USDA United States Department of Agriculture

Forest Service

Rocky Mountain Research Station

Research Note
RMRS-RN-2

September 1998

Data Estimation and Prediction for Natural Resources Public Data

Hans T. Schreuder
Robin M. Reich

Abstract—A key product of both Forest Inventory and Analysis (FIA) of the USDA Forest Service and the Natural Resources Inventory (NRI) of the Natural Resources Conservation Service is a scientific data base that should be defensible in court. Multiple imputation procedures (MIPs) have been proposed both for missing value estimation and prediction of non-remeasured cells in annualized forest inventories such as the Southern Annual Forest Inventory System (SAFIS). MIPs generate clean-looking data bases that are easily used but hide a serious weakness: under different assumptions made by reasonable people, very different data bases and conclusions can be generated. A MIP is an interesting idea for prediction but should only be used for analyses by users, not for filling in data in a public data base. Simple illustrations are given to make our points.

To maintain a defensible data base, FIA and NRI should only provide algorithms to facilitate user-generated data for prediction of non-remeasured cells. Users, not FIA and NRI, should be responsible for generating data bases that utilize these algorithms or other algorithms of their choosing, incorporating assumptions that they are willing to make. But they should be encouraged to work with FIA and NRI personnel in utilizing such algorithms.

Keywords: forest inventory, forest analysis, database management

The purpose of this document is to assess the utility of multiple imputation procedures (MIPs) for missing data and predicting non-remeasured data points in multiresource inventories such as Forest Inventory and Analysis (FIA) of the USDA Forest Service (USFS) and Natural Resources Inventory (NRI) of the Natural Resources Conservation Service (NRCS). We focus on MIPs since they are well documented and often accepted for imputing missing data. They have been proposed for use in the Southern Annual Forest Inventory System (SAFIS) for predicting data in non-remeasured plots. Traditionally FIA and NRI have replaced units that could not be measured because access to the site was refused or for other reasons. Both FIA and NRI are considering adopting an annual inventory in each population of interest. Should MIPs be used to estimate data for the plots that are not remeasured in any given year? Are better alternatives available?

Hans T. Schreuder is Mathematical Statistician, Rocky Mountain Research Station, Fort Collins, CO. Robin M. Reich is Associate Professor of Forestry, Colorado State University, Fort Collins, CO.

Review of Literature

One goal of imputation procedures is to provide statistically valid inference in real-world situations where data base collectors/constructors and key users are different people with different objectives for the data and there is no accepted approach for imputing data.

Rubin (1996, p. 473) recommends MIPs for missing data imputation. Understanding the theory in Rubin (1987) requires expertise in and acceptance of randomization-based and Bayesian inference. Rubin (1996) and Meng (1994) are based on this theory and attempt to sell MIPs to users. MIPs for a set of missing values result in multiple sets of possible values for the true values (p. 476) which can reflect uncertainty across one or several models for nonresponse. Each set of imputations creates a complete data set, each of which can be analyzed using standard complete data analysis software to yield estimates or analyses, i.e., estimates $\hat{Q}$ with associated variance-covariance matrices $\hat{U}$, and significance values p. No matter how $\hat{Q}$ and $\hat{U}$ are calculated with complete data, once missing data are generated by imputation, the estimates can be calculated as if the data sets were complete.

One form of multiple imputation is repeated imputation (Rubin 1996, p. 476) with values obtained from the posterior predictive distribution of the missing values under a specific model, i.e., a specific Bayesian model

for both the data and the missing data estimation. The m complete data analyses with m imputations under one model yield m repeated completed data-statistics $(\hat{Q}_{*1}, \hat{U}_{*1}, \ldots, \hat{Q}_{*m}, \hat{U}_{*m})$. These are combined to draw one repeated imputation inference that adjusts properly for nonresponse to create the repeated imputations. What is "proper" in the context of a MIP depends on the complete data estimates $\hat{Q}$ and associated variance $\hat{U}$. Rubin (1996, p. 477) defines a MIP as proper if

X = an array of all background information fully available and used in a population such as in stratification,

Y = an array of outcome information in a population that is to be sampled in the survey, and

$\hat{Q}$ = estimand, a function of X and Y, i.e., $\hat{Q} = \hat{Q}(X,Y)$.

