific advantages of the robust methods when compared to the fully parametric techniques and to other relevant techniques such as predictive mean matching. To summarize, transformations based upon parametric densities are shown to distort several data characteristics in circumstances where the parametric model is ill fit; however, no circumstances are found in which the transformations based upon parametric models outperform the nonparametric transformations. As a result, the transformation based upon the empirical distribution (which is the most computationally efficient) is recommended over the other transformation procedures in practice.

Key words: Missing data; multiple imputation; empirical CDF; kernel density; ARMS; Markov chain Monte Carlo.

1. Introduction

Missing data are a particularly common and particularly troublesome problem in establishment surveys. A large portion of the statistical literature has been devoted to the analysis of data that contain missing values, and as a result a myriad of approaches exist. Pertinent techniques include calibration weighting (Kott and Chang 2010) and the EM algorithm (Dempster et al. 1977); however, imputation (for a summary, see Rubin 1987) is often the preferred method for handling missing data since it yields a completed dataset on which classical tools for analysis may be applied. Additionally, multiple (or repeated)

[1]Associate Statistician, RAND Corporation, Pittsburgh, PA 15213 U.S.A. Email: mrobbins@rand.org
Acknowledgments: The author acknowledges partial funding from USDA grant #58-3AEU-2-0065, from the Cross-Sector Research in Residence Program between the National Institute of Statistical Sciences (NISS) and National Agricultural Statistics Service (NASS), and from the University of Missouri Research Board. The author acknowledges and thanks Sujit Ghosh, Barry Goodwin, Joshua Habiger, Darcy Miller and Kirk White for their contributions to the research project associated with the work presented here. The author thanks the editoral staff and anonymous reviewers whose helpful comments greatly improved the article. The views expressed are those of the author and do not necessarily represent the views of RAND, NASS or the USDA.

imputation (Rubin 1996) may be used to quantify imputation error. Despite the ubiquity of missing data problems and methodology designed to address them, existing imputation algorithms have many drawbacks, largely with respect to robustness and computational efficiency.

Multivariate imputation techniques tend to be fairly restrictive with respect to the types of model assumptions. Techniques that impute via a multivariate normal model (Schafer 1997; Robbins et al. 2013) are popular and theoretically justified. Techniques that use fully conditional specification (a.k.a. SRMI, as outlined in Raghunathan et al. 2001), which is implemented in several software packages including IVEware (Raghunathan et al. 2002), MICE (Van Buuren and Oudshoorn 1999), and mi (Su et al. 2011), can be used to create imputations in data that contain categorical and discrete variates but lack theoretical justification due to the use of a potentially incompatible Gibbs sampler. However, each of the aforementioned procedures is best suited to sample (i.e., draw) imputations for continuous variables from a normal distribution.

Multivariate techniques that do not sample imputations for continuous items under Gaussian assumptions are relatively sparse. Algorithms which employ fully conditional specification can be modified so that imputations are generated via a conditional modeling/sampling technique known as predictive mean matching (PMM, Little 1988). PMM is a nearest-neighbor procedure; imputations are sampled from observed data values. However, PMM is computationally burdensome in comparison to its Gaussian counterparts and thus can have little utility in high dimensional settings. The IRMI algorithm (Templ et al. 2011) is similar in structure to SRMI-type procedures with the added functionality of estimating conditional models through robust regression; however, steps are not taken to ensure that imputations are sampled from the true conditional distribution, which implies that IRMI imputations will likely distort complex distributional characteristics (further justification for this claim is provided in Section 5). To increase the robustness of traditional normality-based methods, many authors recommend the use of marginal transformations of continuous variates prior to the application of imputation methodology. For example, Raghunathan et al. (2001) suggest a power transformation, whereas Robbins et al. (2013) suggest a density-based transformation (specifically, a skew-normal density is used).

The practicality of the aforementioned procedures is muddled by their computational complexity. The growing ubiquity of multiple imputation, the prevalence of iterative sampling techniques (e.g., Markov chain Monte Carlo) for imputation, and the high dimensional nature of modern statistical analyses result in algorithms that mandate a substantial computational burden. Such issues become increasingly problematic under the guise of the benefits provided by the use of a wide-ranging imputation model (Robbins and White, Forthcoming).

Here, the transformation-based schemes of Robbins et al. (2013) are extended, resulting in the introduction of robust techniques for transformation. In particular, a transformation based on the kernel density is suggested. Woodcock and Benedetto (2009) use a kernel density to generate data values for the purpose of creating a public use dataset from confidential data. Additionally, a fully empirical transformation (which uses a modified empirical distribution) is presented here. The empirical transformation yields a hot-deck

(or nearest-neighbor) technique that may be applied jointly with commonly used multivariate imputation algorithms (such as IVEware, MICE or mi) in a very computationally efficient manner. The proposed methodologies yield simple tools which uphold the ability to preserve complex distributional structures provided by PMM while maintaining the computational efficiency of techniques which mandate Gaussian assumptions.

In this article, imputations for a widely-used data product are generated via the aforementioned transformation techniques. The marginal and multivariate efficacy of the resulting imputations, as well as the inadequacies of imputations generated using a fully parametric model, are illustrated. Specifically, in Section 2, the dataset that will be used throughout, and the technique that will be used to generate imputations (following transformation), are introduced. The robust methods of transformation are presented in Section 3, and data analysis is provided in Section 4. Further, Section 5 presents a simulation study (performed using real and synthetic data) that illustrates the effectiveness of the proposed transformation schemes. The article concludes by providing comments and practical advice in Section 6.

2. The ARMS and Associated Imputation Technique

In June 2009, a research project commenced with the goal of creating a new imputation method for the US Department of Agriculture's (USDA) Agricultural Resource Management Survey (ARMS). Partial findings of the research project are outlined in Robbins and White (2011), Robbins et al. (2013) and Robbins and White (Forthcoming); this article relates additional findings of the project. Although the methodologies presented here are widely applicable, the problem of interest is motivated here through a discussion of the ARMS and its recently developed imputation technique.

ARMS data are a key source of information for congressional decisions that allocate billions of dollars in farm subsidies (Robbins et al. 2013). The survey provides the USDA's most comprehensive view of the American farm household; ARMS data contain 30,000–40,000 units (observations) with 1,000–2,000 items (variables). The ARMS has a multiphase, dual-frame, stratified, probability-weighted sampling design. Design weights are calibrated, and the calibrated weights are used to calculate key survey indications (U.S. Department of Agriculture 2011). Calibration of design weights also accounts for unit nonresponse; the rate of unit nonresponse tends to hover around 30% (National Research Council 2008). Analyses presented herein use data from the 2010 ARMS.

Aside from being high dimensional, ARMS data have a complex distributional structure – the majority of ARMS variables have semicontinuous distributions. To elaborate, a portion of units will report a zero for a given variable, whereas the responses for the remaining observations for that variable are sampled from some strictly positive and (theoretically) continuous distribution.

The new ARMS imputation procedure handles semicontinuous variables via a commonly used mixture model (see Javaras and van Dyk 2003, for example). Specifically, a semicontinuous variable $\boldsymbol{Y}$ is broken down into two latent variables, $\boldsymbol{B}$ and $\boldsymbol{Y}^*$, where $\boldsymbol{B}$ is an indicator variable denoting whether or not $\boldsymbol{Y}$ is positive, and $\boldsymbol{Y}^*$ is a strictly

continuous variable that indicates the positive portion of $\boldsymbol{Y}$. The imputation algorithm treats $\boldsymbol{Y}^*$ as missing whenever $\boldsymbol{Y}$ is missing or 0. All semicontinuous ARMS variables are transformed in this manner, and all ARMS variables with missing values are assumed to be semicontinuous. See Su et al. (2011) for an example of an extant procedure that utilizes similar approaches for handling semicontinuous data. Another key characteristic of the missingness in ARMS data is that all missing values are assumed to be positive. Thus, $\boldsymbol{B}$ is fully observed for all variables.

The positive portions of ARMS variables (i.e., the $\boldsymbol{Y}^*$s) tend to be highly skewed. Since all imputation procedures that are practical in high-dimensional settings link variables through a multivariate normal model, each $\boldsymbol{Y}^*$ is transformed in order to achieve normality. Letting $\boldsymbol{X}$ (which is theoretically Gaussian) represent a transformed version of $\boldsymbol{Y}^*$, Robbins et al. (2013) provide the following procedural outline of the algorithm for imputation in ARMS data:

1. Break each semicontinous variable $\boldsymbol{Y}$ into $\boldsymbol{B}$ and $\boldsymbol{Y}^*$ (observed 0s are treated as missing).
2. Transform: $\boldsymbol{Y}^* \Rightarrow \boldsymbol{X}$ for each variable.
3. Impute: Find $\hat{\boldsymbol{X}}$ (the imputed version of $\boldsymbol{X}$) for each variable.
4. Untransform: $\hat{\boldsymbol{X}} \Rightarrow \hat{\boldsymbol{Y}}$ (the imputed version of $\boldsymbol{Y}$) for each variable (values that are originally observed as 0 are reset to 0).

The imputed data also undergo an editing process to ensure that imputations satisfy all data constraints prior to release. Most variables are not subject to such constraints, and the editing process does not damage the quality of the imputations with regards to analytic properties.

Robbins et al. (2013) focus on Step 3 above. For that purpose, they introduce a dynamic imputation procedure, the so-called iterative sequential regression (ISR) method, that builds a multivariate (normal) model for the $\boldsymbol{X}$s (and respective covariates) through a sequence of conditional linear models while allowing flexibility in the form of each conditional model. For the purpose of transformation, they apply a skew-normal model (Azzalini 1985) to the logged versions of the $\boldsymbol{Y}^*$s. It had been established that such a transformation is sufficient for the majority of ARMS variables (Miller et al. 2010). However, for certain ARMS variables (and surely data from most any other survey) such a model is insufficient.

