THE WALL STREET JOURNAL.
WSJ.com

THE NUMBERS GUY | FEBRUARY 6, 2010

Census Bureau Obscured Personal Data -- Too Well, Some Say

By CARL BIALIK

Errors in some U.S. Census Bureau data are sending researchers inside and outside government scrambling to check whether some key findings need to be reassessed.

After the Census Bureau compiles overall counts in its decennial population surveys and other studies, it releases additional details about respondents to outside researchers. But in order to protect respondents' privacy, the bureau masks some of the personal information in these so-called microdata.

Now, a study has found that the agency went too far hiding individual identities, introducing errors that might lead economists and demographers astray.

By relying on the microdata, researchers would have found, for example, evidence of a steep drop-off in marriage rates for women at age 65, or of a big rise in the proportion of women in their early 70s who are working—both false conclusions.

The anomalies highlight how vulnerable research is to potential problems with underlying numbers supplied by other sources, even when the source is the federal government. And they illustrate how tricky it can be to balance privacy with accuracy.

The Census Bureau, which needs Americans to cooperate with its survey, is understandably leery about privacy breaches.

"We have this tension in our lives," says Robert M. Groves, director of the Census Bureau. "We want to preserve confidentiality, and we want to maximize utility of our data. This tension is inherent in everything we do. We're always talking about it."

No decennial census—the big survey that is used to allot congressional seats and allocate government spending—is affected by the problems, because the once-a-decade counts contain only aggregated data. The agency publishes breakdowns by age, gender, race and other factors, but the subgroups are so large that in most cases no one respondent can be identified.

Many researchers, though, want to dig deeper into the wealth of information collected by the census every 10 years, as well as in the monthly American Community Survey and Current Population Survey. To help them, the bureau releases microdata, which are a subset of all survey responses. Thus researchers can study the income of, say, married, 65-year-old women in North Dakota.

But slicing the data so thinly raises privacy concerns. Suppose there is only one 65-year-old married woman attending college in North Dakota, and that her response was released by the Census Bureau. Then researchers would know everything else she told the agency, including, perhaps, her income and

her parents' birthplace.

To protect the privacy of such unusual individuals and households, the government manipulates data, using several techniques that were described in a 2005 Census Bureau paper. Numbers are rounded, so incomes of $80,600 and $81,400 would both be recorded as $81,000. What statisticians refer to as "noise" is added to some ages—a year or two older or younger, perhaps.

Also, outlier values are averaged together, and that average is assigned to every one of those outliers. For instance, the top half-percent of earners would each be assigned the average income of that wealthy subgroup, so that, say, Warren Buffett's census questionnaire can't be identified. And people with especially unique characteristics might be moved across the country, in a kind of statistical witness protection program, so that entry for the North Dakotan woman might be changed to show her living in Alabama.

When using microdata from the Census Bureau, economist Betsey Stevenson at the University of Pennsylvania's Wharton School stumbled upon what seemed like an intriguing trend: The marriage rate for women at age 65 was 50%, eight percentage points lower than at age 64. She wondered what might explain such a precipitous drop.

"Then I realized this is not a matter of women changing behavior," Prof. Stevenson says. "This is a problem with the data."

A broader look at other census data sets found similar problems, such as miscounting men and woman at certain ages above 65. In a paper released this week by the National Bureau of Economic Research, Prof. Stevenson and co-authors J. Trent Alexander from the Minnesota Population Center at the University of Minnesota and Michael Davern, from the University of Chicago's National Opinion Research Center, argue that the privacy-protection techniques had introduced substantial deviations in the microdata, when compared with overall census counts.

Flawed software programming appears to be at fault. Laura Zayatz, chair of the Census Bureau's disclosure-review board, says code designed to add the statistical noise to the subset of older respondents should have offset those changes with opposite adjustments made elsewhere in the data sample. This didn't happen as it should have, so that ages and other attributes were skewed.

Before the data were released in 2003, the Census Bureau's diagnostic tools flagged the problem, but it "didn't seem large enough in the judgment of our analysts to stop the release," says Dr. Groves, the Census Bureau director.

Since the research by Prof. Stevenson's group emerged, the Census Bureau has been re-evaluating policies on disclosure risk and diagnosing errors in data before they are released. "I'm going to ask for a very careful cracking of this code so we find out what data are possibly affected," Dr. Groves says.

Researchers elsewhere are revisiting their work to learn how it might have affected their findings. Though the Census Bureau had released notes to data users addressing some of the problems, the agency hasn't yet corrected all data sets, and some users of the microdata were unaware of the problems before this week. Representatives for the Social Security Administration and the Office of Management and Budget—agencies mentioned as heavy users of microdata in the paper—said they are checking to see if any of their findings were affected.

J. Michael Brick, a statistician at research company Westat, says the problems could alter opinion polls of older Americans that are weighted according to the microdata. "There could be some very misleading

analysis," Dr. Brick says.

University of Utah sociologist Claudia Geist says that she used some of the affected data in a 2008 paper she co-wrote about geographical mobility. So far, she hasn't found any problems, but she adds, "We will continue to examine the data, because this is a wake-up call to all who use these data."

The findings raise a broader worry: that even when the numbers in microdata match the broader census, all the swapping, rounding and other adjustments made to protect privacy might obscure links researchers are examining, such as between income and marital status.

And some researchers think the agency's efforts to protect privacy might not be worth the trouble, because identity thieves and others who hunt down personal data with illicit intentions might be able to find the information elsewhere.

"There is nothing in the regular census...data that could not be learned far more easily than by using supercomputers and data mining to try to hack into the census," says Steven Ruggles, director of the Minnesota Population Center.

Write to Carl Bialik at numbersguy@wsj.com