Statistica Sinica Preprint No: SS-2018-0207	
Title	On algorithmic and modeling approaches to imputation in large data sets
Manuscript ID	SS-2018-0207
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0207
Complete List of Authors	Roderick Little
Corresponding Author	Roderick Little
E-mail	rlittle@umich.edu
Notice: Accepted version subject to English editing.	

On Algorithmic and Modeling Approaches to Imputation in Large Data Sets

Roderick J. Little

University of Michigan

Abstract

The machine learning and statistical modeling cultures provide contrasting approaches to statistical analysis. Loh, Eltinge, Cho and Li compare these approaches in the setting of imputation of large data sets, recommending machine-learning methods. All the compared methods make assumptions, and I note that these assumptions receive more critical assessment for the model-based approaches than for the tree-based machine-learning methods. I discuss in particular the assumptions about the missing-data mechanism implied by the differing approaches. I question the extent to which general conclusions can be drawn from their simulation study, given the relatively strong performance of the method that discards the incomplete cases, and the limited exploration of the relevant design space.

1. Introduction: Machine learning and statistical modeling approaches to imputation.

There is a spirited competition between classical statistical modeling and algorithmic machine learning-based approaches popular in computer science, particularly in the context of "big data" not collected according to a purposeful statistical design. Lively references include Tukey (1962), Donoho (2017) and, in particular, Breiman (2001). The latter considers the prediction of an outcome variable Y based on a set of predictors X, and contrasts classical parametric statistical modeling with machine-learning approaches, where the relationship between Y and X is viewed as a "black box", and the focus is on algorithms like trees, forests and neural nets that lead to good predictions.

Loh, Eltinge, Cho, and Li (2019, henceforth LECL) consider this competition in the context of large data sets with a general multivariate pattern of missing data. The authors assemble a wide array of imputation approaches, and compare them in the context of a real (although somewhat narrowly defined) missing data problem from the Consumer Expenditure Survey. The comparison is worthwhile, but I think the simulation in the paper provides limited information about the strengths and weaknesses of the approaches. I think the paper reflects a widespread tendency to minimize the importance of assumptions in "automatic" machine-learning approaches, particularly when compared with more explicitly model-based alternatives. These

concerns motivate this commentary.

Imputation of missing data is a form of prediction, and since tree and forest methods seem to do well for prediction, they seem good candidates for imputation. The likelihood-based approaches in LECL are in the general realm of classical statistical modeling, although their imputations are not always based on the likelihood for a coherent joint distribution of the variables. This is the case for the "chained equations" method discussed further below. So-called "doubly-robust" methods of imputation that incorporate inverse probability weights are more accurately termed "quasi-likelihood" methods.

Tree-based regression methods have a long history, dating back at least to Belson (1959). A popular early tree method in the social sciences was Automatic Interaction Detection (Morgan and Sondquist, 1963). Tree-based methods are algorithms, but I would argue that underlying them are statistical models, which have their own set of assumptions, strengths and weaknesses. Categorizing continuous variables requires a choice of cut-points, and often relationships of continuous predictors are smooth rather than the step functions implied by categorization. Alternatives such as splines retain smoothness without imposing strong assumptions on the form of the relationship between outcome and predictor. The data-driven approach to forming trees is akin to forward selection methods in regression, and shares some of the weaknesses of that approach. It provides ample allowance for interactions that might be neglected in

additive regression models, but it may be more reasonable to give main effects and low-order interactions higher priority than higher order effects in a regression model; in Bayesian modeling this can be achieved by assuming flat prior distributions for low-order effects, but proper prior distributions for high-order effects that allow shrinkage of their coefficients towards zero.

Forests can be viewed as arising from mixtures of tree models, sacrificing interpretability of the black box for improved prediction. Indeed, mixing over a set of plausible models seems to work well in prediction competitions like Netflix (Bell, Koren and Volinsky, 2008). From my Bayesian perspective, Bayesian model mixing is a desirable alternative, as models are weighted by their posterior plausibility, with weights that provide useful information about the plausibility of each model.

