nonresponse stratum of the true estimates of the mean, variance, and covariance with the estimates obtained solely from the imputed data set, and (7) comparisons of distributions in the nonresponse stratum--that is, comparisons of distributions generated from "true" data with those generated solely from imputed data. Finally, every effort should be made when constructing the simulation data sets to approximate the actual patterns of missing data in the data base, including various nonresponse rates for analytically important subgroups.

Spruill

The Spruill paper is encouraging because it represents a step taken by two Federal agencies, the Small Business Administration and the Internal Revenue Service, to address an issue of substantial importance to researchers both inside and outside the government. Several years ago the Office of Federal Statistical Policy and Standards in the Department of Commerce sponsored work on statistical disclosure and disclosure-avoidance techniques [5]. Since that time, however, not much has been done in this area by the government agencies concerned about inadvertent disclosure. Developing a research agenda for the three different sizes of firms is a sensible approach to the issue of inadvertent disclosure since the problem of publishing the identity of large firms cannot possibly be the same as the problem for medium and small size firms. I am skeptical of the ability to protect the identity of some large firms even with a statistically sound "contamination" strategy. Overlap between variables released and those in publicly available data is another important way to look at the problem; although in reality the overlap surely must be considerable.

My preference for additional work on this subject is to place greater emphasis on the case of variables not normally distributed, based on my guess that many of the types of variables being considered here are not normally distributed. I suspect that the overlap of variables from file to file is extensive; however, it is my belief that the nature of the variables in common is another parameter which should ultimately be considered in the analysis. The duPont Corporation and General Motors Corporation examples are relevant here--the analyst's knowledge of industry and geography is probably sufficient to identify these large corporations successfully with a high probability; if industry and geography were not available, even in contaminated form, more variables would probably be needed to identify the corporations successfully.

An assumption not stated in the paper is whether a one-to-one relationship exists between a publicly available data item and the administrative data item. It is quite likely that definitional differences exist among the various data sets. An additional assumption not stated explicitly is that the types and extent of errors in the measurement of publicly available data are the same as those found in Federal administrative record systems. These are simplifying assumptions not likely always to be true--the actual data problem has an added degree of disclosure protection. Finally, it would have been useful to have been given a more developed discussion of difficulties associated with analyzing contaminated data because most users would have considerable difficulties. Data-related problems confronting analysts become even more severe if contaminated microdata files are released to the public since the pool of potential users would be extensive, comprising a wide range of academic training.

REFERENCES

[1] U.S. Department of Commerce. Office of Federal Statistical Policy and Standards. "Report on Statistical Uses of Administrative Records." Statistical Policy Working Paper 6. 1980.

[2] Alvey, Wendy and Scheuren, Fritz. "Background for an Administrative Record Census." 1982 American Statistical Association Proceedings, Social Statistics Section.

[3] The issues raised in this discussion are general ones and many of the comments could be made about any Federal statistical data collection sysstem.

[4] Kalton, G. and Kasprzyk, D. "Imputing for Missing Survey Responses." 1982 American Statistical Association Proceedings, Section on Survey Research Methods.

[5] U.S. Department of Commerce. Office of Federal Statistical Policy and Standards. "Report on Statistical Disclosure and Disclosure - Avoidance Techniques." Statistical Policy Working Paper 2. 1978.

REJOINDER

This reply is in response to the discussion given by Daniel Kasprzyk on five papers dealing with methodological research currently underway in the Internal Revenue Service's Statistics of Income and Research Divisions. The authors of the papers would like to thank Dr. Kasprzyk for his many sound and thoughtful comments. As further clarification on the issues he has raised, we have provided the remarks below.

Bahnke-Wheeler

Dr. Kasprzyk's comments on the Bahnke-Wheeler paper offer some helpful criticisms on how the paper might better explain the Statistics of Income processing system, particularly the comment that our discussion of the resources and time needed to complete different processing stages should have been expanded. Also, the topics of studying tolerance levels prior to production and the magnitude of nonresponse, both item and whole unit, in the editing realm, were addressed in our research, but not to any extent of the paper itself. We will try to bring out these issues in the paper we intend to write for next year's meetings. Finally, a discussion of imputation was not included in our paper, because the Hinkins paper covered that topic.