Then for the sample I, the values of the complete-data statistics $\hat{Q}$ and $\hat{U}$ created by filling in the missing Y values, $\hat{Q}_{*1}$ and $\hat{U}_{*1}$, for large m are:

$$E(\hat{Q}_{\infty} \mid X, Y, I) \approx \hat{Q} \tag{1}$$

and

$$E(\overline{U}_{\infty} \mid X, Y, I) \approx \hat{U} \tag{2}$$

and $\hat{B}_{\infty}$, the variance-covariance of the $\hat{Q}_{*1}$ across the m imputations, is approximately unbiased for the randomization variance of $\overline{Q}_{\infty}$ stated as

$$E(\hat{B}_{\infty} \mid X, Y, I) \approx \mathrm{var}(\overline{Q}_{\infty} \mid X, Y, I). \tag{3}$$

Only (3) has no direct analogue in ensuring validity for complete-data randomization theory that we think should be insisted upon for public data bases. It means that $\hat{U}$, an ancillary complete-data estimand, is approximately unbiasedly estimated after imputation.

The m pairs of estimates $(\hat{Q}_{*1}, \hat{U}_{*1}, \ldots, \hat{Q}_{*m}, \hat{U}_{*m})$ are then combined under a Bayesian paradigm for survey inference from repeated imputations. Basically this amounts to the Bayesian result: posterior mean of Q = average (repeated complete-data posterior mean of Q and posterior variance of Q = average (repeated complete data variances of Q) + var (repeated complete-data posterior means of Q) where variance refers to variance over the repeated imputations.

The repeated imputation estimator is:

$$\overline{Q}_{\mathrm{m}} = \sum_{i=1}^{m} Q_{*i} / m \tag{4}$$

and

$$\mathrm{var}(\overline{Q}_{\mathrm{m}}) = \overline{U}_{\mathrm{m}} + (m+1) B_{\mathrm{m}} / m \tag{5}$$

where $\overline{U}_{\mathrm{m}}$ = within imputation variability and B_{m} = between imputation variability.
Rubin notes that as $m \to \infty$,

$$(Q - \overline{Q}_{\infty}) \sim N(0, \mathrm{var}_{\infty}) \tag{6}$$

where $\mathrm{var}_{\infty} = \overline{U}_{\infty} + B_{\infty}$ and the eigenvalues of B_{∞} relative to var_{∞} measure the fractions of information missing about Q due to nonresponse.

When the multiple imputations are proper for $(\hat{Q}, \hat{U})$ and the complete-data inference based on $(\hat{Q}, \hat{U})$ is randomization-valid for Q, then (4) is randomization-valid for Q no matter how complex the survey design. Note: the key catch here is being "proper" for $(\hat{Q}, \hat{U})$.

Rubin (1996, p. 479) addresses two concerns about MIPs: the operational difficulty for the data base constructor and ultimate users and the acceptability of answers obtained partially through simulation; and the validity of repeated-inference imputation in the classical statistics (frequentist) sense of when multiple imputations are not proper although perhaps reasonable.

Rubin's reaction to the first criticism is quite reasonable: simulation methods are much more accepted now in statistics, and several are accepted and worthy of theoretical investigations and routine practical applications. A simulation deals only with the missing information, leaving the rest of the information to the inference method, either analytic- or simulation-based, that assumes all sample data are available. Therefore the acceptable number of imputations can be quite small if the fraction of missing information, g, is modest as is usually true in public use surveys. Rubin recommends g < 30%. He thinks that five multiple imputations are often adequate for each nonresponse model. His response to other parts of the first concern is to dismiss them as not relevant anymore. With regard to the second concern about invalid MIP inferences he does not handle this very persuasively. Meng (1994, p. 547) recognizes the key issue by stating: "The validity of assumptions is fundamental to any inference, and thus is always of great concern. Creating multiple imputations for public-use data files magnifies this concern, because the validity of the imputation model affects virtually all the subsequent analyses." He follows this up on p. 553 with, "The imputer's task is easy to state but hard to implement: to create multiple imputations for missing values that properly reflect uncertainty about these values given all the available information." So far so good. However, he thenadds (p. 553), "The key step here is to construct a probability model for predicting the missing values, for which Bayesian prediction is the only sensible general approach." A problem is that even two scientifically honest imputers can construct very different probability models and this is aggravated when the data are used for controversial issues. Clearly, this is an even more serious problem in prediction where more values are to be predicted than observed, relative to imputing some missing values, the original objective of MIP.

We think Meng (1993, p. 553) is wrong in his expressed desire: "Sensibly using all available information has been a key guideline in practice for constructing imputation models and has been emphasized repeatedly in the literature ... ." Using available information is good. But it may require important assumptions that are unlikely to hold such as assuming equal probabilities of selection for a historical sample for

which the actual probabilities of selection are lost (Schreuder and Alegria 1995) or is of unknown or arguable quality (an interesting example: some people would argue that the area of forest in a state can be more accurately estimated from aerial photographs than from ground sampling, whereas others would disagree strongly). Meng (1993, p. 554) suggests that even when a good effort is made to ensure the generality of the imputation model, the model's form and underlying assumptions should still be reported. As he notes, this helps the analyst judge whether the model is misleading for a specific analysis. But how often can one judge this correctly?