As a result, the focus here turns to Steps 2 and 4 above: the mechanisms for transformation. We present robust nonparametric methods for transformation that will retain the applicability of the ISR procedure while ensuring that imputations preserve the marginal structure of complex survey variables (as will be illustrated in the sections that follow). It is emphasized that the methods presented in the following are widely applicable; these techniques may be applied to any data that contain theoretically continuous (or semicontinuous) variables and may be applied in conjunction with a wide array of imputation procedures.

To help illustrate the applicability of the methodology presented here to general imputation problems, statistical analyses that require the specific ARMS design are not the focus here. Regardless, the survey design is not expected to have a substantial influence on the choice of transformation scheme.

In this article, the standard errors of estimators derived using imputed data are adjusted for imputation error via multiple imputation (MI, Rubin 1987;1996). MI involves the generation of multiple datasets which have been imputed independently of one another; imputations are presumed to have been randomly sampled from the posterior distribution of the missing data given the observed data. Rubin's rules for combining information across datasets have been provided in a number of references (including the two given above). The validity of MI inferences in settings where complex survey data are used has been called into question frequently (Kott 1995; Fay 1996; Kim et al. 2006). Although MI has demonstrated utility for analysis of ARMS data (Robbins and White, Forthcoming), MI is used here primarily due to its simplicity and effectiveness in comparing imputation error across datasets imputed via differing methods.

3. Transformation Techniques

Let the length-n vector $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}'$ denote a survey variable, where n is the sample size (i.e., number of experimental units). To develop a transformation scheme that attains normality, consider the fact that any continuous random variable with a known cumulative distribution function (CDF) can be transformed into a standard normal variate. Specifically, let X be any scalar random variable with known CDF $F(x)$, and let

$$T(x) = \Phi^{-1}(F(x)) \tag{1}$$

represent the transformation function, where $\Phi(\cdot)$ denotes the standard normal CDF, then

$$T(X) \sim N(0, 1)$$

It is noted that when variables are transformed via (1) and then linked through a multivariate normal distribution (which is the model used for imputation here), the resulting model may be considered a Gaussian copula (Nelsen 2009).

The impasse with respect to application of the above transformation scheme is the fact that in practical circumstances, the CDF $F(x)$ tends to be unknown. Thus, in order to apply the above transformation to the positive portions survey, it is necessary to first develop a manner for determining (or approximating) the CDF of these positive portions. As mentioned above, a log-skew-normal model suffices for the majority of ARMS variables. That is, in accordance with (1); Robbins et al. (2013) suggest that if

$$T_1(y) = \Phi^{-1}(F(y|\hat{\xi}, \hat{\omega}, \hat{\alpha})) \tag{2}$$

then $T_1(\log Y_i)$ should have (or approximately have) a standard normal distribution for all relevant i. In the above, $\{\hat{\xi}, \hat{\omega}, \hat{\alpha}\}$ represent consistent estimators of the skew-normal parameters. Clearly, such marginal transformations provide no general implication that joint normality will be obtained; however, Robbins et al. (2013) illustrate rigorously that for ARMS data multivariate normality is (adequately) achieved through marginal transformation to normality. It is noted that these conclusions also hold when the nonparametric transformations proposed herein are used.

As was also mentioned above, the transformation in (2) is inadequate for certain ARMS variables. For instance, labor variables, where the response indicates the number of weekly

Table 1. List and description of ARMS variables pertinent to this study. The number of positive and observed (n_{obs}) values and the number of missing values (n_{mis}) is provided for each listed variable. Within the simulation study of Subsection 5.1, additional missingness is imposed in the variables marked with an asterisk

Name	Description	n_{obs}	n_{mis}
P758*	Operator's expenditure for hired labor	9,354	0
P764*	Operator's wage expenditure for operator	1,296	0
P784*	Contractor's expenditure for contract labor	151	0
P828*	Operator's on-farm labor (in hrs/wk) for Jan.–Mar.	19,285	1,296
P829*	Operator's on-farm labor (in hrs/wk) for Apr.–Jun.	19,342	1,438
P830	Operator's on-farm labor (in hrs/wk) for Jul.–Sept.	19,274	1,474
P831	Operator's on-farm labor (in hrs/wk) for Oct.–Dec.	19,114	1,517
P832*	Spouse's on-farm labor (in hrs/wk) for Jan.–Mar.	8,991	533
P833*	Spouse's on-farm labor (in hrs/wk) for Apr.–Jun.	9,298	513
P834	Spouse's on-farm labor (in hrs/wk) for Jul.–Sept.	9,298	529
P835	Spouse's on-farm labor (in hrs/wk) for Oct.–Dec.	9,097	559
P884	Estimated value of farm credit stock on Dec. 31	4,273	1,025
P952*	Operator and spouse off-farm labor	10,462	1,081

hours worked, tend to observe onerous marginal distributions. Names and descriptions of ARMS variables that will be discussed in this study are given in Table 1. Names of ARMS variables are formed by placing a "P" in front of the numeric item code seen on the survey questionnaire.

As an example, Figure 1 provides a histogram of log(P829) with the best-fitting skew-normal density curve. Only positive responses for this variable are included in this graph (and similar plots that follow). A scatter plot of log(P829) and log(P830) is also provided in the figure to illustrate the bivariate dispersion of the data points. Likewise, only units that report positive values for both variables are plotted in this graph (and similar ones that follow). These labor variables are analyzed on the log scale because logged values are closer to being Gaussian than the untransformed values.

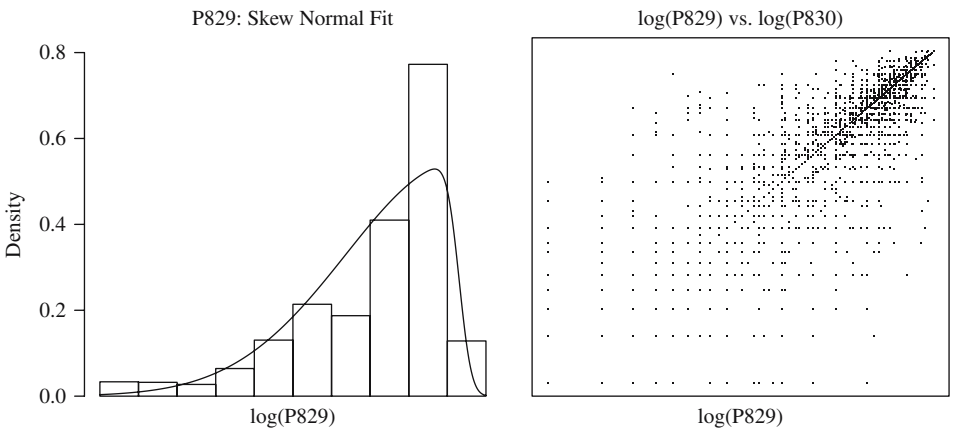

Fig. 1. Histogram of log(P829) (left) and scatter plot of log(P829) versus log(P830) (right). The left plot has the best fitting skew-normal density curve overlaid. Axis values are suppressed to avoid disclosure where necessary

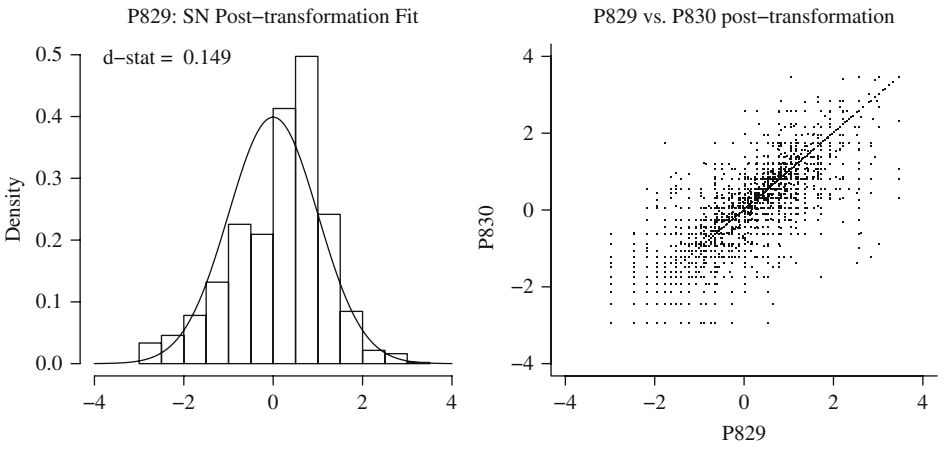

Fig. 2. Histogram of log(P829) (left) and scatter plot of log(P829) versus log(P830) (right) following skew-normal transformation. The left plot has the standard normal density curve overlaid

To illustrate further the specific deficiencies of the skew-normal (SN) transformation for the labor variables, Figure 2 provides a histogram log(P829) and a scatter plot of log(P829) versus log(P830), all following skew-normal transformation. As the transformed data should observe a standard normal distribution, the standard normal density is plotted over the histogram of the transformed data. Additionally, a Kolmogorov-Smirnov (KS) test under the assumption of a standard normal distribution is applied to the transformed values shown in the left graph in Figure 2, and the distance statistic (d-stat) is given in the upper-left corner of the plot. Labor variables such as P829 tend to have repeating values, which makes the KS test theoretically inappropriate, but such results are given here and in further plots for a comparison of goodness of fit.

The power (or Box-Cox) transformation is often applied within imputation procedures (e.g., Raghunathan et al. 2001). However, the Box-Cox transformation show no increase in utility over the log-skew-normal transformation described above; therefore it is not discussed further. A more robust transformation scheme is clearly warranted. Accordingly, nonparametric models for $F(x)$ are considered.

