DISCUSSION

Stephen P. Chilton, Congressional Research Service

About Dr. House's paper I have little to say except some praise for the idea and the careful way the preliminary steps have been taken. The problem which the paper addresses is a severe one, as my own reflection and USDA's experience indicate: the population of establishments selling to farmers is very poorly approximated by a list frame drawn (primarily) from the Yellow Pages. Dr. House's paper discusses clearly the problems of incompleteness, the inclusion of establishments not selling to farmers, the difficulty in weighting the establishments, and the costliness of maintaining the list.

The idea developed to handle these problems--asking a sample of farmers from whom they purchase the commodities of interest--appears to be a good one, and Dr. House again takes care to describe the method and its potential advantages. The section on its practical feasibility appears to demonstrate that the method could be used in practice. The preliminary results comparing the coverage of the new frame and the old show that, as expected, the new frame has greater coverage than the old. Whether or not this makes a difference to the final results of the establishment survey--average prices paid by farmers for supplies--remains to be seen.

I have two quibbles with the paper. The first is that there was no discussions of the difference between regarding the sampling scheme as a cluster sample with the possibly overlapping clusters or regarding it as a random sample from a population of possible frames. The latter viewpoint is implied in the paper's discussion of ratio estimator bias, but the former seems the method of interest. If these viewpoints are not equivalent to one another, then how do we choose between them?

My other quibble is that in the last paragraph of the conclusions section, Dr. House makes an assertion which was not discussed in the text itself. The conclusions section is not the place to state new findings.

Turning now to the Ford, Kleweno and Tortora paper, I like it for going through the types of analyses which should be undertaken in any careful study of imputation methods. First, I am glad to see work being done comparing a variety of imputation methods. Work like this and that of Herzog (1980) is necessary for us to understand what classes of imputation methods are most useful. Second, the paper studies the accuracy of the imputation methods not only in terms of their effects on summary statistics but also in terms of their accuracy in predicting individual data items. Much of my work is done using microsimulation models, where complex social program rules are applied to sample cases one-by-one. These rules can be highly non-linear, and thus I require accuracy on a case-by-case basis. The authors' concentration on the individual case is therefore welcome. Third, and in a similar vein, the paper also studies the effect of the simulation process on the covariance structure. This aspect of the problem is often ignored, but it is clear that we often wish to do contingency table or correlational analysis of our data, and so the imputation process must preserve that structure as well as it can. Finally, the paper also looks at how the methods work when data is not missing at random--when respondents differ systematically from non-respondents. We know that this occurs in practice, and it is very useful to look at how sensitive the different methods are to such patterns. I find the missingness algorithms used to be rather artificial and perhaps less complex than real missingness patterns, but they are at least an attempt to provide some challenge to the missing-at-random assumption.

I do have a number of problems with the paper, however. Aside from some specific technical problems which I mention at the end of this discussion, I have three general criticisms of the paper: that it attacks the problem of imputation in a specific survey without having laid the groundwork for it; that it uses imputation methods which perpetuate the problems of variance estimation (discussed by Rubin (1977) and Herzog (1980)); and that the paper is cast at too abstract a level to be useful.

The first problem is that the authors are using data from a survey in which the missing data are already imputed and are not identified as such. They thus have no idea how many data values are missing, where the missing values are, what the true values are, or how accurate the existing imputation methods are. How is it possible for us to judge the practical utility of the imputation methods they study without having them applied to real data? (To take an extreme example, suppose that all values are missing for every case and have been imputed by the field officers (the current practice). In such a situation this study would only be showing how well the methods could reproduce the field officers' guesses.) I am surprised at this, because there are several simple studies which could be done to alleviate these problems. For example, a field could be added to the questionnaire for each item to indicate which values are missing and have been imputed. Another study might involve interviewing the field agents to discover their own imputation methods, and/or erasing data and asking them to impute it.

The second problem is that most of the methods of imputation used in the paper do not allow the proper estimation of the variance of a statistic. This has been set forth in Rubin (1977), but the matter is complex and important enough to deserve restatement here. The first point to make is that imputation methods which repeatedly impute a value (e.g., a mean value for the variable) spuriously decrease the estimated variance of a statistic. As an extreme example, consider a survey with two cases with known values of x and the remainder of the cases with imputed value AVG(x) = (x1+x2)/2. The variance of x itself will be spuriously decreased because we should be including some variance component reflecting the variance of the non-respondents' values. To impute AVG(x) for the unknown x values is implicitly to say that we know exactly the correct value of the missing x. Clearly this is not the case, and our variance computations need to reflect that uncertainty. The same is true when one uses a regression equation for imputation: imputing the

predicted value of x says implicitly that the independent variables can tell us exactly what the true value should be--that there is no residual unexplained variance. This charge of ignoring the response variability of the non-respondents can be lodged against all the methods in the paper except PRINCOMP.

A second point to make is that we are not certain of whether and to what extent non-respondents differ from respondents. All the methods in the paper implicitly assume that non-respondents' values are distributed exactly like those of respondents. A Bayesian approach could be used to reflect our uncertainty about these respondent/non-respondent differences and to assess its effect on the variance (and bias) of our statistics.

These points all bear on one theme: that missing data makes us less certain of our results, while current imputation methods, including those used in the paper, spuriously decrease it. This is a widespread problem in survey research and other areas of statistical analysis which impute for missing data, and this paper does nothing to address it.

The third problem is that in many ways the paper is too abstract to be useful. For example, the reader is never told what the variables y1, y2 and w are. This leaves me thinking that the results might be due to the nature of the survey itself (vaguely referred to as a "hog survey") or to the nature the variables themselves. Similarly, we are told scarcely anything about the ESTMAT procedure. As a consequence, I learn nothing from the paper about the class of models which includes ESTMAT; all I know is that one specific procedure had such-and-such a set of properties. In addition to these examples, there is also my general feeling that the results are ill-digested and hence present us with much less guidance than they could. For example, all six methods appear to bias the mean values downward. Why? Because the data are skewed? Because the imputation models always have that property? Without knowing this, my inclination is to avoid these models completely because I don't know what's going on with them. These problems are ones of style rather than substance, of course.

A few final comments: (1) There is no indication in the description of the PRINCOMP procedure that the variables to be imputed load on the principal component. If they had only weak loadings, the procedure would at best be imputing a random number. (2) The PRINCOMP procedure uses one dimension of similarity among cases, but there is no reason to throw away the other dimensions. I imagine that an imputation procedure using a generalized distance function (e.g., Vacek & Ashikaga (1980)) or hot deck procedure using several dimensions would provide increased accuracy. (3) The Zero Spike procedure estimates a missing value for one variable based on an average ratio between it and some other variable for which the value is known. The other variable is chosen as that having the greatest correlation with the variable to be imputed on all cases for which neither are missing. However, correlation does not imply direct proportionality between the variables because of the constant term in the regression line. The procedure should either be modified to allow for that constant term or else the "correlations" should be computed from regression lines with intercept 0.

Thomas N. Herzog (1980). "Multiple Imputation Modeling for Individual Social Security Benefit Amounts" (these Proceedings, 1980).

Donald Rubin (1977). "Formalizing Subjective Notions About the Effect of Non-respondents in Sample Surveys," JASA, 72, 538-543.

Pamela M. Vacek and Takawaru Ashikaga (1980). "An Examination of the Nearest Neighbor Rule for Imputing Missing Values" (American Statistical Association, Proceedings of the Statistical Computing Section, 1980).