2. The Comparisons in LECL.

LECL compare imputation methods for estimating the mean of Y = amount of interest and dividend income, for people with this income type, in the Bureau of Labor Statistics Consumer Expenditure Survey (CES). Methods compared include various tree and forest methods of imputation, and imputation based on "model-based" approaches. Simulations are conducted to mimic the CES missing data pattern. The paper favors tree and forest methods over model-based alternatives, as can be seen in the following extract from the abstract (italics are mine): "Standard adjustments based on item imputation and on propensity weighting... can be challenging when

auxiliary variables are numerous and are themselves subject to incomplete data problems. This paper shows how classification and regression trees can overcome these problems...The results show that if the number of auxiliary variables is not small or if they have substantial missingness rates, likelihood methods can be impracticable or inapplicable. Tree or forest methods are always applicable, are relatively fast, and have higher efficiency under real-data situations with incomplete patterns similar to that in the aforementioned [Consumer Expenditure] Survey."

The thirteen imputation methods compared in the paper can be grouped into three broad classes: (a) Trees and forests to predict the response indicator R. The inverse of the resulting predicted probabilities are then used as weights (IPW). Specifically, GCT and RCT model R using GUIDE and CART classification trees, and GCF is a forest version of GCT. (b) Trees and forests to impute missing values of Y. Specifically, GRT and RRT impute the conditional mean of Y using GUIDE and RPART classification trees, and GRF is a forest version of GRT. (c) Other "existing" approaches. Specifically, SIM = mean imputation; MICE = chained equation multiple imputation (MI), which regresses each variable in turn on all the others, with missing values replaced by most recent imputations. LECL use MICE software (van Buuren and Groothuis-Oudshoorn, 2011). AME = Amelia; AIPW = Amelia-augmented IPW; DRT, DRF = doubly robust methods, with the mean of Y predicted using GRT, GRF.

3. Applicability of the Methods, and their Underlying Assumptions.

Regression methods, based on trees or parametric models, assume that the regressions estimated on cases with Y observed are well-specified. They also assume that the predictions based on the observed data apply to cases with Y missing. This in turn involves assumptions about the missingness mechanism, since nonresponse is not under our control.

The assumptions of parametric models, and in particular the assumed form of the mean function relating Y to the X's, tend to be explicit – which variables are included, assumed functional form, which interactions are included, and so on. The fact that assumptions are explicit in parametric models can be seen as a strength, not a weakness, since the statement allows for model criticism and refinement. The assumptions of parametric methods, both concerning what predictors are included in the mean function and the missing at random (MAR) assumption (Rubin, 1976; Little and Rubin, 2002) for the missingness mechanism, are mentioned prominently in LECL.

Of the "model-based" methods, I like the chained equations MI, represented in LECL by MICE (van Buuren and Groothuis-Oudshoorn, 2011). MI allows the propagation of imputation uncertainty, and the chained equation modeling of a sequence of conditional distributions is very flexible. In particular, regressions can be tailored to variable type – linear regression for continuous outcomes, logistic regression for binary outcomes, Poisson regression for count data, and so on; splines can be used

to model nonlinear relationships with continuous variables; and interactions can be included to model lack of additivity. This is achieved at the expense of a lack of a coherent joint distribution of the variables, but I think flexibility trumps theoretical cohesiveness in applications. Concerning chained equation MI, LECL write that "Little is known about the performance of the methods in real-world settings where variables are not normally distributed (e.g., categorical variables) and probabilities of missingness are not determined by logistic regression... To our knowledge, only three published simulation studies used real data ... None had more than 20 X variables and only one had missing values in X."

I would respond that MI have been extensively applied, to both large and small data sets –Google Scholar currently lists over 2300 citations to MICE, and over 1600 citations to the alternative IVEware (Raghunathan, Lepkowski, vanHoewyk, and Solenberger, 2001). This count ignores applications using other software. Van Buuren and Groothuis-Oudshoorn (2011) list about 80 references to the application of chained equation MI in real applications, many of which are not restricted to normal variables.

The earliest application of chained equations MI was to the 1989 Survey of Consumer Finances (Kennickell, 1991). Khare, Little, Rubin, and Schafer (1993) apply MI based on a joint model to multivariate missing data in the (quite large) National Health and Nutrition Examination Survey (NHANES). The method does very well in

an associated simulation study on real NHANES data (Ezzati-Rice et al., 1993, 1995), which assesses confidence coverage as well as point estimation. Large surveys that use chained equation MI include the Consumer Expenditure Survey (Paulin, Fisher, and Reyes-Morales 2006), and the Health Interview Survey (Schenker et al, 2006). Lee and Carlin (2010) assess chained equation MI in a simulation study that draws random samples from a large synthetic population created to resemble data from the US National Longitudinal Study of Adolescent Health. Akande, Li and Reiter (2017) compare MI methods for categorical data in a simulation based on the American Community Survey, favoring default regression tree and Bayesian mixture model approaches over default chained equations approaches based on additive models.