3.1. Transformation Via the Kernel Density

Next, consider the Gaussian kernel, which is used to estimate the probability density function (PDF). Similarly, Woodcock and Benedetto (2009) use kernel densities for marginal transformation to normality. The kernel density (using a Gaussian kernel) of $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}'$ is

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \phi\left(\frac{x - Y_i}{h}\right),$$

where $h > 0$ is a bandwidth parameter, and $\phi(\cdot)$ represents the standard normal PDF. The CDF of $\boldsymbol{Y}$ may be approximated with

$$\hat{F}_h(y) = \int_{-\infty}^{y} \hat{f}_h(x)dx = \frac{1}{n} \sum_{i=1}^{n} \Phi\left(\frac{y - Y_i}{h}\right).$$

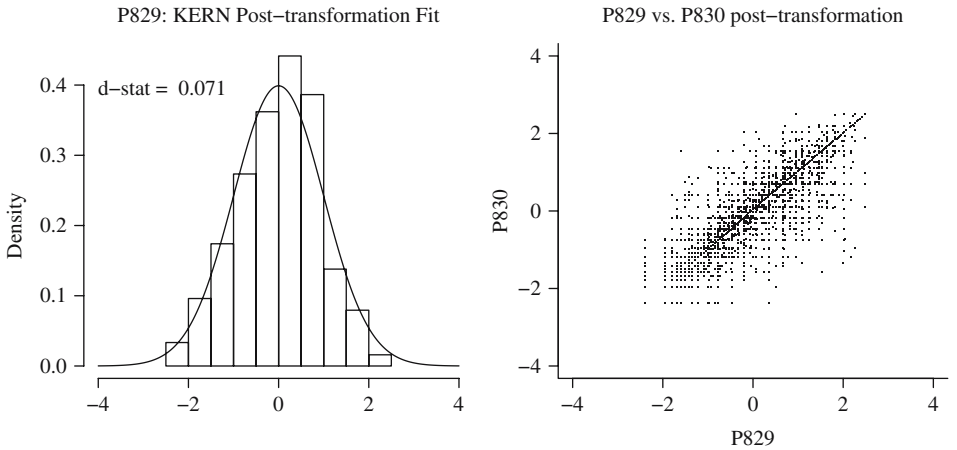

Fig. 3. Histogram of log(P829) (left) and scatter plot of log(P829) versus log(P830) (right) following kernel-density transformation. The left plot has the standard normal density curve overlaid

Therefore, the kernel-density transformation for $\boldsymbol{Y}$ is

$$T_2(y) = \Phi^{-1}(\hat{F}_h(y)), \tag{3}$$

and $T_2(Y_i)$ should appear to have been sampled from a standard normal distribution.

Figure 3 provides a histogram log(P829) and a scatter plot of log(P829) versus log(P830), all following the kernel-density (KERN) transformation. Clearly, the figure provides an instance where the kernel density offers a transformation to normality that is superior to that of the skew-normal family – the plots indicate that normality assumptions appear reasonable (in both the univariate and multivariate sense).

Selection of the bandwidth parameter, h, in kernel-density functions is a well-studied issue (Silverman 1986; Sheather and Jones 1991; Scott 2009). Selection algorithms often return small values of h for ARMS variables; such choices of h fail to adequately differentiate the KERN transformation from the EMP transformation described below. To avoid this issue, a bandwidth parameter of $h = 0.2$ is used whenever the KERN transformation is applied to ARMS data herein; this value offers adequate smoothing for the ARMS variables used.

3.2. *Transformation Via the Empirical Distribution*

The empirical distribution function of $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}'$ is now considered:

$$\tilde{F}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\},$$

where $\mathbf{1}\{A\}$ is the indicator of event A. We, however, focus on

$$\bar{F}(y) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{1}\{Y_i < y\} + \frac{1}{2}\mathbf{1}\{Y_i = y\}\right),$$

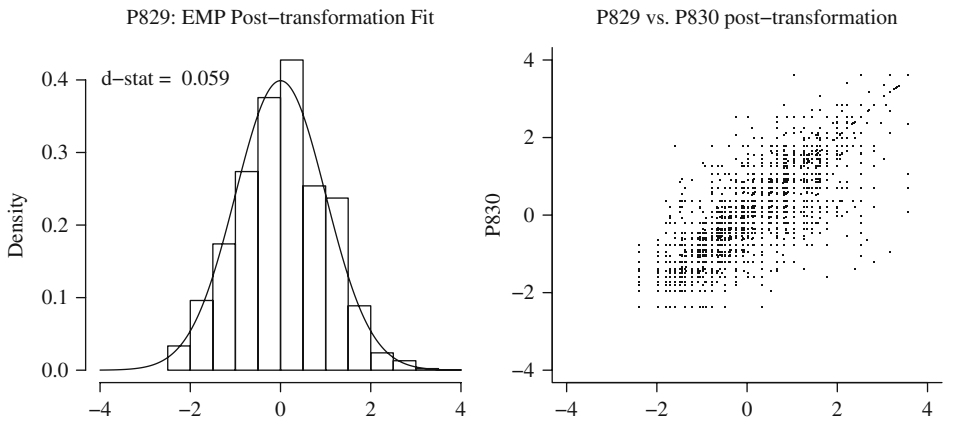

Fig. 4. Histogram of log(P829) (left) and scatter plot of log(P829) versus log(P830) (right) following empirical distribution transformation. The left plot has the standard normal density curve overlaid

since $\bar{F}(y) = lim_{h \to 0} \hat{F}_h(y)$ whenever $y \in \mathbb{R}$. Particularly, $\bar{F}(y)$ is preferable to $\tilde{F}(y)$ in cases where n is small or where $\boldsymbol{Y}$ contains repeating values (which is common for theoretically continuous portions of ARMS items). The empirical distribution (EMP) transformation for $\boldsymbol{Y}$ is

$$T_3(y) = \Phi^{-1}(\bar{F}(y)), \tag{4}$$

and $\tilde{\boldsymbol{Y}} = \{T_3(Y_1), \ldots, T_3(Y_n)\}$ should appear to have been sampled from a standard normal distribution. Note that $T_3(y)$ does not exist if $\bar{F}(y) = 0$ or 1. However, for all $y \in \boldsymbol{Y}$, $\bar{F}(y) \in (0, 1)$, meaning the observed values can be transformed via (4) without issue. Nonetheless, it is recommended to set $\bar{F}(y) = 1/(2n)$ if $y < min_i\{Y_i\}$, and $\bar{F}(y) = (2n - 1)/(2n)$ if $y > max_i\{Y_i\}$.

Figure 4 provides a histogram of P829 and a scatter plot of P829 versus P830, all following the EMP transformation. Repeating values of P829 prevent the EMP transformation from achieving exact normality. Regardless, the figure indicates that the EMP transformation is also clearly superior to the SN transformation in the circumstances illustrated here.

Since $\bar{F}(y) \xrightarrow{a.s.} F(y)$, the transformation in (4) is preferable when there is enough observed data to ensure that the empirical data provide a sufficient scope of the full distribution (including, most importantly, the tails).

3.3. Untransformation

Let $\boldsymbol{X}$ represent a transformed version of $\boldsymbol{Y}$ following application of one of the aforementioned schemes. Imputations will then be created for $\boldsymbol{X}$, resulting in $\hat{\boldsymbol{X}}$, an imputed version of the transformed data. However, the imputations must be "untransformed" (i.e., returned to their original scale). If a transformation of the type in (1) has been applied to $\boldsymbol{Y}$, the following inverse transformation may be applied to the imputed values:

$$T^{-1}(z) = F^{-1}(\Phi(z)), \quad \text{for } z \in (-\infty, \infty), \tag{5}$$

where $F^{-1}(u)$, for $u \in (0, 1)$, represents the inverse of the $F(y)$, for $y \in (-\infty, \infty)$. The CDF found using skew-normal assumptions, $F(y|\xi, \omega, \alpha)$, and the CDF found using

the kernel density, $\hat{F}_h(y)$, are both continuous and one-to-one mappings defined over $\mathbb{R}$. Thus, their respective inverses, $F^{-1}(u|\xi, \omega, \alpha)$ and $\hat{F}_h^{-1}(u)$, exist for $u \in (0, 1)$.

Let $\hat{x}$ represent an imputation for $\boldsymbol{X}$, which represents the transformed version of $\boldsymbol{Y}$. If the skew normal transformation seen in (2) was applied to this variable, $\hat{x}$ can be untransformed by calculating

$$\hat{y} = T_1^{-1}(\hat{x}) = F^{-1}(\Phi(\hat{x})|\hat{\xi}, \hat{\omega}, \hat{\alpha}),$$

and if kernel transformation seen in (3) was applied, the inversion requires the calculation of

$$\hat{y} = T_2^{-1}(\hat{x}) = \hat{F}_h^{-1}(\Phi(\hat{x})).$$

Computations involving the above two expressions (the latter, in particular) can be quite intensive.

The empirical CDF, $\bar{F}(y)$, is neither continuous nor one-to-one. Thus, its inverse, $\bar{F}^{-1}(y)$, does not exist, and (5) is not directly applicable. Hence, inversion of the empirical distribution transformation works as follows. Let $\boldsymbol{U} = \{U_1, \ldots, U_n\}$, where

$$U_i = \bar{F}(Y_i) \quad \text{for } i = 1, \ldots, n.$$

Note that the U_i should resemble uniform variates. For $\hat{x} \in (-\infty, \infty)$, let $u_x = \Phi(\hat{x})$, and after setting

$$i_x = \underset{i}{\operatorname{argmin}}|U_i - u_x|,$$

untransform imputations in variables requiring the empirical transformation by calculating

$$\hat{y} = T_3^{-1}(\hat{x}) = Y_{i_x}.$$

Inverting the empirical distribution in this manner ensures that any imputation of values in variables transformed using (4) will be sampled directly from observed values. Accordingly, an imputation method that utilizes the empirical method can be considered a "hot-deck" technique (Little 1988; Little and Rubin 2002). The EMP transformation is also advantageous due to its computational simplicity. However, the KERN transformation scheme is very demanding computationally (as it requires numeric integration).

4. Analysis of Imputed Data

The ISR algorithm of Robbins et al. (2013) is applied to the complete 2010 ARMS dataset using the imputation model described therein, where only the transformation technique is varied. For instance, five completed datasets were independently created (in the vein of multiple imputation) where the skew-normal (SN) transformation in (2) is used for all variables requiring transformation. This process is then repeated using the kernel-density (KERN) transformation in (3) and the empirical distribution (EMP) transformation in (4). Discussion is limited to the ARMS variables described in Table 1. The table also lists the number of positive and observed values (n_{obs}) and the number of missing values (n_{mis}) for each variable.

