

USDOT-Census Bureau -- Disclosure Review Board Meeting

January 27, 2006

Attendees: Phil Steel, Marie Pees, Celia Boertlein, Clara Reschovsky (Census Bureau), Neil Russell and Mike Cohen (BTS), Elaine Murakami, Ed Christopher, and Nanda Srinivasan (FHWA), Ed Weiner (OST)

Ed Christopher called the group to order and thanked the Census Bureau (CB) staff for the opportunity to meet again. He then reviewed the following agenda. It should be noted that the agenda was not distributed prior to the meeting so the opinions expressed are only the initial reactions on the part of the CB , rather than a considered opinion.

Topics:

1. Follow-up on the memoranda on September 13, 2005 and December 13, 2005
2. Disclosure Risk
3. Potential Joint Investigative Team
4. Other Issues

1. Follow-up on the memoranda dated September 13 and December 13, 2005

Q1. Will the “rule of 50/300 apply to all workplace tables or just the ones using “mode to work”?”

Phil Steel said that the Census Bureau and DRB was concerned with the “mode to work” tables because it is crossed with so many other variables. Ed C suggested that consistency would suggest that same rules be applied to all variables and tables

Phil S noted that the 50-weighted or the 300 unweighted rules for 5-year data are **only intended for Place of Work (POW) tables**. Areas that are primarily a residence area with a few workers are a disclosure issue for POW. Elaine M said that 17 percent of tracts are affected by the rule of 300, even though these tracts only represent 2 percent of workers. Phil S. said that the Census Bureau could revisit the rounding rule, once they have a track record for ACS weighting but speculated that they would not it sufficient to discard the threshold.

Q2. Does collapsing and suppressing apply to residence tables AND/OR to workplace tables?

At first there seemed to be some disagreement over this. Celia B said that it applied to just work place tables but then someone pointed out that the memo did in fact say it applied to “mode to work” tables for both the residence and workplace tables. Ed C asserted that it looked like the CB was discriminating against the “mode to work” tables.

Celia B went on to talk about collapsing and suppressing: Celia B said that there are two trials of Coefficient of Variation (CV) tests applied to the base tables:

1. Complete distribution
2. Collapsed version

Celia B. said that CB has not yet decided the application of collapsing and suppression rule for the 5-year average. The test for the CV is applied to the complete distribution first and then to the collapsed version, and if the test does not pass, the table is suppressed. Along with the CV test, the rule of 3 un-weighted records for the mode to work variable would apply both at the residence and the work end.

Elaine said we may need to have different combinations of collapsing so that urbanized areas and areas with complex transit are able to obtain mode data, because otherwise, public transit is more likely to be suppressed.

It was reiterated that the concern with the “mode to work” question was the many variables with which it was cross-tabulated. Celia B explained that the cross tabs for the mode to work tables were based on subject table design. Nandu S noted that it was crossed with some 22 other variables. Elaine M pointed out that educational attainment was crossed with 17. Ed C asked if there was a point or minimum number of crosstabs that could be done that would avoid the collapsing schema. Celia B (I think it was her) suggested that it was not just transportation who had interest in mode to work crosstabs

Elaine M asked if disclosure rules would be reconsidered if some of the cross tab variables for mode were reduced. Elaine M listed a few variables of interest to transportation planners including earnings, poverty status, occupation, industry, class of worker, travel time, and vehicles available. Phil S. said reduction in the number of variables cross-tabulated with mode to work may reduce the need for the rule of 3. No resolution was reached and Ed C suggested that transportation take the question back and further analyze the issue.

Celia B., Phil S., and Marie Pees said that there will be a marker to show data that don't pass the rule of 50 or 300.

Q3. What does the phrase “otherwise the data must be collapsed or suppressed and complementary suppression must be applied” mean?

