DISCUSSION

Charles H. Alexander

Two of these papers - Biemer&Woltman and Thibaudeau - are about latent class analysis. The Perritt and Crouse paper is an outlier in this respect, but together the three papers illustrate some common themes about the benefits and limits of automation and of statistical models. This has been an enjoyable discussion to prepare, because these are three good papers and because along the way I get to ask the age-old question Awhat is truth?@

The first two papers both make a important contribution just by adopting an explicit probability model for a long-standing and hard-to-pin-down problem. For the reinterview paper, this extends a development that goes back to work of Morris Hansen and his colleagues in the 1960s; Biemer and Woltman have been leaders since then in continuing this development and finding fruitful applications for response error models. The Thibaudeau paper contains an initial statement of a formal probability model for the demographic methods used to produce intercensal population estimates. This is part of a substantial new effort involving statisticians and demographers at the Census Bureau. The need to formally integrate the statistical and demographic approaches to this problem was pointed out in a 1979 paper by Purcell and Kish, but little progress has been made since, so this is an important development.

I=ll quickly give my understanding of latent class analysis, as a background for my comments. In each paper, four categorical variable are observed. For Thibaudeau these are migration status, age, tenure, and a salary indicator variable. For Biemer and Woltman, the observed variable are race on an original interview, race on a reinterview, Hispanic origin, and Dress Rehearsal site. The data of interest are a cross-tabulation or contingency table, giving the estimated proportion of the population with each possible combination of these observed variables.

These data can be analyzed by a hierarchical log-linear model. The model includes main effects for each variable, as well as two-way interactions indicating the dependence or association between any two variables, three-way interactions that measure how the relationship of two variables can differ depending on the value of a third variable, and higher-order interactions representing still more complex patterns of dependency. Leaving any of these interaction terms out of the model makes the assumption that the particular interaction is not present for those variables.

The Alatent classes@ come in with the assumption that there is an additional, unobserved variable X, which has specified interactions with the observed variables. In the Thibaudeau paper, this is assumed because there actually is an important variable, tax filer status, which is unobserved but is known to be associated with the observed variables. In the Biemer and Woltman paper, X is introduced as a hypothetical Atrue race@ measurement, to provide understanding about the differences between the original and reinterview measurements of race, based on their interactions with other variables.

It may seem surprising that the parameters relating to an unobserved, and possibly fictitious, variable can be estimated. This is done by assuming that some of the interaction terms among the original, observed variables are zero, so that patterns which would appear to be due to those interactions are actually the result of the dependency of the observed variables with some unobserved X. From the observed dependencies, one can then deduce what interactions X must have with the original variables in order to produce the observed patterns, and also deduce the probability of the different values (Alatent classes@) that the categorical variable X can take. There is no claim to estimate the actual value of X for specific households or people, but under the assumed model all the relevant aspects of the probability distribution of X can be estimated.

For all of us trained as statisticians, there is a temptation to believe that once we have assumed a model to be true, then reality must follow it. However, there are two fundamental uncertainties about the latent class model, which must be kept in mind:

1. Even if we have in mind a particular meaning for X, in reality X includes the effects of all variables that have been omitted from the model. So even if the dependencies are due to unobserved variables, the X may not be what we think it is.

2. The latent class analysis is nothing but a way to interpret the dependencies among the observed variables. What we attribute to an unobserved X may be only a higher order interaction among the original variables, which we erroneously assumed to be zero.

Note that a good fit of the latent class model does not rule out the second possibility. The successful test of fit does show that once X has been included in the model, any omitted interactions are not significant. But it cannot prove that X is the true explanation, rather than the interactions that would have been significant if X had not been included in the model. It may be a useful exercise to see what those interactions would be, and see if there are substantive grounds for judging them less plausible than X.

Thibaudeau admits his temptation to believe his model is truth. It=s tempting because both the relative frequency of the two latent classes, and the deduced patterns of associations, do resemble what is known about filing status. But he virtuously resists temptation and emphasizes that we cannot make formal inferences about tax filers. Nevertheless, his ultimate conclusion is persuasive: it is not safe to base intercensal estimates on a simple model using data that exclude tax filers. He shows that if the two latent classes do overlap heavily with filers/non-filers, then there would be a serious bias due to leaving out filers. Even if that=s not what=s going on, the results of his analysis show that the model needs to include additional variables, or higher-level interactions among the observed variables, before it can adequately describe the population.

I think Biemer and Woltman do lead us into temptation, to excessive belief in the model. The problem is the clearest when considering the estimate of bias in the interviews and reinterviews. In one of their tables, there is a 3.4 percentage point difference between the initial interview and the reinterview as far as the percentage of people who are Asome Other Race@. Their analysis concludes that the interview is 2.3 points above the truth and the reinterview is 1.1 points below the truth. At this point we must ask Awhat is truth?@ How can they know?