The full imputation model includes many additional variables beyond those listed in Table 1. Many of these variables contain missing values; the others are used as fully observed covariates. The respective transformation scheme is applied to all continuous or semi-continuous variables within the imputation algorithm. A list of variables included in the full model is given in Robbins et al. (2011).

To begin, analysis of marginal data characteristics of the variables (which contain missingness) in Table 1 is considered. Table 2 provides the unweighted sample mean ($\bar{x}$) and sample standard deviation (s) of the nonzero values, in addition to the between imputation variance (B) and upper bound (U_{MI}) and lower bound (L_{MI}) for the 95% confidence interval (as found using Rubin's combining rules for multiple imputation) for the population mean of each variable. The reported values of $\bar{x}$ and s represent the mean of their respective values when calculated in each of the five imputed datasets. Table 2 presents the results in "cells", where the top, middle, and bottom value in each cell is the respective estimate found using the SN, KERN and EMP transformations, respectively. Table 2 indicates that the choice of transformation method may result in differing values of means and variances. The discrepancies do not appear to be substantial, although it is noted that they are not explained by randomness in the imputations alone. Further, the lack of influence of the transformation type is likely due to relatively small missingness rates. It is also noted that other quantities (e.g., a 90% quantile) may be more heavily influenced by the transformation technique; however, the objective here is to present statistics that are of practical relevance.

To further examine marginal characteristics of imputations, discussion is now restricted to the variable P884. This variable is of particular interest because a large portion of positive and observed responses take on a single value (the specific value may not be disclosed here). This phenomenon is illustrated by the histogram of the positive and observed values of P884 which is provided in the left plot in Figure 5. The middle plot shows the positive and observed values of P884 following the EMP transformation. The plot provides visual evidence that the EMP transformation imposes "separation" between values that are frequently repeated and neighboring values. This separation ensures that there is a relatively high probability that an imputed value will equal the repeating value. For instance, 16.6% of positive and observed responses for P884 take on the frequently occurring value, and 9.3% of all EMP imputations take on that value (whereas 0% of SN and KERN imputations take on the value). The right plot in Figure 5 provides kernel-density plots of observed and imputed values (for each of the three transformation schemes), which further illustrates the need for a nonparametric transformation procedure.

There are alternative approaches for imputing P884. For instance, a three-level mixture model which includes two indicator variables (the first one indicating the occurrence of an observation equaling zero and the second indicating the observation taking on the frequently occurring value) may be more appropriate. However, such a procedure would have to enable the second indicator variable to have missing values (since it is not known whether or not the missing values of P884 take on the frequently occurring value). Therefore, the use of the marginal transformations (as opposed to higher-level mixture models) permits the convenience of a multivariate normal imputation model while producing high-quality results.

Table 2. Summary statistics for imputation of various ARMS variables. The top, middle and bottom values in each cell are calculated using SN, KERN and EMP imputations, respectively

	P828	P829	P830	P831	P832	P833	P834	P835	P884	P952
$\bar{x}$	36.42	47.37	46.45	42.41	19.84	23.21	23.39	21.51	46400	59400
	36.47	47.43	46.50	42.44	19.82	23.22	23.41	21.50	44600	59500
	36.47	47.47	46.55	42.48	19.80	23.19	23.40	21.48	45300	59300
s	0.0317	0.0406	0.0401	0.0376	0.0415	0.0492	0.0497	0.0452	6.69e6	5.70e5
	0.0316	0.0404	0.0399	0.0374	0.0415	0.0496	0.0502	0.0454	5.25e6	6.20e5
	0.0316	0.0405	0.0399	0.0375	0.0411	0.0492	0.0499	0.0450	5.44e6	5.90e5
B	6.55e-4	1.44e-3	3.43e-3	1.52e-3	2.05e-3	8.68e-3	1.33e-3	1.92e-3	6.09e5	6.71e4
	6.27e-4	2.10e-3	9.81e-4	1.57e-3	2.54e-3	2.74e-3	2.57e-3	1.53e-3	1.03e6	2.22e4
	1.74e-4	3.42e-3	1.62e-3	4.11e-3	1.31e-3	1.98e-3	1.39e-3	2.12e-3	8.92e5	3.01e4
L_{MI}	36.07	46.97	46.04	42.04	19.42	22.77	22.95	21.08	40900	57800
	36.11	47.03	46.10	42.05	19.40	22.77	22.96	21.07	39600	58000
	36.12	47.06	46.14	42.07	19.40	22.74	22.95	21.05	40200	57700
U_{MI}	36.77	47.78	46.87	42.80	20.25	23.65	23.84	21.93	51900	60900
	36.82	47.84	46.90	42.83	20.23	23.68	23.87	21.92	49700	61100
	36.82	47.89	46.95	42.88	20.21	23.63	23.84	21.91	50300	60800

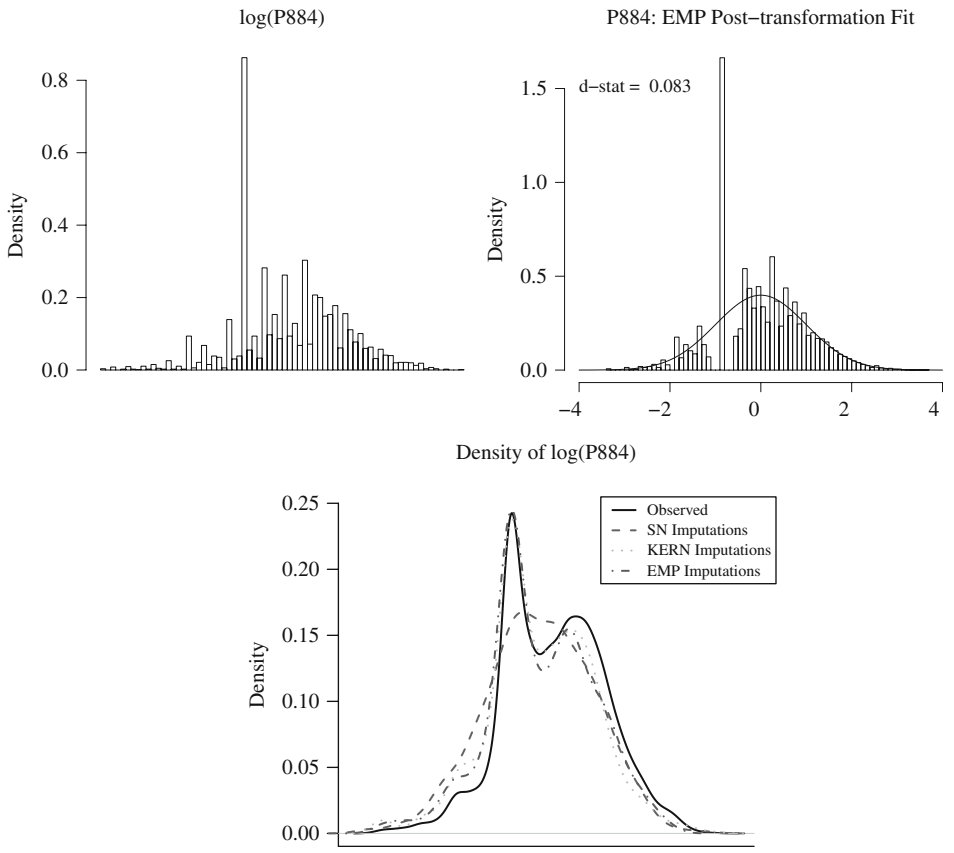

Fig. 5. Histogram of positive and observed values of log (P884) before (top left) and after (top right) an empirical transformation and densities for observed and imputed values of P884 (bottom)

To monitor the multivariate influence of imputations sampled using the various transformation schemes, consider scatter plots. Figure 6 provides scatters plots of log(P829) versus log(P830) for each of the three transformations where pairwise positive and observed pairs are marked with an '×' and imputed pairs are marked with a '+'. Lines of best fit for observed and imputed pairs are also included. Plots are given on the log scale in order to emphasize the differences between methods. The plots appear to indicate that bivariate extremes are underimputed, which may (partially) be a result of imputed values tending to be smaller than observed values for both variables in the plots. This phenomenon is to be expected for the labor variables; data indicate that "hobby" farmers, who are less likely to work on-farm full time, are more likely to refuse response for labor items. Regardless, the EMP transformation is clearly the most likely to preserve the underlying bivariate structure.

To further gauge the multivariate quality of the imputations, consider an econometric model motivated by the following. Farm operators often pursue off-farm sources of income; the on- and off-farm labor decisions of farmers have been well scrutinized in the economic literature. Economic theory suggests that the amount of time a farm operator (and the operator's spouse) choose to work on the farm is heavily influenced by factors