LECL claim that MICE "fails" with many X's. I think this means that they couldn't get it to work, since there is no obvious reason why regression on a large set of units and predictors can't succeed. See, for example, Stuart, Azur, Frangakis and Leaf (2009), who successfully apply chained equation MI to a data set with 9000 cases and 400 variables. LECL's description of missing data – "About 20% of these variables have missing values; 67 of them have more than 95% missing values" – seems to confuse missingness with "not applicable" (NA), which is not really missing data.

LECL suggest multicollinearity (for normal regression) and quasi-complete separation (for logistic regression) as explanations for failure of chained equation methods in their setting. These are much studied problems. Multicollinearity can be overcome

by simple approaches such as stepwise selection, which, although flawed, may not be too bad for prediction (e.g. Dempster, Schatzoff and Wermuth 1977), particularly given that LECL focuses on point estimation rather than the impact of variable selection on assessing imputation uncertainty.

I am currently engaged with colleagues in the Institute for Social Research at Michigan in a project to multiply impute wealth variables in the Health and Retirement Survey, using IVEware. This is a much larger and more complex problem than that addressed by LECL, involving simultaneously imputing both recipiency and amount of many wealth-related variables, in a longitudinal setting with large numbers of predictors, some also missing. We have encountered problems with multicollinearity, often because more than one slightly different version of the same variable exists in the data file. Excluding such duplicates, and applying methods like stepwise selection, overcomes these problems. More sophisticated alternatives to stepwise regression are regularization methods like ridge regression or the lasso, which are applied to chained equations in Deng, Chan, Edo and Long (2016).

Concerning logistic regression, Clogg, et al. (1991) describe a simple remedy for quasi-complete separation in the application of logistic regression to multiple imputation for the (extremely large) U.S. Census industry and occupational recoding project. Computing power has advanced exponentially since that application, which is now more than 25 years old.

The impression given by LECL is that the tree algorithm will automatically lead to good predictions of missing values. I think this uncritical assessment is common for algorithmic methods, where the underlying model is treated as a "black box" and not explicitly scrutinized. But tree methods do make assumptions about the form of the mean function, as discussed, for example, in James et al. (2013). In particular, the categorization of continuous predictors assumes that the relationship with the outcome is a step function that is flat for the intervals within each category, and has jumps between the categories. The set of predictors available at each split is determined by forward selection, a method that is known to have limitations as a variable selection method (e.g. Dempster, Schatzoff and Wermuth, 1977).

Concerning the missingness mechanism, LECL argue that the MAR assumption that underlies chained equation methods is "artificial", but the missingness assumptions underlying tree-based methods is "natural". The tree methods assume a missing not at random (MNAR) mechanism, because they include indicators of missingness of predictors as covariates. In their simulations, LECL create a complete data set from the CES, and then use a GUIDE forest to estimate the missingness probability for each case and create missing values. This method of creating missing data applies the same MNAR mechanism as that assumed by the GUIDE methods, biasing the simulations against methods that assume other mechanisms, including MAR. LECL write: "A major feature of the experimental design is the novelty of ensuring that

predictor variables are naturally missing, i.e., not constrained to be MAR, in the simulation population." Generating data under a particular MNAR mechanism does not address the performance of methods for the multitude of other possible MNAR mechanisms. Moreover, the specific form of MNAR assumed by the tree methods is not scrutinized. It can be illustrated in the simple case where both the outcome Y and one of the predictors (say X) have missing values, and other variables (say Z) are fully observed. This leads to four patterns of missing data, as in Figure 1.

Figure 1 about here

Let R_X and R_Y denote response indicators for X and Y respectively, with value 1 if the corresponding variable is observed and 0 if it is missing. Any imputation method assumes (implicitly or explicitly) a model for the joint distribution for (X, Y, R_X, R_Y) given Z. A general statement of the MAR assumption (see Example 1.13 in Little and Rubin, 2002) is

$$\Pr(R_X = R_Y = 0|Z, X, Y) = g_{00}(Z)$$

$$\Pr(R_X = 0, R_Y = 1|Z, X, Y) = g_{01}(Z, Y)$$

$$\Pr(R_X = 1, R_Y = 0|Z, X, Y) = g_{10}(Z, X)$$

$$\Pr(R_X = R_Y = 1|Z, X, Y) = 1 - g_{01}(Z, Y) - g_{10}(Z, X) - g_{00}(Z),$$

with functions $g_{00}(Z), g_{01}(Z,Y), g_{10}(Z,X)$, that do not need to be explicitly modeled under the MAR model-based approach. In terms of the predictive distributions for imputation, these conditions imply that