Celia B pointed out and both Marie P and Phil S agreed that there would only be one pass through the collapsing. It was explained that the test would be run once and if a variable did not get collapsed on that run it would be suppressed. Ed C said he didn't understand why there was only one attempt on collapsing and that would be applied across all geographies within a given area. He said that differences in transportation services across different cities made one collapsing schema not useful. Ed Weiner then asked if the collapsing order would be known or released before it was done. The CB has not yet determined the collapsing structure.

It was also explained by Celia B and Marie P that complementary suppression was where the next cell with the smallest non-zero value would be suppressed in addition to the original cell that was suppressed, to avoid disclosure by subtraction. In other words, every time one cell in a table is suppressed another one that would have otherwise passed the test would also be automatically suppressed.

2. Disclosure Risk

Elaine M. presented some work on calculating disclosure risk. She went through a few examples where she believed that the risk of disclosing an individual should be reduced as one cumulates records over time. In the first example, Elaine pointed out that ACS is a sample of housing unit addresses, but the number of households and persons for any given geographic unit in a five-year period is much larger than in a point-in-time sample. For example, a census tract with 2,000 workers in each year, might have a sample frame of 2,800 workers in a 5-year period (assuming 10 percent of workers move every year). If the number of workers sampled at the rate of 1.4% is 28 workers for year, the sample size is $140/2800 = 5$ percent. The decennial census sampled at the rate of 17%, and we are now down to 5 percent.

Compounding the fact that populations shifts occur over time, the actual information is aging “right before your eyes.” For example although age will be reported in cohorts like 35-44 the actual year of birth of the people in sample would be in a 15 year range, instead of 10. Elaine’s belief is that not only is the universe of people highly fluid over a 5 year period but behavioral variables like travel mode to work is changing too. Ed C wondered out loud at what point the risk would become so small that for 5 year small area data we could stop all this disclosure proofing and begin to publish the data.

Work will continue on this front. Elaine is working with BLS to examine National Longitudinal Survey of Youth (NLSY) data to examine the percent of people who change both residence and workplace in a 5-year frame.

Phil S. appreciated Elaine’s efforts and said the CB could use the information on the aging of the data on negotiating the rule of 50 or 300, but the rule of 3 would not be affected by this argument. The data span affects the nominal size of the geography – because of migration, the number of individuals to whom the data might belong is greater than the population at any one time. The 50 and 300 thresholds are geographic, so should decrease some from this viewpoint though there is still the problem that migration affects some areas more than others. The threshold of 3 is a rule addressing table dimension: a one in a frequently crossed variable allows the characteristics to be strung together, in effect merging parts of the tables. Two is thrown in so that one of the two cannot subtract out his/her characteristics to reveal construct the remainder.

Elaine M. said that DOT is trying to get Nanda to examine the microdata and evaluate the size of geography for which flows can pass the disclosure rules, and believes we need to consider synthetic data approaches for small geographic tabulations.

- ACTION:** 1. Elaine will continue to work with BLS.
2. Nandu will continue to work on TAZ aggregation issues.

3. Potential Joint Investigative Team (JIT)

Mike C. explained that for CFS, the US DOT and the CB have established teams to jointly examine the research leading into the 2007 CFS including development of the questionnaire, methodology issues, and adding noise to the CFS data to avoid disclosure.

ACTION: Clara will work with Phil Salopek and investigate the next steps to establish a potential JIT.

4. Other Issues

Phil S mentioned that the RDCs might be a place for transportation to look when considering examining the ACS data. At that point the entire DOT contingent let out a groan and Ed C explained some of the past history with working with the RDC's where one project for \$25K was only able to produce a working paper on why the researcher could not get into the DRB (not the intent of the project) and another, working through a different RDC, has now been trying for over 20 months to get through the door. Ed offered that the stumbling block seems to be in how the CB is interpreting the rules when have the researchers have to show "how the study would improve the CB process." Marie P said that she thinks the new ACS staff and others are working on how this is being interpreted.