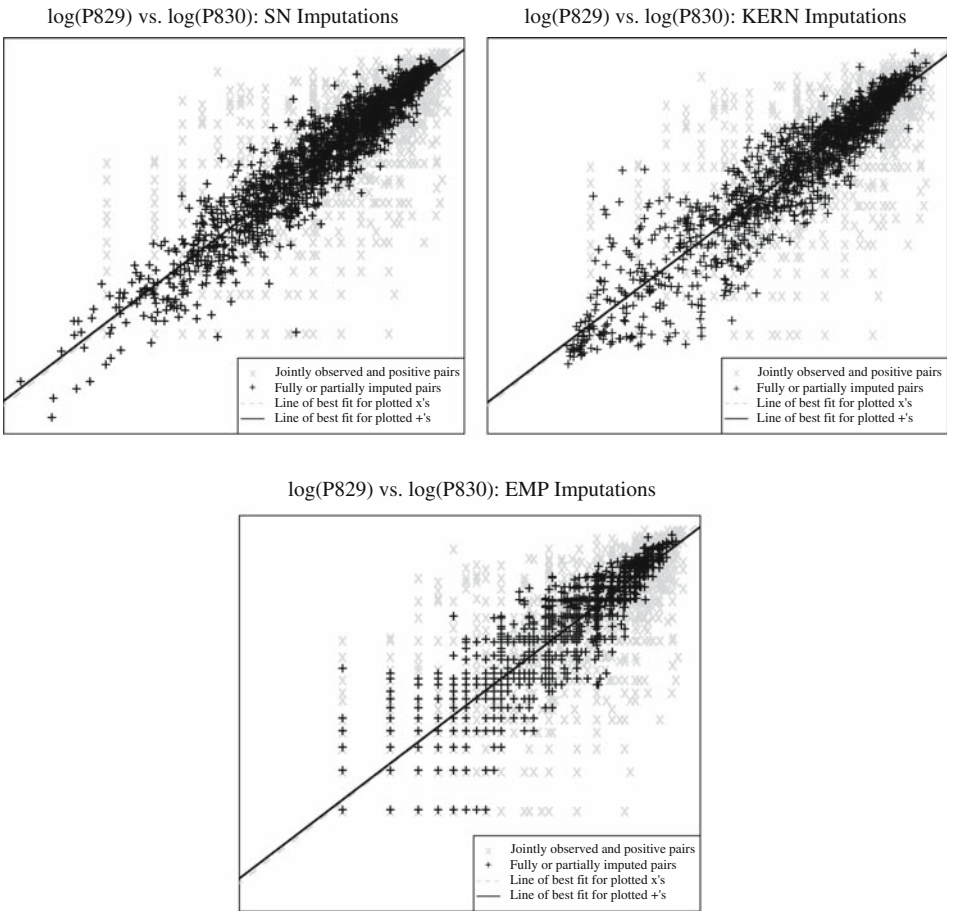

Fig. 6. Scatter plots of imputed and observed pairs of log (P829) and log (P830) for the various transformation schemes

such as the hours worked off farm by the operator and spouse, an on-farm wagerate, off-farm wage rate, the operator's age and level of education, and so forth. Econometric models investigating this concept have been considered by Huffman (1980); Sumner (1982); Huffman and Lange (1989); Mishra and Holthausen (2002) and Kwon et al. (2006) among many others. Here, consider the following linear model:

$$\text{OPHR} = \beta_0 + \beta_1 \text{OPOFFHR} + \beta_2 \text{OFFRATE} + \beta_3 \text{P1242} + \beta_4 \boldsymbol{Z} + \varepsilon, \qquad (6)$$

where $\boldsymbol{Z}$ represents a set of additional categorical covariates and ε is a mean zero error term. In the above, OPHR is the number of on-farm hours worked weekly by the farm operator (calculated as the average of P828, P829, P830 and P831). Likewise, OPOFFHR is the number of hours worked off-farm by the farm operator. OFFRATE is calculated as P952/(OPOFFHR + SPOFFHR) where SPOFFHR is the number of hours worked off farm by the operator's spouse. That is, OFFRATE represents the combined off-farm wage rate for the operator and spouse. Further, P1242 is the operator's age. Estimated values of coefficients are found using least squares while isolating to units that report nonzero values

Table 3. Summary information for the econometric model. The top, middle and bottom values in each cell are calculated using SN, KERN and EMP imputations, respectively

	β_0	β_1	β_2	β_3
Coefficient	2237	−0.432	−8.02e-3	−8.96
	2224	−0.427	−8.05e-3	−9.08
	2240	−0.433	−7.06e-3	−8.90
se(*Coef*.)	9733	2.90e-4	1.96e-6	1.575
	9688	2.89e-4	1.94e-6	1.569
	9642	2.87e-4	1.90e-6	1.559
B	775.5	2.08e-5	7.51e-7	0.1794
	843.7	3.14e-5	7.09e-7	0.2711
	876.6	1.38e-5	8.45e-7	0.1106
L_{MI}	2034	−0.467	−1.14e-2	−11.60
	2039	−0.462	−1.14e-2	−11.80
	2037	−0.468	−1.11e-2	−11.46
U_{MI}	2440	−0.397	−4.60e-3	−6.329
	2445	−0.391	−4.68e-3	−6.357
	2443	−0.400	−4.12e-3	−6.349

for all pertinent variables. A model similar to (6) which involves the hours worked on farm by the spouse was also considered throughout this study, but the findings are redundant and thereby omitted.

Table 3 provides results for these two models. The format of this table is similar to that of Table 2, as are the findings: The choice of transformation method may have a noticeable (but in this case not substantial) impact on the estimations found using econometric modeling.

5. A Simulation Study

This section presents simulation analyses which evaluate the efficacy of the proposed transformation techniques. Ideally, all assessments would be performed using real data, since synthetic data are not guaranteed to adequately mimic the complex structures encountered in practice – the motivation behind the proposed techniques is to capture such structures. Accordingly, when possible, evaluations are performed with observed ARMS data; in circumstances where such analyses do not offer sufficiently clear conclusions; a small-scale study using entirely synthetic data is used to inform the discussion.

5.1. Simulations Involving ARMS Data

A preferable technique for simulation involving real data would be to draw a sample of respondents from the observed units while treating the full dataset as a population from which population parameters can be ascertained; implementations of this scheme are seen

in Reiter (2005) and Manrique-Vallier and Reiter (2014). However, there are not enough available data for this approach to be feasible within the ARMS. ARMS data are high dimensional; nonetheless, the effective sample size (the number of positive values) can be quite small for some variables. Instead, a jackknife-type study is executed here.

As setup, a completed ARMS dataset is created using the imputation scheme outlined in Robbins et al. (2011). Specifically, the full-scale ISR algorithm and model are used in conjunction with various transformation schemes. It is not feasible to use complete cases only since there are an insufficient number of complete cases. This single completed dataset is used to create all of the benchmark values required within the simulation study. Next, missingness is randomly imposed in eight of the ARMS variables according to a probabilistic model. Imputations are then created for these newly missing values and the values of desired metrics as found using the imputed data are compared to values found using the original benchmark dataset. It is worth noting that the rate of missingness that is imposed will vastly exceed the original rate of missingness in ARMS data. The eight variables in which holes are poked are marked in Table 1 with an asterisk; some of these variables originally contained missingness, whereas others did not.

In addition to the eight variables in which missingness is imposed, there are 18 additional variables used as covariates within the imputation model for ISR. The imposed missingness is completely at random (MCAR, in the terminology of Little and Rubin 2002). Specifically, any positive value is imposed as missing with a probability of 0.5, and the occurrence of imposed missingness is independent across all values. Since the imposed rate of missingness is far higher than the missingness rate in the original dataset, the influence of imputations within the benchmark study should be filtered out. The performance of ISR with density transformations has been analyzed in great detail under other missingness mechanisms (e.g., MAR and NMAR – for details, see the supplemental material of Robbins et al. 2013). Analyses under MAR and NMAR are not expected to yield information regarding the influence of the transformation type beyond what is learned under MCAR missingness; for brevity, only MCAR is considered in these ARMS-based simulations. Since ISR is iterative (as it is a form of Markov chain Monte Carlo), each completed dataset is sampled using a burn-in period of 200 iterations.

The goal is to assess the potential for bias (in any point and interval estimates calculated from the ARMS data) caused by the choice of transformation method. The performance of the methodology is measured in terms of the relative change of a metric post imputation. Missingness is randomly imposed in the completed benchmark dataset 100 different times. Each time missingness is imposed, imputations are independently created five times (in the vein of multiple imputation) for each method. The methods used are as follows.

1. SN – The skew-normal transformation of (2) is used for all variables.
2. KERN – The kernel-density transformation of (3) is used for all variables.
3. EMP – The empirical distribution transformation of (4) is used for all variables.
4. EMPABB – EMP with an approximate Bayesian bootstrap.

The transformation schemes discussed in Section 3 will result in imputations that understate variability due to the fact each transformation scheme requires that any variable's CDF, $F(x)$, be treated as known despite the fact that $F(x)$ is, in fact, estimated.

To address this issue, Woodcock and Benedetto (2009) suggest an approximate Bayesian bootstrap (ABB), where $F(x)$ is estimated using a bootstrapped pool of observations as opposed to the actual pool of observations. Here, ABB is used together with the EMP method, resulting in EMPABB as above.

Let $\boldsymbol{\chi}$ denote the benchmark dataset, and let $\boldsymbol{\chi}_k^{[d]}$ denote the d^{th} completed dataset ($d = 1, \ldots, 5$) as imputed for the k^{th} artificially incomplete dataset (where $k = 1, \ldots, 100$). Finally, let $\theta(\cdot)$ denote a metric of interest (where the argument represents the dataset used to compute the metric). The percent change in the metric is computed via

$$\Delta\theta(k) = 100\left(\frac{\bar{\theta}_k - \theta(\boldsymbol{\chi})}{\theta(\boldsymbol{\chi})}\right),$$

where $\bar{\theta}_k = \sum_{d=1}^{5} \theta(\boldsymbol{\chi}_k^{[d]})/5$. Results are presented in the form of box plots of the 100 values of $\Delta\theta(k)$.

Metrics tracked in this simulation study include the sample mean and standard error of the sample mean as calculated over the *nonzero* values of each variable in which missingness is imposed in addition to the regression coefficients in (6) and their respective standard errors. Note that the standard error of a sample mean equals the sample standard deviation times a constant (i.e., $n_{\text{obs}}^{-1/2}$). Covariances were also monitored but yielded results that mimic those of the regression coefficients (accordingly, those results are omitted from the discussion). Confidence intervals for the sample means and regression coefficients can be calculated using Rubin's combining rules for MI, although the details are omitted here.

Findings are shown in Figure 7 for P784, P829, β_1 and β_2. The results indicate that for certain variables (e.g., P829) whose marginal distributions cannot be modeled with an appropriate parametric density, biases in basic marginal characteristics may be induced if one does not utilize a nonparametric transformation. Further, the nonparametric transformations result in imputations that appear to adequately preserve the quantities studied here (though there may be evidence of a moderate decrease in the variance of P784 caused by the nonparametric methodology). Likewise, there does not appear to be an advantage to using the EMPABB method in place of the EMP method.

Finally, since the empirical distribution transformation is designed to handle repeating values, it has the potential to be applied to variables that are binary or ordinal (though not strictly categorical with more than two categories). However, such efficacy of the transformation for such a purpose has not been thoroughly investigated.