$$f(X|Z,Y,R_X=0,R_Y=1,\theta)=f(X|Z,Y,\theta) \tag{1}$$

$$f(Y|Z,X,R_Y=0,R_X=1,\theta)=f(Y|Z,X,\theta) \tag{2}$$

$$f(X,Y|Z,R_X=0,R_Y=0,\theta)=f(X,Y|Z,\theta) \tag{3}$$

where densities f are distinguished by their arguments. In MICE and other model-based MAR approaches, imputations of missing values of Y and X are linked to the complete-data distributions on the right sides of (1-3) by iteration. On the other hand, suppose that, as in a tree algorithm, R_X is included as a predictor in the regressions to impute missing values of Y, and R_Y is included as a predictor in the regressions to impute missing values of X. Resulting imputations make the following MNAR assumptions, since the regression on the right side of each equation is estimated on the data and then used to impute the missing values of the outcomes of the regression on the left side:

$$f(Y|Z,X,R_Y=0,R_X=1,\theta)=f(Y|Z,X,R_Y=1,R_X=1,\theta) \tag{4}$$

$$f(Y|Z,R_Y=0,R_X=0,\theta)=f(Y|Z,R_Y=1,R_X=0,\theta) \tag{5}$$

$$f(X|Z,Y,R_X=0,R_Y=1,\theta)=f(X|Z,Y,R_X=1,R_Y=1,\theta) \quad (6)$$

$$f(X|Z,R_X=0,R_Y=0,\theta)=f(X|Z,R_X=1,R_Y=0,\theta). \quad (7)$$

This is a form of pattern-mixture model (Little, 1993). It is not clear to me why this particular MNAR assumptions (4-7) are better or "more natural" than the MAR assumptions (1-3). In particular, if X is predictive of both Y and missingness of Y (R_Y) after conditioning on Z, then the failure to condition on X when imputing Y for cases where X is missing ($R_X = 0$) leads to bias. Empirically, we cannot tell whether (4-7) is better or worse than (1-3) –there is no information in the data to decide this question – but tree and forest methods do make an assumption about the missingness mechanism, that is, they are not "assumption-free."

Statements in LECL about inefficiency of MAR-based methods are based not on theory, but on (to my mind) questionable implementations of these methods in the simulation study. I conjecture based on considerations in Little (1993) that methods based on (1-3) are likely to be more, not less, efficient than methods based on (4-7). The reason is that MAR bases imputations of missing values on the distribution of all the data, whereas (4-7) bases imputations on the distribution of subsets of the data.

Of course, the patterns in LECL are more complex than Figure 1, but insight is conveyed by looking at simpler cases.

4. The LECL Simulation Findings.

Getting particular software implementations of methods to work, with acceptable

computing times, is an important issue, but it is a moving target, because software and computing power are constantly evolving. These practical considerations aside, the main tools for assessing the properties of statistical methods are theory and simulation studies. Since LECL do not advance theoretical arguments in favor of their tree methods, the main basis for comparison of methods is their simulation study.

Simulation studies are experiments, and good ones adopt the classical ideas of experimental design, going back to R.A. Fisher. That is:

(a) decide on the factors that potentially affect the relative performance of the methods

(b) Manipulate these factors in a (fractional) factorial statistical design that attempts to cover the relevant design space.

(c) Apply analysis of variance of the results for key outcomes to assist in interpreting conclusions.

Many factors seem important in the LECL imputation setting, including sample size, fraction of missing information, missingness mechanism, form and strength of the true relationship between the variable with missing data and predictors (and in particular the assumed functional forms and prominence of main effects relative to interactions), form and strength of the true relationship between missingness and its predictors, degree of association between the propensity to respond and the variable

with missing values, and degree of misspecification of the true models for missingness and the survey variables.

Viewed from this perspective, the simulation study in LECL studies the mean of a single variable in a single population, manipulates just one factor, namely the number of predictors X, and considers just two outcomes, bias and root mean squared error (RMSE) of the estimates. I question whether general conclusions about the compared methods can be justified from such a limited exploration.