Schwartz

Following are some comments on a number of items that the discussant questioned or felt need additional considerations, namely SOI quality levels, lack of source document use in data correction processing, consideration of item criticality in the development of quality control procedures, and measurement of the quality at the intermediate and the final processing stages.

The absence of quality levels for various processing phases is by no means unique to SOI programs, but is a general occurrence in the production of statistical data from administrative documents [1]. Lack of specific information provided by the data user on quality needs is mainly responsible for this situation. If there are no specified requirements, almost any quality level is theoretically acceptable. A sense of ethical obligation and responsibility and pride in the work will often lead the data producer to implement various quality control procedures (with the general purpose of finding errors and improving the quality) which in a sense results in a certain quality level. This may or may not be adequate depending on how the data are used. It is as likely, as it is unlikely, that too much may currently be done in SOI quality control, but in the absence of designated quality goals, this cannot be determined.

The lack of use of the source document in a number of data correction or adjustment processes poses a number of quality problems, and some of these must be tolerated due to operational considerations. However, attempts are made to keep these problems to a minimum by ensuring as much as possible that manual or computer adjustments made without the source document are procedurely and technically sound. A number of adjustments interestingly enough are made to correct errors on the original document, not in the statistical editing or other processing. For very significant documents, such as very large corporation returns, microfilm copies are made and are referenced in resolving certain error conditions.

The criticality of items is being given more attention in the development of quality control procedures. This will be considered particularly in defining what constitutes a defective document.

The quality of data in the final product (published tabulations) can be determined in several ways and the one utilized will depend on the extent of quality control coverage of the various processing phases and the reliability of the resulting data. A number of past efforts to measure the quality of the final product had to be abandoned because most of the quality resources were spent on controlling and measuring the quality at the intermediate processing phases. If the quality is properly controlled and/or measured at all major intermediate processing phases, the quality of the final product can be derived from these data. However, if there are missing links, a review of a sample of documents in the final computer file is necessary to determine this level. In complex programs, subject to potentially high error rates, the process control approach (which eliminates the need for error measurement in the final product) is generally more cost-effective whereas for simple low-error programs error measurement of the end product is generally more cost-effective.

Harte

I apologize to Dr. Kasprzyk because the paper presented differs in an important way from the paper sent to him. That paper featured a Monte Carlo study which was not discussed today. It was replaced by a discussion of a full scale study based on actual data conducted for the IRS by Westat, Inc. Their study provided better evidence that post-stratification by industry is a promising approach. Our further research will be based on the full scale study of 1979 and 1980 tax year return information.

Hinkins

I certainly agree with the discussant that these preliminary results cannot necessarily be transferred to other industry of asset size classes. The results for only one combination of factors were described in this paper, but the intention is to continue this work as a factorial experiment, time and money permitting. Evidently this was not made clear in the paper. I certainly do not want to leave the impression that this represents the final conclusion of our work in this area. As the discussant mentions, the traditional hot deck procedure is most effective when the nonresponse rate is relatively low. While this is the case in most of our asset and industry classes, there are several classes with nonresponse rates around 50%. For such problem classes, and for classes considered sufficiently important, we do need to consider enhancements of the imputation procedure. We are currently considering the viability of using information from the previous year's (corporate) return.

I would like to thank the discussant for his suggestions of criteria for measuring the effectiveness of our imputation procedure.

While my complete report contains several of his suggested comparisons, I now plan to include several more.

Spruill

I want to thank Dr. Kasprzyk for his thoughtful comments on a longer version of my paper. (The comments concerning the three different sizes of firms and the DuPont/General Motors examples are relevant to the longer version, not the version included in this Proceedings.) In response to his suggestions, I plan to look at non-normal test data (gamma-distributed data, in particular) and at actual IRS business tax data. My paper now includes several references that discuss how to use contaminated data in analyses; however, I need to do more in this area.

CONCLUDING COMMENTS

In conclusion, we would again like to thank Dr. Kasprzyk for his helpful comments, but hope he is aware that progress in any of these areas is difficult and not as rapid as each of us would like to see. Nevertheless, we are committed, and will continue in our efforts, to improve and make more readily available data from the Statistics of Income program, as recognized by Dr. Kasprzyk in his discussion.

REFERENCE

[1] Kilss, Beth and Scheuren, Fritz. "Statistics from Individual Income Tax Returns: Quality Issues." 1982 American Statistical Association Proceedings, Section on Survey Research Methods.