If the paper had dealt with a variable such as salary, where there really is a Atruth@ which an accountant could establish, the temptation to claim a knowledge of the truth would be less. It would be clear that the latent class analysis, looking only at interview and reinterview results, and Hispanic origin and Dress Rehearsal site, could not tell us what the accountant would find. However, since race is defined by self-identification, there is no external truth, so what is wrong with defining truth to be the X from the latent class analysis? The authors reasonably observe that X is Adevoid of influences that would cause instability in responses to the race question,@ at least as far as the influences included in the model are concerned.

I see two concerns:

1. This Atruth@ may be very dependent on the choice of the auxiliary Agrouping@ variables, as well as the assumption about which interactions in the models are zero. For example the estimate of bias in race in Sacramento might be quite different if different sites than the two from South Carolina had been included in the analysis.

2.This truth is not the result of a Apreferred method@ in the sense that it has anything to do with a more valid way of asking the race questions.

I don=t want to seem negative about the overall paper, based on this relatively narrow concern. The potential value of latent class analysis for analyzing reinterview data is clear. As the authors note, it avoids making the assumption of equal distributions needed for the old method, and by introducing the additional grouping variables it may reduce the effect of nonhomogeneity as the estimates of the reliability R. This is valuable work that obviously should continue. However, I do think that the authors need to define more carefully the concept of bias that their analysis yields.

The empirical results in the paper about the unreliability of race for Hispanic respondents, as well as ASome Other Race@ and AMore than One Race@ respondents, are dramatic. They foreshadow what was found in the subsequent comparisons of Census 2000 race data to data collected using American Community Survey methods, namely that slightly different data collection methods dramatically affected the race response. Next time we need to apply the authors= methods earlier in the decade as part of testing the race question.

In the latent class analysis papers, we see the strengths and weaknesses of modern computational approaches. High-speed computers make it easy to use methods to fit a variety of models, which is good. However, not having a closed- form solution makes it harder to visualize what relationships of the observed variable lead to certain conclusions about the unobserved Atruth@.

In the Perritt and Crouse paper, we see mostly advantages of automation in the way it is being applied to the Agriculture Census processing system, including automating and standardizing edit, imputations, and checking of the data. This is an idea whose time has come, and it parallels efforts at the Census Bureau, Statistics Canada, and other U.S. and international survey organizations.

Unlike the previous two papers, there is not a formal probability model for any of the these analyses. This is understandable, since those papers looked in depth at a few variables, while the system for the Agricultural Census must handle a huge number of variables, without time for studying the theory for any one variable in great depth. Still, I think we can all aspire to a future when we have explicit theoretical models for the edit and imputation methods that we use in production. The steps toward standardization that are evident in the Perritt and Crouse paper are steps in that direction.

They use a number of different techniques to detect and fix problems in their data, choosing the techniques based on what information is available in different situations. They use deterministic Decision Logic Tables when there is information to be sure about obvious errors; these errors are fixed based on the Decision Logic Table, or by using previously reported data. When that fails, they localize errors by checking linear consistency constraints, when such constraints can be specified. These errors are fixed deterministically if the constraints imply a unique solution, and otherwise the system fixes the problem by imputation from a direct donor or the donor of a ratio.

After the automated clean-up is done, a less structured review by subject matter experts is used for things where analytic judgements can detect errors in the data, helped by a Data Review Tool. These errors are fixed by analytic judgements. Special attention is paid to highly influential observations.

The paper does not explain specifically how the system developers identified which variables are treated by which of these methods. Further discussion of this would be helpful. Reading between the lines, the approach presented in the paper appears to have been a compromise between one camp that would have stayed with Amanual reviews of all reported data by subject matter experts as the primary means of data editing@ if that had been feasible, and another camp that would have banned manual reviews in favor of Athe automated minimal change philosophy@ if they hadn=t been thwarted by the barrier of Acultural acceptance.@ I think this actually is a fortunate compromise, which is better than either of the extremes would have been. The automated portion of the compromise approach adds a valuable orderliness and standardization to the process, while still providing the flexibility to take advantage of information and insight that is available in specific situations.

Something the authors obviously confronted, but do not discuss, is the training of the subject matter experts in using the data review tool, as well as the process of feedback from these experts into the development of this tool. This standardization of the procedures for data review (macro or micro) is something we at Census are finding very important, not to limit the exercise of judgement and insight but to share the best practices among analysts. This is especially important as some of the Aold hands@ in the federal statistical system near retirement.

For future work, these are some generalizations that might improve the performance of these methods. The strict sequencing of models limits the ability to use information about Alater@ variable to improve the edits of Aearlier@ variables. The Euclidean distance measure may not work as well for categorical variables, especially zero-one variables, as for continuous variables. There may be value in looking at non-linear constraints, especially when combining ratio variables with direct variables. However, these things are easier to suggest then to implement. The NASS methodologists have set their sights high already, and will have made a major advance with the successful implementation of the methods described in this paper.