This is particularly true given the finding that SIM – imputing the unconditional mean of Y, a method equivalent to complete-case analysis – seems to do about as well as any of the other methods in terms of RMSE (see Figure 4 in LECL). The relative absence of correctable bias in SIM suggests that the simulated mechanism, as it relates to this particular Y, is not far from missing completely at random (MCAR, see Rubin, 1976), notwithstanding the factors conditioned by the GUIDE method of creating missing data. Deviations from MCAR are generally needed to differentiate alternative imputation methods, which aim to use information on X to reduce bias and increase precision relative to SIM.

From the LECL results in Figure 4, it seems that CART weighting is poor – perhaps a problem with extreme weights? Also, MICE either does not work or is terrible... but imputing Y using MICE with no covariates at all is equivalent to SIM, aside from simulation error from imputing draws, which can be rendered negligible

by increasing the number of multiple imputations. This equivalent "null" version of MICE should do about as well as SIM. Inclusion of any good predictors of Y should improve the precision of estimates relative to this null version, leading to reduced RMSE. These observations conflict with the reports that MICE either does not work or has much higher RMSE than SIM.

5. Conclusions

Tree and forest methods may indeed be useful tools for imputation – imputation is a form of prediction, and these methods can be good at this. The methods are algorithmic and "atheoretical", but they are not assumption-free, making assumptions about the missingness mechanism and the form of the predictive distributions of the missing values. The "automatic" nature of these methods tends to divert attention from the question of whether the assumptions underlying them are reasonable.

MI methods like MICE impute draws from a predictive distribution, and create $D > 1$ filled-in data sets with different values imputed. Imputing draws is useful when statistics other than means are of interest, like extreme percentiles of the distribution, regression parameters, or nonlinear quantities. Multiple imputation is useful for improving estimation efficiency and propagating imputation error, which is important for valid statistical inferences. None of these considerations are addressed by LECL, who restrict attention to point estimation of a mean.

I have not focused here on methods that use of propensity weights to improve

the robustness of imputation. Besides the methods discussed by LECL, an approach not discussed by these authors but worthy of consideration is penalized spline of propensity prediction (Little and An, 2004, Zhang and Little 2009).

Comparisons of alternative imputation approaches, as in LECL, are of interest. Assessment of underlying assumptions, and systematic and even-handed simulation experiments, may provide better information about the relative strengths and weaknesses of methods.

Acknowledgements

This work is supported by Grant 1R21HD090366-01A1 from the National Science Foundation. I appreciate the suggestions of an associate editor and three referees on earlier drafts of this article.

References

Akande, O., Li, F. and Reiter, J. (2017). An Empirical Comparison of Multiple Imputation Methods for Categorical Data. Am. Statist., 71, 2, 162-170.

Bell, R.M., Koren, Y. and Volinsky, C. (2008). The BellKor 2008 Solution to the Netflix Prize. AT&T Labs, Research Florham Park, NJ.

Belson, W. A. (1959). Matching and prediction on the principle of biological classification. Appl. Statist., 8, 65-75.

Breiman, L. (2001). Statistical Modeling: Two Cultures. Statist. Sci., 16,3, 199-231.

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B. and Weidman, L. (1991). Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression. J. Amer. Statist. Assoc., 86, 413, pp. 68-78.

Dempster, A.P., Schatzoff, M. & Wermuth, N. (1977). A Simulation Study of Alternatives to Ordinary Least Squares. J. Am. Statist. Assoc., 72, 77-91

Deng, Y., Chang, C., Ido, M.S. & Long, Q. (2016). Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data. Scientific Reports, 6, Article number: 21689

Donoho, D. (2017). 50 years of Data Science. J. Comp. Graphical Statist., 26, 4, 744-766.

Ezzati-Rice, T.M., Khare, M., Rubin, D.B., Little, R.J.A. & Schafer, J.L. (1993). A Comparison of Imputation Techniques in the Third National Health and Nutrition Examination Survey. Proc. Survey Res. Methods Section, Am. Statist. Assoc. 1993, 303-308.

Ezzati-Rice, T., Johnson, W., Khare, M., Little, R., Rubin, D., & Schafer, J. (1995). A Simulation Study to Evaluate the Performance of Model-Based Multiple Imputations In NCHS Health Examination Surveys. Proc. 1995 Annual Research Conference, U.S. Bureau of the Census, 257-266.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R, Springer: New York.

Kennickell, A.B. (1991). Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation. Mimeo, Board of Governors of the Federal Reserve System.

