DISCUSSION

Robert E. Fay, U.S. Bureau of the Census[1]
Statistical Methods Div., U.S. Bureau of the Census, Washington DC 20233

The paper of Susan Hostetter, "Measuring Income for Developing and Reviewing Individual Tax Law Changes: Alternative Concepts," explores the problem of what to measure from an administratively based statistical system in order to provide the most effective information for the analysis of policy and as a statistical representation of income in the United States. The careful presentation of concepts and comparison of their implications will surely be of interest to many who use these data.

I am reminded of a sentence in Thomas Pynchon's Gravity's Rainbow: "Roger is only a statistician." I have recalled this sentence on occasion when asked to discuss a well-presented analysis of primarily substantive content. Except for complementing the author on this work, I am unable to add to what she has accomplished. A paper of this sort represents an important resource to the statistical public, and I am pleased that it was included in this session.

The next paper, "Family Data from the Canadian Personal Income Tax File," by Edouard Auger, takes the problem of using administrative data for statistical purposes in a different direction: use of tax data to approximate census concepts in Canada. The census concept of family is one of great substantive importance but also some challenge to derive from these records. The results the paper reports appear quite encouraging, and the careful work it represents is exemplary.

The paper does not consider in detail what would seem to be the next step in the process: the development of models to adjust the tax data further to increase comparability of the results with the census concepts. I presume such work will be considered. I think that the direction represented by the paper, namely to investigate carefully improvements and adjustments to the tax data, is the correct first step to take before resorting to models. I wish the author continued success in his research.

The paper of John P. Hiniker, "The Selection of Returns for Audit by the IRS," represents a solution to the "cocktail party problem," namely how does a statistician describe the profession to someone with limited exposure to statistical analyses? The general solution to this problem usually takes the form of interesting examples, and the selection of IRS returns for audit certainly commands interest. The paper effectively illustrates application of statistical concepts in a practical setting.

The basic approach is sound. To add to the scope of the paper, however, I will mention some possible different analyses that could be considered for this or similar problems.

The method described in the paper, discriminant analysis, was derived from an original assumption that a vector of characteristics, $\mathbf{x} = \{x_i\}$, followed a normal distribution with mean m_g and variance $\mathbf{V}$, for observations from group g. As long as the group means differ from each other, a discriminant function in the form of a linear function or functions (in the case of more than two groups) of the components of $\mathbf{x}$, can be used to classify the membership in the groups. This function can be estimated from sample data from observations for which group membership is known. Under these assumptions, the sample means for each group and the pooled estimate of the within-group covariance matrix, $\mathbf{V}$, are sufficient statistics for the linear discriminant function.

Logistic regression is also related to the problem of classification. In the simplest case, logistic regression represents a model of the form

$$\ln \frac{P(g=1;\ \mathbf{x})}{P(g=2;\ \mathbf{x})} = b_0 + b_1 x_1 + b_2 x_2 + \ldots$$

No restrictions are required on the distribution of $\mathbf{x}$, but maximum likelihood estimation of the parameters of this model generally requires iteration.

Under the assumptions of linear discriminant analysis just described, the logarithm of the ratio of group probabilities, conditional on the observed $\mathbf{x}$, can be shown to have the form of the logistic regression model. Because of the more specialized assumptions of discriminant analysis, the form of maximum likelihood estimator differs from that for logistic analysis, but the two problems have a close underlying connection.

The paper acknowledges the inappropriateness of the underlying assumptions of discriminant analysis for these data, but makes reference to the general robustness of this approach in motivating its application. In my view, this position is reasonable. As a direction for possible future research, however, the logistic model bears investigation.

The paper describes use of likelihood ratios in the development of the classification model; log likelihood ratios are a more appropriate choice for this problem, since the solution would then become invariant under reordering of group membership, unlike the current version. With this modification, the problem becomes closely connected to the Fellegi-Sunter procedure considered in the paper by Drew, Armstrong, and Dibbs.

The two-stage procedure represents a practical adaptation rather than a necessary consequence of the general theory for the classification problem. Its apparent effect is to compensate for lack of fit of the first model in the region of key interest, namely, returns with a high probability of revision from audit. This sort of evidence suggests that improvements in the model may be possible through a different formulation.

It is usually easier to think of additional models for a set of data than to implement them and make them work. Consequently, I will add to the preceding remarks my appreciation of the careful and creative effort in developing and implementing the models and presenting them in this session.

The next paper, "Research into a Register of Residential Addresses for Urban Areas of Canada," by J. Douglas Drew, John B. Armstrong, and M. Ruth Dibbs, represents a detailed consideration

of a procedural issue with important implications. The development of such a register has both costs and benefits that the paper considers in a balanced fashion.

A similar issue arises in conducting the censuses in the United States, where address registers were prepared for most of the housing units in 1980, building upon experience in 1970 and 1960. In the U.S., the registers began in computer readable form in some areas before the start of the census, but the process of additions, deletions, and other modifications of the registers were not captured subsequently. Except for test censuses, the first experience comparable to the Canadian plan nationally will be in 1990.

The matching work described in the paper is methodologically important for administrative records research. Each large matching project seems to encounter its own unique aspects, but there is much of general interest in the work that they report.

The paper "Osculatory Interpolation Revisited" has both aesthetic and methodological merit. The authors, H. Lock Oh, Fritz Scheuren, and Beth A. Kilss, succeed at making this potentially obscure (although clearly important!) topic lively and interesting. The geometric interpretation is both pretty and useful for generating new variations and insights.

As a discussant, I can find nothing to add to the treatment of the problem the authors discuss directly. Consequently, a few tangential observations follow.

The paper discusses the problem for highly accurate counts, based on complete enumerations or very large samples. Additional technical issues may arise when sampling error is a significant consideration, and this issue may represent a direction for possible future research.

The paper also considers administrative data that are presumably essentially free from rounding effects. The use of the additional information represented by the means within the intervals, which is advantageous in the problem considered in the paper, may be less clearly helpful in the presence of substantial rounding effects. This question should be considered by researchers considering application of these methods to income data reported by respondents in surveys.

John L. Czajka blends the issues of sampling and missing data in his paper, "Predicting Edit Outcomes: The Strategic Use of Imputation in Estimating Corporate Income Statistics." His concern is the validity of small domain estimates when missing data strategies usually employed for nonresponse are used as estimators following a sampling operation. The central issue is whether missing data adjustments should be based on information from other domains or only each specific domain in question. The general theme of the paper appears to be to compromise between these options. His paper details the concern with the previous method for small domain estimates. A complete evaluation of the proposed alternatives will be an excellent topic for a future paper. Overall, the presentation is thoughtful and inventive.

Wendy Alvey deserves particular thanks for organizing this session.

Footnote:

[1] The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.