Of interest is P784; this variable was included in this study since it has a particularly low number of positive and observed values (151 in the true dataset and thereby approximately 75 prior to imputation within the simulation study – see Table 1). Parametric and nonparametric transformations (when the former are well fit) are expected to perform equivalently on large samples (wherein sufficient data are available to adequately approximate the CDF under all transformation types); discrepancies between transformations are anticipated to be most visible when there are few observations available. To that end, it is noted that the SN transformation results in a substantially wider confidence interval for the mean of the nonzero observations of P784 (approximately three

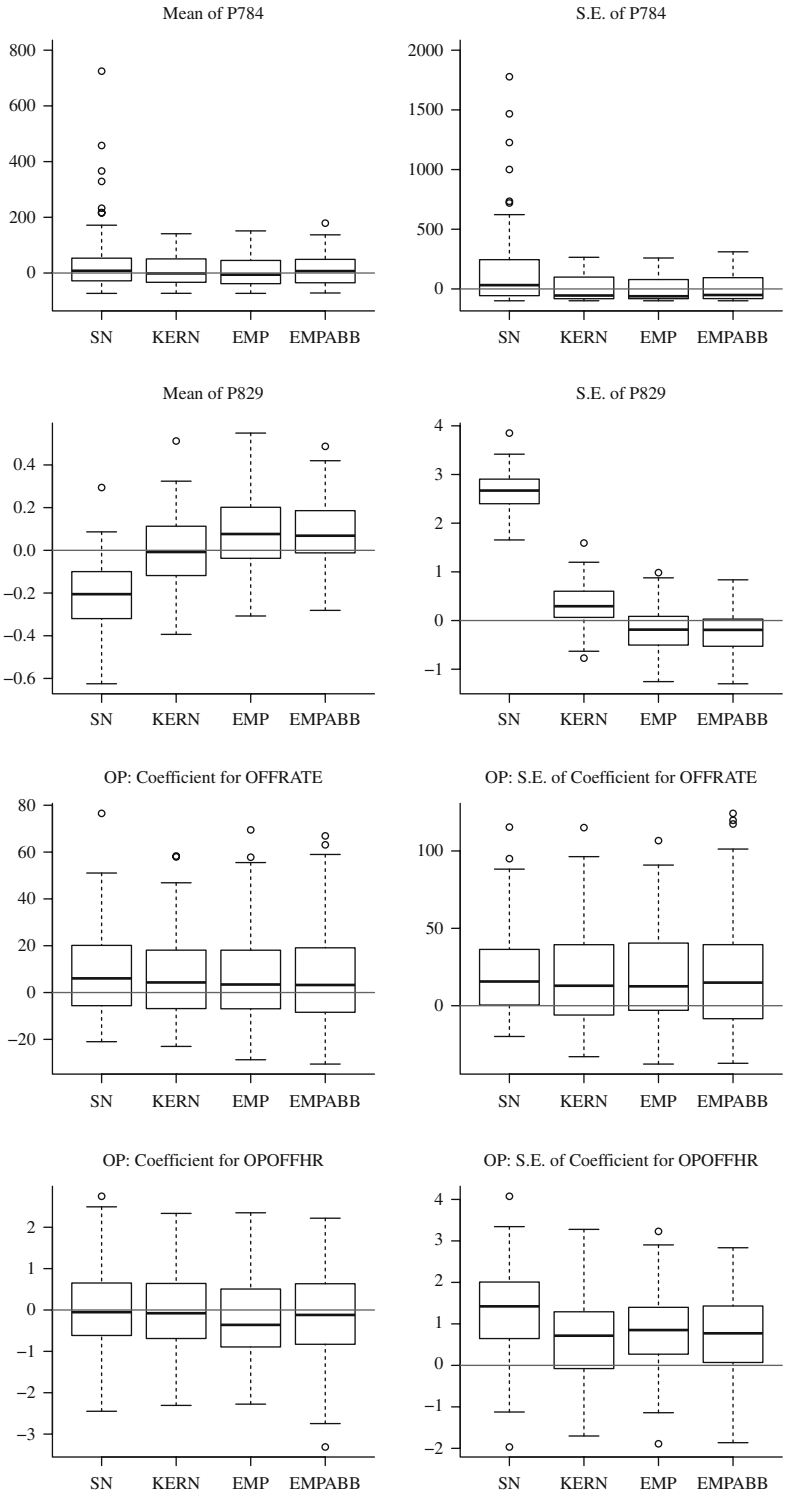

Fig. 7. Box plots of % change in various metrics

to four times wider on average than the KERN and EMP transformations) within the simulations used to generated Figure 7. Since the SN model seems appropriate for this variable (the KS test yields a p-value of 0.801 when a skew-normal distribution is assumed), since it seems unreasonable to assume that 75 observations can sufficiently quantify a CDF, and since Figure 7 implies that the nonparametric transformations may decrease the variance of this variable, it is suggested that the SN transformation is more appropriate than the nonparametric transformations for P784.

Ideally, comparisons to predictive mean matching (PMM, Little 1988) could have been presented in this study. PMM is a popular technique that builds a predictive model for imputations through regression, and then samples imputations from observed data – making it similar to (and useful in the same settings as) the methods presented here. However, direct comparisons to PMM within the simulations above (wherein such comparisons would be most useful due to the unknown distributional structure of ARMSdata) cannot be made here due to computational constraints. For instance, one iteration of ISR takes 1.15 seconds, and one iteration of MICE with PMM takes 15 minutes when run on the group of variables used above. These computations are executed on a 64-bit Windows machine with a 3.3 GHz processor and 8.0 GB of RAM.

To summarize, the above study helps to verify the efficacy of the proposed methodology on real data, but it has some notable shortcomings. For instance, it is desirable to investigate the comparative performance of the proposed techniques against other methods such as PMM, and to present results for a variety of missingness structures. Many of these shortcomings are the consequence of computational issues. Furthermore, the above simulations leave unanswered the question as to whether or not a parametric transformation is preferable in settings involving small samples. A small-scale study involving fully synthetic data is thus presented below.

5.2. *Simulations Involving Synthetic Data*

The small scale of the following simulation study (only two variables are used for various sample sizes) makes it computationally feasible to consider a variety of methods and missingness mechanisms. Specifically, the four transformation techniques mentioned above (SN, KERN, EMP, and EMPABB) are used in conjunction with ISR. As needed, skew-normal MLEs are used, and the kernel bandwidth parameter is estimated via the method of Sheather and Jones (1991). Further, PMM is considered (while used in conjunction with `mice`) as well as IRMI (Templ et al. 2011); no transformation is used when these methods are applied.

Data are generated as follows. Let $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ represent a random sample from a skew-normal distribution with parameters $\xi = 4$, $\omega = 2$ and $\alpha = -2$. Additionally, let $\tilde{\boldsymbol{X}} = \{\tilde{X}_1, \ldots, \tilde{X}_n\}$ represent the version of $\boldsymbol{X}$ that has been transformed in accordance with (2) while using the true parameter values, and define $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$, where $Y_i = 1 + 0.5\tilde{X}_i + \varepsilon_i$ for $i = 1, \ldots, n$, and where $\boldsymbol{\varepsilon} = \{\varepsilon_1, \ldots, \varepsilon_n\}$ is a random sample of length n from a standard normal distribution.

Missingness is imposed in the values of $\boldsymbol{X}$ through the following mechanisms. Under MCAR missingness, each observation of $\boldsymbol{X}$ is missing with probability 0.5. For MAR missingness, X_i is missing with a probability equal to $1/(1 + exp(-\tilde{Y}_i))$, where $\tilde{Y}_i$

represents a standardized version of Y_i. NMAR missingness was also considered, but the results are excluded for brevity since they provided no additional information regarding the choice of transformation scheme beyond what is learned from the other mechanisms. Imputations in $\boldsymbol{X}$ are generated via the techniques mentioned above; the elements of $\boldsymbol{Y}$ are not transformed at any point. Further, $m = 5$ imputed datasets are created, and no burn-in period is necessary since missingness is restricted to one variable. MI point and interval estimates are generated for a handful of parameters, and the entire process is replicated independently 1,000 times for various values of n.

For a given imputation method, missingness mechanism, and value of n, let $\hat{\theta}_k$ denote the MI point estimate of a generic parameter θ calculated following the k^{th} replication ($k = 1, \ldots, 1,000$). The percent bias in the multiple imputation estimate of θ is approximated by calculating $\bar{\Delta}\theta = 100\sum_{k=1}^{1,000}[(\hat{\theta}_k/\theta) - 1]/1,000$. Similarly, the sequence of 1,000 values of $\hat{\theta}_k$ can be tested to see if the percent bias is statistically nonzero. Further, the empirical coverage of the MI interval estimate of θ is calculated via the portion of the 1,000 replications in which the true value of θ is contained within its 95% confidence interval.

First, we consider the basic univariate parameters $\mu = \mathrm{E}[X_1]$ and $\sigma^2 = \mathrm{Var}(X_1)$; results are given in Table 4. All transformation methods offer strong performance in terms of bias and coverage for these parameters, as does the PMM procedure. However, the IRMI procedure shows some evidence of bias and observes poor coverage for these simple quantities. It appears that all methods induce a small amount of bias (which mostly disappears with increasing n) under MAR missingness; the fact that this bias tends to be negative is a consequence of the form of the function that generates the MAR missingness. Moreover, the results imply that the use of the approximate Bayesian bootstrap does not improve the results. Finally (and most importantly), all transformation schemes appear to offer equivalent performance.