Khare, M., Little, R.J.A., Rubin, D.B., & Schafer, J.L. (1993). Multiple Imputation of NHANES III. Proc. Survey Res. Methods Section, Am. Statist. Assoc. 1993, 297-302.

Lee, K.J. & Carlin, J.B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. Am J Epidemiol., 171, 5, 624-32.

Little, R.J. (1993). Pattern-Mixture Models for Multivariate Incomplete Data. J. Am. Statist. Assoc., 88, 125-134.

Little, R.J. & An, H. (2004). Robust Likelihood-Based Analysis of Multivariate Data with Missing Values. Statist. Sinica, 14, 949-968.

Little, R.J. & Rubin, D.B. (2002). Statistical Analysis with Missing Data, 2nd ed. New York: Wiley

Loh, W-Y., Eltinge, J., Cho, M. J. & Li, Y. (2017). Classification and Regression Trees and Forests for Incomplete data from Sample Surveys. Online preprint, Statistica Sinica.

Morgan, J. A. & Sonquist, J. N. (1963). Problems in the Analysis of Survey Data: and a Proposal. J. Amer. Statist. Assoc., 58, 415-434.

Paulin, G., Fisher, J. & Reyes-Morales, S. (2006). Multiple Imputation Manual: Supplement to 2004 Consumer Expenditure Interview Survey Public Use Microdata Documentation, US Department of Labor, Bureau of Labor Statistics, Division of Consumer Expenditure Surveys.

Raghunathan, T., Lepkowski, J. VanHoewyk, M., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, Survey Methodol. 27, 1, 85-95. For associated IVEWARE software see http://www.isr.umich.edu/src/smp/ive/.

Rubin, D. B. (1976). Inference and Missing Data. Biometrika, 63, 581-592.

Schenker, N., Raghunathan, T.E., Chiu, P-L., Makuc, D.M., Zhang, G. and Cohen, A.J. (2006). Multiple Imputation of Missing Income Data in the National Health Interview Survey. J. Amer. Statist. Assoc., 101, 924-933.

Stuart, E.A., Azur, M., Frangakis, C. and Leaf, P. (2009), Multiple Imputation with Large Data Sets: A Case Study of the Children's Mental Health Initiative. Am. J. Epid., 169, 9, 1133–1139.

Tukey, J.W. (1962). The Future of Data Analysis. Ann. Math. Statist., 33, 1, 1-67.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. J. Statist. Software, 45, 1–67.

Zhang, G. & Little, R. J. (2009). Extensions of the Penalized Spline of Propensity

Prediction Method of Imputation. Biometrics, 65, 3, 911-918.Applicability of the Methods, and their Underlying Assumptions.

Regression methods, based on trees or parametric models, assume that the regressions estimated on cases with Y observed are well-specified. They also assume that the predictions based on the observed data apply to cases with Y missing. This in turn involves assumptions about the missingness mechanism, since nonresponse is not under our control. The assumptions of parametric models, and in particular the assumed form of the mean function relating Y to the X's, tend to be explicit – which variables are included, assumed functional form, which interactions are included, and so on. The fact that assumptions are explicit in parametric models can be seen as a strength, not a weakness, since the statement allows for model criticism and refinement. The assumptions of parametric methods, both concerning what predictors are included in the mean function and the missing at random (MAR) assumption (Rubin, 1976; Little and Rubin, 2002) for the missingness mechanism, are mentioned prominently in LECL. Of the "model-based" methods, I like the chained equations MI, represented in LECL by MICE (van Buuren and Groothuis-Oudshoorn, 2011). MI allows the propagation of imputation uncertainty, and the chained equation modeling of a sequence of conditional distributions is very flexible. In particular, regressions can be tailored to variable type – linear regression for continuous outcomes, logistic regression for binary outcomes, Poisson regression for count data, and so on; splines

can be used to model nonlinear relationships with continuous variables; and interactions can be included to model lack of additivity. This is achieved at the expense of a lack of a coherent joint distribution of the variables, but I think flexibility trumps theoretical cohesiveness in applications. Concerning chained equation MI, LECL write that "Little is known about the performance of the methods in real-world settings where variables are not normally distributed (e.g., categorical variables) and probabilities of missingness are not determined by logistic regression... To our knowledge, only three published simulation studies used real data ... None had more than 20 X variables and only one had missing values in X." I would respond that MI have been extensively applied, to both large and small data sets –Google Scholar currently lists over 2300 citations to MICE, and over 1600 citations to the alternative IVEware (Raghunathan, Lepkowski, vanHoewyk, and Solenberger, 2001). This count ignores applications using other software. Van Buuren and Groothuis-Oudshoorn (2011) list about 80 references to the application of chained equation MI in real applications, many of which are not restricted to normal variables. The earliest application of chained equations MI was to the 1989 Survey of Consumer Finances (Kennickell, 1991). Khare, Little, Rubin, and Schafer (1993) apply MI based on a joint model to multivariate missing data in the (quite large) National Health and Nutrition Examination Survey (NHANES). The method does very well in an associated simulation study on real NHANES data (Ezzati-Rice et al., 1993, 1995), which assesses confi-