Meng (1994) states, "From an inferential point of view, perhaps the most fundamental reason for imputation is that a data collector's assessment and information about the data, both observed and unobserved, can be incorporated into the imputations." This can be quite useful for the purposes of individual users who use the data to make decisions but could be dangerous for a scientific data base because of the subjectivity introduced this way. As noted by Meng (p. 539): "Multiple imputation is motivated from the Bayesian perspective, yet survey inferences, the primary area of application so far, are traditionally dominated by frequentist analyses."

Fay (1991, 1992) and Kott (1992) question the validity of inferences based on MIP. For example, Fay demonstrates that the variance estimator from repeated-imputation combining rules does not agree asymptotically with the sampling variance of the repeated-imputation estimator even for the correct imputation model. Meng (1994, p. 539) argues that this is due to "uncongeniality" which means basically that the analysis procedure does not correspond to the imputation model. He attributes this to the analyst and imputer having access to different amounts and sources of information with different assessments about both response and nonresponse.

Meng (1994, p. 540) quotes from Fay (1991, p. 437) that a design-based approach "...first makes inferences from a sample with missing data to a census with missing data, and then evaluates the uncertainty in making inferences from the uncertain census to the population." He then makes a point that we don't like (Meng, p. 540): "...it seems to move opposite to the intended direction of multiple imputation by shifting substantial burdens to the users of survey data." We believe that burden should be on the users. Meng then proceeds to make the following scary comment that does not take public data bases into account: "Multiply imputed data can be better than observed data."

Rubin (1996) discusses alternatives to MIPs. The most attractive one seems to be a procedure that weights adjustments for nonresponse which can be useful in obtaining approximately unbiased estimates. Each unit receives as weight the inverse probability of obtaining its pattern given (X,Y) information. The nonresponse probabilities have to be estimated if patterns of occurrence are affected by nonresponse. However, the complete-data analyses of many users do not allow for sampling weights; nonresponse adjustments in weights estimated from the data are not usually accounted for in constructing confidence intervals and p-values, and special analyses and software need to be developed. Also, weighting adjustments focus on unbiased estimation and de-emphasize efficiency. For example, weighting by inverse probabilities near the boundary of the convex hull of observations can generate estimates with large variances. But we see this as a potentially useful warning, not a serious concern as Rubin does. These alternatives suffer from the same problem as MIPs, i.e., subjectivity is involved in generating missing data. So again, different users could end up with substituting very different values for missing data.

Kott (1995), a critic in some ways of MIP, states that for univariate statistics based on complex survey data in the presence of nonresponse, jackknife estimation has greater theoretical promise than repeated-imputation inference. But repeated imputation has no serious competition in handling multivariate statistics based on complex survey data with complicated patterns of nonresponse.

Discussion and Recommendations

Public data bases should be defensible in court in regard to their representativeness of populations of interest and lack of bias in use for inferences. Traditionally FIA and NRI have replaced sample locations when access was denied or was considered too dangerous. Denied access is becoming a serious problem in the United States for such surveys. We recommend that a separate stratum be set up for "denied access or too dangerous access" so that the magnitude of the problem will be known through the acreage represented by this stratum. This is in line with a recommendation made for forest health monitoring (FHM) by the USFS (Bill Smith 1996, personal communication). New samples should not be substituted for such sample units unless needed to obtain an adequate sample size. It should be safe to substitute values for some limited number of missing values using easily explained techniques such as jackknife estimation (Kott 1995). If such missing values are common, they should not be estimated, but again a stratum should be formed with estimates of land area indicating the seriousness of the problem. This is in line with Fay (1994, p. 437) who recommends that the design-based approach should first be used to make inferences from a sample with missing data to a census with missing data. Then the

uncertainty in making inferences from that census to the population should be evaluated.

We believe that published models or imputation techniques such as MIPs should not be used in improving public data bases but are useful in inferences by analysts willing to supplement these data bases. It is highly unlikely that such estimation will be widely accepted as giving reliable predictions for all parameters of interest. Even the most promising scientific models, those for growth and mortality, have been shown to be quite unreliable for FIA data. MIP and models used for improving the utility of the public data base for individual users should be encouraged by the development of easy-to-use algorithms readily adaptable to knowledge and needs of users. Considerable ancillary information is often available even though reliability is often unknown. Users should be encouraged to work closely with the people responsible for collecting and maintaining the public data bases to minimize misuse of the data because of misunderstanding the strengths and limitations of such data. The beauty of annual inventories is that promising prediction techniques can be and should be readily and rigorously tested every year. The main emphasis in estimating missing data and predictions for nonremeasured plots should be directed at developing and testing procedures acceptable to a wide range of users. Such acceptable procedures should then be recommended to users/analysts with appropriate cautions.