In order to provide parallels to the log-skew-normal distributions that positive portions of ARMS data observe, we also study summary statistics of the transformed variable $U_i = \exp(X_i)$. Specifically, we use multiple imputation to develop point and interval estimates of $\gamma = \mathrm{E}[U_1]$ and $\nu^2 = \mathrm{Var}(U_1)$ by applying Rubin's combining rules to the sequence $\{\hat{U}_1, \ldots, \hat{U}_n\}$, where $\hat{U}_i$ represents a version of U_i that contains imputations of missingvalues. The ability of an imputation algorithm to preserve such quantities is a strong indication that the distribution of the imputed data matches that of the actual data had they been fully observed (since γ and ν^2 follow from the specificform of the MGF of X_1). Results for these two quantities are shown in Table 5. The table indicates that IRMI imputations provide biased estimates of γ and ν^2 under all missingness mechanisms. This observation is not surprising, since IRMI does not take steps to ensure that the full distributional structure is captured in the imputation process. Although all methods are more imprecise in their estimation of γ and ν^2 than of μ and σ^2, Tables 4 and 5 both yield the same conclusions regarding the comparative performance of the techniques.

In summary, the key findings of the simulation studies presented in this subsection are that all methods involving transformation are comparable to PMM and that the choice of transformation technique does not have a significant influence on bias or coverage probabilities. The latter finding is noteworthy because the SN method, which is ideally suited to this setting, shows no gains over the nonparametric methods, whereas the

Table 4. Empirical bias and coverage probabilities (the latter are in parentheses) of the point estimates and 95% confidence intervals (found using MI) of two parameters involving the synthetic random variable X. An asterisk indicates that the bias is statistically nonzero at the 0.01 significance level

% bias and % coverage for $\mu = \mathrm{E}[X]$							
	n	SN	KERN	EMP	EMPABB	PMM	IRMI
MCAR	50	0.51 (91.3)	0.45 (91.9)	−0.16 (91.2)	−0.08 (90.6)	−0.34 (90.5)	2.22* (68.7)
	100	0.27 (93.2)	0.12 (93.7)	0.25 (93.5)	0.71* (92.4)	0.18 (93.2)	1.95* (70.4)
	250	−0.15 (92.3)	0.15 (93.5)	0.00 (94.6)	0.26 (94.0)	−0.15 (95.1)	1.66* (67.7)
	500	0.02 (93.7)	−0.16 (95.3)	−0.16 (93.3)	−0.03 (94.7)	0.12 (94.0)	1.54* (68.9)
	1,000	−0.02 (93.7)	0.01 (94.1)	−0.10 (94.3)	0.02 (94.5)	0.01 (95.2)	1.49* (62.0)
	2,500	0.09 (94.7)	0.00 (94.5)	0.01 (94.0)	−0.04 (94.4)	0.09 (93.4)	1.50* (54.2)
MAR	50	−2.37* (90.1)	−3.42* (89.4)	−3.13* (90.0)	−2.12* (90.3)	−2.19* (90.0)	−0.65 (68.3)
	100	−1.31* (92.0)	−1.40* (89.4)	−2.02* (89.2)	−1.25* (91.8)	−2.33* (91.7)	0.21 (66.2)
	250	−0.95* (91.2)	−0.59* (93.5)	−0.50* (92.6)	−0.55* (91.6)	−0.72* (91.5)	0.96* (67.2)
	500	−0.41* (91.5)	−0.28 (92.3)	−0.35* (91.7)	−0.38* (91.9)	−0.34* (91.4)	0.94* (63.5)
	1,000	−0.31* (91.6)	−0.26* (93.5)	−0.39* (93.2)	−0.19 (92.9)	−0.36* (92.4)	1.27* (62.3)
	2,500	−0.09 (92.8)	−0.11 (93.0)	0.04 (92.1)	0.05 (92.0)	−0.19* (91.3)	1.29* (55.6)
% bias and % coverage for $\sigma^2 = \mathrm{Var}(X)$							
	n	SN	KERN	EMP	EMPABB	PMM	IRMI
MCAR	50	−5.85* (81.9)	−6.60* (81.6)	−5.74* (82.1)	−7.50* (78.7)	−4.78* (81.2)	−40.4* (35.4)
	100	−2.95* (85.0)	−3.26* (86.2)	−3.07* (86.6)	−5.20* (84.6)	−3.29* (85.2)	−40.3* (18.8)
	250	−1.44* (87.5)	−1.24* (88.8)	−1.19* (88.3)	−2.12* (86.6)	−1.21* (91.0)	−40.0* (2.90)
	500	0.01 (91.3)	−0.45 (90.7)	−0.12 (91.6)	−0.70 (90.1)	−0.94* (89.2)	−39.9* (0.20)
	1,000	−0.20 (90.3)	0.01 (91.6)	−0.49 (92.5)	−0.27 (90.6)	−0.28 (91.8)	−40.5* (0.00)
	2,500	−0.22 (91.0)	−0.05 (91.9)	−0.11 (91.3)	0.07 (90.9)	−0.21 (92.6)	−40.2* (0.00)
MAR	50	−5.96* (83.0)	−6.57* (82.3)	−7.05* (80.4)	−9.31* (75.2)	−7.95* (78.7)	−38.8* (38.8)
	100	−3.63* (86.9)	−3.93* (85.5)	−4.26* (85.3)	−4.97* (83.0)	−4.25* (85.3)	−38.7* (24.2)
	250	−2.10* (88.0)	−1.83* (87.6)	−1.48* (88.0)	−1.70* (87.1)	−0.90 (88.6)	−38.2* (6.50)
	500	−1.39* (87.1)	−1.02* (87.4)	−1.50* (88.3)	−1.42* (86.7)	−0.69 (86.5)	−37.7* (0.40)
	1,000	−0.37 (90.0)	−0.68* (88.4)	−0.36 (87.5)	−0.69* (86.8)	−0.91* (85.5)	−37.0* (0.00)
	2,500	−0.24 (90.0)	−0.45* (90.2)	−0.37 (89.1)	−0.50* (88.3)	−0.17 (87.7)	−37.2* (0.00)

Table 5. Empirical bias and coverage probabilities (in parentheses) of the point estimates and 95% confidence intervals of two parameters involving the synthetic random variable $U = \exp(X)$. An asterisk indicates that the bias is statistically nonzero at the 0.01 significance level

% bias and % coverage for $\gamma = \mathrm{E}[U]$ where $U = \exp(X)$							
	n	SN	KERN	EMP	EMPABB	PMM	IRMI
MCAR	50	−0.34 (84.2)	−0.44 (82.9)	−1.67 (82.5)	−2.97* (80.6)	−0.11 (81.8)	−18.5* (53.2)
	100	0.30 (86.5)	−0.76 (86.8)	−0.38 (88.1)	−0.30 (86.9)	−0.29 (86.6)	−19.5* (48.8)
	250	−0.89 (89.0)	0.00 (90.0)	−0.32 (89.9)	−0.36 (88.0)	−0.92 (88.8)	−20.3* (29.8)
	500	0.25 (90.8)	−0.39 (92.1)	0.05 (92.2)	−0.41 (89.0)	0.03 (92.3)	−20.9* (12.0)
	1,000	0.01 (91.8)	0.26 (91.0)	0.14 (93.1)	−0.02 (92.0)	0.01 (93.2)	−21.3* (1.40)
	2,500	0.11 (92.1)	0.01 (91.8)	0.18 (93.6)	−0.09 (92.5)	0.14 (93.2)	−21.1* (0.00)
MAR	50	−7.94* (73.1)	−10.1* (73.1)	−8.67* (70.6)	−8.65* (71.7)	−8.88* (72.2)	−25.3* (40.6)
	100	−4.88* (79.8)	−5.23* (77.6)	−6.99* (74.3)	−5.41* (75.1)	−7.16* (75.4)	−26.0* (28.7)
	250	−3.89* (78.9)	−3.11* (81.3)	−1.86* (80.8)	−2.66* (79.9)	−1.94* (81.7)	−25.4* (14.5)
	500	−1.81* (81.3)	−1.76* (83.3)	−1.53* (80.4)	−2.51* (81.8)	−1.38* (83.0)	−25.8* (4.70)
	1,000	−1.39* (84.3)	−1.12* (83.7)	−1.52* (84.1)	−1.02* (82.2)	−1.88* (82.1)	−25.1* (0.70)
	2,500	−0.36 (86.4)	−0.76* (84.7)	−0.30 (84.9)	−0.45 (83.2)	−0.78* (84.8)	−25.3* (0.00)
% bias and % coverage for $\nu^2 = \mathrm{Var}(U)$ where $U = \exp(X)$							
	n	SN	KERN	EMP	EMPABB	PMM	IRMI
MCAR	50	−11.1* (43.2)	−5.13 (43.8)	−8.34 (44.2)	−17.8* (34.1)	1.06 (43.9)	−45.6* (4.50)
	100	−2.55 (52.2)	−3.71 (48.3)	−3.12 (51.5)	−7.06 (43.1)	−0.97 (50.4)	−44.1* (1.80)
	250	−5.22 (57.5)	1.46 (58.0)	−1.35 (56.1)	−8.18* (52.3)	−4.34 (56.6)	−43.6* (0.80)
	500	−0.82 (63.3)	−0.73 (64.0)	0.68 (62.4)	−5.05* (55.5)	2.40 (62.2)	−45.2* (0.30)
	1,000	2.74 (64.5)	2.11 (66.9)	1.30 (66.7)	−1.57 (63.5)	−0.36 (67.8)	−46.0* (0.10)
	2,500	−0.74 (72.1)	−0.54 (71.1)	−0.08 (72.0)	−0.18 (65.4)	0.01 (71.0)	−44.9* (0.00)
MAR	50	−26.1* (34.5)	−29.3* (33.9)	−22.1* (32.8)	−29.1* (25.0)	−33.4* (33.2)	−53.5* (5.40)
	100	−19.4* (41.9)	−20.0* (39.5)	−23.4* (38.4)	−17.9* (35.0)	−20.2* (39.4)	−63.1* (1.60)
	250	−15.1* (45.0)	−12.5* (45.0)	−6.10 (45.6)	−16.8* (43.9)	−7.43 (49.8)	−63.5* (0.10)
	500	−6.26 (49.7)	−9.49* (52.4)	−2.86 (52.0)	−15.6* (45.9)	−4.77 (51.1)	−65.3* (0.10)
	1,000	−8.89* (53.3)	4.87 (52.3)	−6.85* (54.3)	−6.41* (49.2)	−10.2* (50.5)	−64.7* (0.00)
	2,500	0.06 (56.4)	−4.06* (58.2)	−1.50 (57.9)	−5.23* (51.6)	−4.17 (57.4)	−65.3* (0.00)

nonparametric methods will certainly provide higher efficacy in settings where the skew-normal assumption is violated. Figure 7 shows that the nonparametric methods yield a decrease in the variance of P784, and Tables 4 and 5 implicate that all methods may have decreased variability in items with small sample sizes. This decrease is not seen by the SN method in Figure 7, perhaps because the skew-normal distribution does not adequately capture the tails of the distribution of P784 (which also helps to explain the outlying values for this variable and transformation method in Figure 7).