dence coverage as well as point estimation. Large surveys that use chained equation MI include the Consumer Expenditure Survey (Paulin, Fisher, and Reyes-Morales 2006), and the Health Interview Survey (Schenker et al, 2006). Lee and Carlin (2010) assess chained equation MI in a simulation study that draws random samples from a large synthetic population created to resemble data from the US National Longitudinal Study of Adolescent Health. Akande, Li and Reiter (2017) compare MI methods for categorical data in a simulation based on the American Community Survey, favoring default regression tree and Bayesian mixture model approaches over default chained equations approaches based on additive models. LECL claim that MICE "fails" with many X's. I think this means that they couldn't get it to work, since there is no obvious reason why regression on a large set of units and predictors can't succeed. See, for example, Stuart, Azur, Frangakis and Leaf (2009), who successfully apply chained equation MI to a data set with 9000 cases and 400 variables. LECL's description of missing data – "About 20% of these variables have missing values; 67 of them have more than 95% missing values" – seems to confuse missingness with "not applicable" (NA), which is not really missing data. LECL suggest multicollinearity (for normal regression) and quasi-complete separation (for logistic regression) as explanations for failure of chained equation methods in their setting. These are much studied problems. Multicollinearity can be overcome by simple approaches such as stepwise selection, which, although flawed, may not be too bad for prediction (e.g.

Dempster, Schatzoff and Wermuth 1977), particularly given that LECL focuses on point estimation rather than the impact of variable selection on assessing imputation uncertainty. I am currently engaged with colleagues in the Institute for Social Research at Michigan in a project to multiply impute wealth variables in the Health and Retirement Survey, using IVEware. This is a much larger and more complex problem than that addressed by LECL, involving simultaneously imputing both recipiency and amount of many wealth-related variables, in a longitudinal setting with large numbers of predictors, some also missing. We have encountered problems with multicollinearity, often because more than one slightly different version of the same variable exists in the data file. Excluding such duplicates, and applying methods like stepwise selection, overcomes these problems. More sophisticated alternatives to stepwise regression are regularization methods like ridge regression or the lasso, which are applied to chained equations in Deng, Chan, Edo and Long (2016). Concerning logistic regression, Clogg, et al. (1991) describe a simple remedy for quasi-complete separation in the application of logistic regression to multiple imputation for the (extremely large) U.S. Census industry and occupational recoding project. Computing power has advanced exponentially since that application, which is now more than 25 years old. The impression given by LECL is that the tree algorithm will automatically lead to good predictions of missing values. I think this uncritical assessment is common for algorithmic methods, where the underlying model is treated as a "black

box" and not explicitly scrutinized. But tree methods do make assumptions about the form of the mean function, as discussed, for example, in James et al. (2013). In particular, the categorization of continuous predictors assumes that the relationship with the outcome is a step function that is flat for the intervals within each category, and has jumps between the categories. The set of predictors available at each split is determined by forward selection, a method that is known to have limitations as a variable selection method (e.g. Dempster, Schatzoff and Wermuth, 1977). Concerning the missingness mechanism, LECL argue that the MAR assumption that underlies chained equation methods is "artificial", but the missingness assumptions underlying tree-based methods is "natural". The tree methods assume a missing not at random (MNAR) mechanism, because they include indicators of missingness of predictors as covariates. In their simulations, LECL create a complete data set from the CES, and then use a GUIDE forest to estimate the missingness probability for each case and create missing values. This method of creating missing data applies the same MNAR mechanism as that assumed by the GUIDE methods, biasing the simulations against methods that assume other mechanisms, including MAR. LECL write: "A major feature of the experimental design is the novelty of ensuring that predictor variables are naturally missing, i.e., not constrained to be MAR, in the simulation population." Generating data under a particular MNAR mechanism does not address the performance of methods for the multitude of other possible MNAR