Examples

Assume we have a sample of 1,000 plots on which at time 1 we have measured plot basal area and number of trees. We limit our interest to these 2 variables. Some year later we select a random sample of 200 plots from the 1,000 and again measure plot basal area and number of trees, so we can calculate plot basal area growth and tree frequency for those 200 plots. There is considerable information available of potential utility in estimating the values to be predicted for the 800 plots that were not remeasured. The information can be used in four contradictory ways:

1. A government agency has available an individual tree growth model to predict the change in basal area (Δx) and number of trees (ΔN) on a given plot. The general form of the models are $\Delta x = f(dbh, bal, site, species)$ and $\Delta N = f(dbh, bal, site, species)$ where dbh is the diameter of a tree at 4.5 ft, bal is the basal area of larger trees, site is a measure of the productivity of the site, and species is the type of tree we are estimating growth for. This type of model would be used to estimate the growth of the individual trees on the plot and then summed to obtain estimates for the plot as a whole. This approach cannot generate multiple estimates as desired in MIP.

2. A forest company has extensive holdings in the area and has good diameter distribution-based growth and yield models of the form $\Delta x = W(age_1, site, dbh_1, N_1)$ and $\Delta N = f(site, age_1, N_1)$ where $\bar{dbh}$ and N_1 are the average diameter and number of trees in the plot at time 1, and $W(\ )$ is a model based on the Weibull distribution. With their statistical models, the company can generate a series of estimates as desired with MIP but the actual models used are kept secret from us since that is company policy.

3. A local environmental organization assures us that it can provide excellent values too but refuses to explain how it will generate the data values after it learns about the forest company policy. In fact, it uses the models $\Delta x = x_1 + C_1$ and $\Delta N = N_1 * C_2$ where C_1 and C_2 are environmental assessment factors generated by an expert biologist in its employ. The errors associated with these estimates are unknown, so the organization generates only best estimates for the 800 plots.

4. Good Landsat Thematic Mapper coverage is available for the two time periods and a research scientist at a nearby university assures us that she can estimate the growth and number of trees for the 800 plots using a combination of double sampling and geostatistical procedures. In this procedure, the growth on the remeasured plots are modeled as a function of the Landsat TM data and then geostatistical procedures such as kring and cokriging are used to estimate the growth on the 800 plots that were not measured in the first time period. Only best estimates can be generated for the 800 plots, however, not a set of estimates as required by the MIP.

Which option(s) do we use? Options (1) and (4) are objective but hard to defend because growth and change in number of tree models are striking in their unreliability and remote sensing is unable to live up to its promise for detailed information. Options (2) and (3) are perhaps reliable but are likely to yield totally different results and the public does not know how the data were generated. Perhaps a proper and congenial Bayesian procedure can encompass all four options but imputations based on them would likely yield very different results—a recipe for confusing users and causing debate. The main alternative to us seems to be to allow the various users to use any or all of these four options, or other ones to generate data, but keep them out of the public data sets.

References

Fay, R. R. 1991. A design-based perspective on missing data variance. Proc. 1991 Annual Research Conference. US Bureau of the Census, Wash. DC., p. 429-440.

Fay, R. E. 1992. When are inferences from multiple imputation valid? Proc. Survey Research Methods Section. Amer. Stat. Assoc., Alexandria, VA, p. 227-232.

Kott, P. S. 1992. A note on a counter-example to variance estimation using multiple imputation. Tech Rep National Agricultural Statistical Service, Wash. DC. 7 pp.
Kott, P. S. 1995. A paradox of multiple imputation. Unpublished manuscript. National Agricultural Service, Wash. DC. 17 pp.
Meng, X. L. 1994. Multiple imputation with uncongenial sources of input (with discussion). Stat. Sciences 9:538-574.
Rubin, D. B. 1987. Multiple imputation for nonresponse in surveys. J. Wiley and Co, NY.
Rubin, D. B. 1996. Multiple imputation after 18+ years. J. Amer. Stat. Assoc 91:473-489.
Schreuder, H. T.; Alegria, J. 1995. Stratification and plot selection rules: misuses and consequences. USDA Forest Service, Rocky Mountain Forest and Range Experiment Station Res. Note RM-536. 4 pp.