6. Comments

Nonparametric transformation of survey data prior to imputation provides a straightforward manner through which unique marginal data characteristics can be preserved throughout the imputation process – such transformations are also shown to maintain multivariate aspects. The empirical transformation described above has the added advantage that imputations are drawn from observed data, which makes a method that utilizes it a nearest neighbor-type technique, and which also increases the probability that complex underlying data structures (that are common in establishment surveys) are maintained. Further, the empirical transformation is advantageous due to its computational simplicity.

The evaluations presented in this article did not unveil circumstances in which a transformation based upon a parametric model (i.e., the skew-normal distribution) is clearly preferable to the nonparametric methods. Further, no settings were found in which a transformation based upon a kernel density outperformed the transformation based upon an empirical distribution – the latter is more computationally efficient. In light of the above, the recommendation is that in practical circumstances the empirical distribution transformation be used when possible (however, further evaluations beyond those presented here may be needed to support this conclusion). With any transformation method, the practitioner should always investigate the validity of the posttransformation multivariate model (a joint normal distribution was used here) prior to generating imputations.

As an additional comment, it is noted that the nonparametric methods are applied here while exclusively using ISR (Robbins et al. 2013). ISR has the restriction that variables with missing values be sampled from continuous distributions. However, the nonparametric transformations are applicable in conjunction with any imputation technique which applies normality assumptions to continuous variables. For instance, these transformations could be employed with IVEware (Raghunathan et al. 2002) or MICE (Van Buuren and Oudshoorn 1999), which include capabilities for imputation of categorical variables.

Furthermore, it is also possible to use the methods discussed here for simulation of fully or partially synthetic datasets for the purposes of data confidentiality (Rubin 1993; Reiter 2002; Raghunathan et al. 2003). Woodcock and Benedetto (2009) use a kernel-density transformationfor this purpose, and it is noted that the empirical transformation has such utility if it is acceptable for synthetic values to be sampled from the observed data.

Finally, we note that one may use the EMP transformation technique for imputation of ordinal or binary variables (though not for categorical variables with more than two categories) since the method samples imputations from the set of observed values. However, the performance of the EMP method for this purpose has not yet been examined thoroughly.

7. References

Azzalini, A. 1985. "A Class of Distributions Which Includes the Normal Ones." *Scandinavian Journal of Statistics* 12: 171–178.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. "Maximum Likelihood From Incomplete Data via the EM Algorithm (with discussion)." *Journal of the Royal Statistical Society* Series B 39: 1–38.

Fay, R.E. 1996. "Alternative Paradigms for the Analysis of Imputed Survey Data." *Journal of the American Statistical Association* 91: 490–498. DOI: http://dx.doi.org/10.1080/01621459.1996.10476909.

Huffman, W.E. 1980. "Farm and Off-Farm Work Decisions: The Role of Human Capital." *Review of Economics and Statistics* 62: 14–23.

Huffman, W.E., and M.D. Lange. 1989. "Off-Farm Work Decisions of Husbands and Wives: Joint Decision Making." *The Review of Economics and Statistics* 71: 471–480. DOI: http://dx.doi.org/10.2307/1926904.

Javaras, K.N., and D.A. van Dyk. 2003. "Multiple Imputation for Incomplete Data with Semicontinuous Variables." *Journal of the American Statistical Association* 98: 703–715. DOI: http://dx.doi.org/10.1198/016214503000000611.

Kim, J.K., J.M. Brick, W.A. Fuller, and G. Kalton. 2006. "On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling." *Journal of the Royal Statistical Society* Series B 68: 509–521. DOI: http://dx.doi.org/10.1111/j.1467-9868.2006.00546.x.

Kott, P.S. 1995. *A Paradox of Multiple Imputation*. Tech. rep., National Agricultural Statistics Service, Fairfax, VA. Presented at the Joint Statistical Meetings, August 1995, Orlando, FL

Kott, P.S., and T. Chang. 2010. "Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse." *Journal of the American Statistical Association* 105: 1265–1275. DOI: http://dx.doi.org/10.1198/jasa.2010.tm09016.

Kwon, C.-W., P. Orazem, and D.M. Otto. 2006. "Off-Farm Labor Supply Responses to Permanent and Transitory Farm Income." *Agricultural Economics* 34: 59–67. DOI: http://dx.doi.org/10.1111/j.1574-0862.2006.00103.x.

Little, R.J.A. 1988. "Missing-Data Adjustments in Large Surveys." *Journal of Business & Economic Statistics* 6: 287–296. DOI: http://dx.doi.org/10.1080/07350015.1988.10509663.

Little, R.J.A., and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley & Sons.

Manrique-Vallier, D., and J.P. Reiter. 2014. "Bayesian Multiple Imputation for Large-Scale Categorical Data With Structural Zeros." *Survey Methodology* 40: 125–134.

Miller, D., M. Robbins, and J. Habiger. 2010. "Examining the Challenges of Missing Data Analysis in Phase Three of the Agricultural Resource Management Survey." In Proceedings of the JSM, Section on Survey Research Methods: American Statistical Association. Alexandria, VA, 816–823.

Mishra, A.K., and D.M. Holthausen. 2002. "Effect of Farm Income and Off-Farm Wage Variability on Off-Farm Labor Supply." *Agricultural and Resource Economics Review* 31: 187–199.

National Research Council. 2008. *Understanding American Agriculture: Challenges for the Agricultural Resource Management Survey*. Washington, D.C.: The National Academies Press.

Nelsen, R.B. 2009. *An introduction to Copulas*. New York: Springer.

Raghunathan, T., J. Lepkowski, J. van Hoewyk, and P. Solenberger. 2001. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology* 27: 85–95.

Raghunathan, T.E., P.W. Solenberger, and J. van Hoewyk. 2002. *Iveware: Imputation and Variance Estimation Software*. Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.

Raghunathan, T., J. Reiter, and D. Rubin. 2003. "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics* 19: 1–16.

Reiter, J.P. 2002. "Satisfying Disclosure Restrictions With Synthetic Data Sets." *Journal of Official Statistics* 18: 531–544.

Reiter, J.P. 2005. "Releasing Multiply Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study." *Journal of the Royal Statistical Society* Series A 168: 185–205. DOI: http://dx.doi.org/10.1111/j.1467-985X.2004.00343.x.

Robbins, M.W., S.K. Ghosh, B. Goodwin, J.D. Habiger, D. Miller, and T.K. White. 2011. *Multivariate Imputation Methods for Addressing Missing Data in the Agricultural Resource Management Survey (ARMS)*. A NISS/NASS collaborative research project, National Agricultural Statistics Service/National Institute of Statistical Sciences.

Robbins, M.W., and T.K. White. 2011. "Farm Commodity Payments and Imputation in the Agricultural Resource Management Survey." *American Journal of Agricultural Economics* 93: 606–612. DOI: http://dx.doi.org/10.1093/ajae/aaq166.

Robbins, M.W., S.K. Ghosh, and J.D. Habiger. 2013. "Imputation in High-Dimensional Economic Data as Applied to the Agricultural Resource Management Survey." *Journal of the American Statistical Association* 108: 81–95. DOI: http://dx.doi.org/10.1080/01621459.2012.734158.

Robbins, M.W., and T.K. White. Forthcoming. "Direct Payments, Cash Rents, Land Values, and the Effects of Imputation in U.S. Farm-Level Data." *Agricultural and Resource Economics Review*.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rubin, D.B. 1993. "Discussion of Statistical Disclosure Limitation." *Journal of Official Statistics* 9: 461–468.

Rubin, D.B. 1996. "Multiple Imputation After 18 + Years." *Journal of the American Statistical Association* 91: 473–489. DOI: http://dx.doi.org/10.1080/01621459.1996.10476908.

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

Scott, D.W. 2009. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Vol. 383. New York: Wiley.

Sheather, S.J., and M.C. Jones. 1991. "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation." *Journal of the Royal Statistical Society* Series B 53: 683–690.

Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. Vol. 26. New York: CRC Press.

Su, Y.-S., M. Yajima, A.E. Gelman, and J. Hill. 2011. "Multiple Imputation with Diagnostics (mi) in r: Opening Windows into the Black Box." *Journal of Statistical Software* 45: 1–31.

Sumner, D.A. 1982. "The Off-Farm Labor Supply of Farmers." *American Journal of Agricultural Economics* 64: 499–509. DOI: http://dx.doi.org/10.2307/1240642.

Templ, M., A. Kowarik, and P. Filzmoser. 2011. "Iterative Stepwise Regression Imputation Using Standard and Robust Methods." *Computational Statistics & Data Analysis* 55: 2793–2806. DOI: http://dx.doi.org/10.1016/j.csda.2011.04.012.

U.S. Department of Agriculture. 2011. *Farm Production Expenditures 2010 Summary*. Washington, D.C.

Van Buuren, S., and C.G.M. Oudshoorn. 1999. *Flexible Multivariate Imputation by MICE*. Leiden: TNO Preventie en Gezondheid. For associated software see http://www.multiple-imputation.com (accessed October 21, 2014).

Woodcock, S.D., and G. Benedetto. 2009. "Distribution-Preserving Statistical Disclosure Limitation." *Computational Statistics and Data Analysis* 53: 4228–4242. DOI: http://dx.doi.org/10.1016/j.csda.2009.05.020.

Received November 2012
Revised September 2014
Accepted September 2014