mechanisms. Moreover, the specific form of MNAR assumed by the tree methods is not scrutinized. It can be illustrated in the simple case where both the outcome Y and one of the predictors (say X) have missing values, and other variables (say Z) are fully observed. This leads to four patterns of missing data, as in Figure 1. Figure 1 about here Let ${{R}_{X}}$ and ${{R}_{Y}}$denote response indicators for X and Y respectively, with value 1 if the corresponding variable is observed and 0 if it is missing. Any imputation method assumes (implicitly or explicitly) a model for the joint distribution for $(X,Y,{{R}_{X}},{{R}_{Y}})$ given Z. A general statement of the MAR assumption (see Example 1.13 in Little and Rubin, 2002) is $\begin{align} & \Pr ({{R}_{X}}={{R}_{Y}}=0|Z,X,Y))={{g}_{00}}(Z) \\ & \Pr ({{R}_{X}}=0,{{R}_{Y}}=1|Z,X,Y)={{g}_{01}}(Z,Y) \\ & \Pr ({{R}_{X}}=1,{{R}_{Y}}=0|Z,X,Y \\ \\ & \Pr ({{R}_{X}}={{R}_{Y}}=1|X,Z,Y)=1-{{g}_{01}}(Z,Y)-{{g}_{10}}(Z,X)-{{g}_{00}}(Z), \\ \end{align}$ with functions ${{g}_{00}},{{g}_{01}},{{g}_{10}}$ that do not need to be explicitly modeled under the MAR model-based approach. In terms of the predictive distributions for imputation, these conditions imply that $\begin{align} & f(Y|Z,X,{{R}_{X}}=1,{{R}_{Y}}=0,\theta)=f(Y|Z,X,\theta) \\ & f(X|Z,Y,{{R}_{Y}}=1,{{R}_{X}}=0,\theta)=f(X|Z,Y,\theta) \\ & f(X,Y|Z,{{R}_{X}}=0,{{R}_{), \\ \end{align}$ (1) where densities f are distinguished by their arguments. In MICE and other model-based MAR approaches, imputations of missing values of Y and X are linked to the complete-data distributions on the right side of (1) by itera-

tion. On the other hand, suppose that, as in a tree algorithm, ${{R}_{X}}$is included as a predictor in the regressions to impute missing values of Y, and${{R}_{Y}}$is included as a predictor in the regressions to impute missing values of X. Resulting imputations make the following MNAR assumptions, since the regression on the right side of each equation is estimated on the data and then used to impute the missing values of the outcomes of the regression on the left side: $\begin{align} & f(Y|Z,X,{{R}_{X}}=1,{{R}_{Y}}=0,\theta)=f(Y|Z,X,{{R}_{X}}=1,{{R}_{Y}}=1,\theta) \\ & f(Y|Z,{{R}_{X}}=0,{{R}_{Y}}=0,\theta)=f(Y|Z,{{R}_{X}}=0,{{R}_{Y}}=1,\theta) \\ & f(X|Z,Y,{{R}_{X}}=0,{{R}_{Y}}=1,\theta)=f(X|Z,Y,{{R}_{X}}=1,{{R}_{Y}}=1,\theta) \\ & f(X|Z,{{R}_{X}}=0,{{R}_{Y}}=0,\theta)=f(X|Z,{{R}_{X}}=0,{{R}_{Y}}=1,\theta). \\ \end{align}$ (2) This is a form of pattern-mixture model (Little, 1993). It is not clear to me why the particular MNAR assumption (2) is better or "more natural" than the MAR assumption (1). In particular, if X is predictive of both Y and missingness of Y (${{R}_{Y}}$) after conditioning on Z, then the failure to condition on X when imputing Y for cases where X is missing $({{R}_{X}}=0)$ leads to bias. Empirically, we cannot tell whether (2) is better or worse than (1) –there is no information in the data to decide this question – but tree and forest methods do make an assumption about the missingness mechanism, that is, they are not "assumption-free." Statements in LECL about inefficiency of MAR-based methods are based not on theory, but on (to my mind) questionable implementations of these methods in the

simulation study. I conjecture based on considerations in Little (1993) that methods based on (1) are likely to be more, not less, efficient than methods based on (2). The reason is that MAR bases imputations of missing values on the distribution of all the data, whereas (2) bases imputations on the distribution of subsets of the data. Of course, the patterns in LECL are more complex than Figure 1, but insight is conveyed by looking at simpler cases.