

CSAC, April 16-17, 2015

Discussion: "Big Data and Modernizing Federal Statistics:
Update" by Bill Bostic and Ron Jarmin

Noel Cressie

National Institute for Applied Statistics Research Australia (NIASRA)
University of Wollongong, Australia
(ncressie@uow.edu.au)



- Goals.
- Research Agenda.
- New Center at Census Bureau.
- New Institute with core members at University of Michigan.

- We cannot (and should not) obtain characteristics (including space-time coordinates) of every person in the US.
- Governments need timely, *aggregated* information to make decisions for its population and sub-populations. Businesses need the same information in order to be competitive and react to the marketplace. The amount of aggregation depends, *inter alia*, on the sub-populations of interest.
- “Representative” data are collected for the sub-populations, mandates are met, services are provided, and planning decisions are made.
- The goal is still to estimate economic and social characteristics in the presence of uncertainty. “Big Data” may or may not reduce that uncertainty.

- The fog of “Big Data.” Size matters but so does noise, missingness of individuals, and missingness of variables.
- Design principles: “Stratify, Cluster, Randomize” to get at the signal. Add “Aggregate” to this list, to help clear the fog.
- Computational considerations: Data archives are distributed; analytics are done at data nodes; Moore’s law needs parallelization; memory size constrains analyses.
- “Big” can also mean “Many” datasets. Confidentiality applies to the whole and should not be treated piecemeal.

“What we measure affects what we do. If we have the wrong metrics, we will strive for the wrong things.” (Joseph Stiglitz in “Towards a better measure of well-being,” Financial Times, September 13, 2009)

- Our metrics have been high accuracy (i.e., small bias) and high precision (i.e., small variance), within a cost constraint.
- Have our metrics changed? Do we now want low cost within a quality constraint (e.g., bias/variance)?
- While the cost of “Big Data” is going down, are they complete and to be trusted? Clearly, no!

- Aggregation for “Big Data” can be a powerful tool.
- Aggregation forces down variance, but it does not affect bias.

Suppose there are n independent estimators for θ : For $i = 1, \dots, n$,

$$\hat{\theta}_i = \theta + \text{bias} + \varepsilon_i, \text{ where } E(\varepsilon_i) = 0 \text{ and } \text{var}(\varepsilon_i) = \sigma^2.$$

Then the aggregated estimate is, $\hat{\theta} \equiv \sum_{i=1}^n \hat{\theta}_i / n = \theta + \text{bias} + \bar{\varepsilon}$,

where $E(\bar{\varepsilon}) = 0$ and $\text{var}(\bar{\varepsilon}) = \sigma^2/n$. As n increases,

$$E(\hat{\theta}) - \theta = \text{bias (fixed)},$$

but

$$\text{var}(\hat{\theta}) = \sigma^2/n \text{ (decreases)}.$$

- *Sample-based* quality metrics such as bias and variance seem to be waning in importance for Federal Statistics. Their *model-based* versions, particularly those based on hierarchical statistical models, are becoming more prominent.
- Just because we have “Big Data,” it does not imply that we have removed the uncertainties surrounding the question being answered. If n is large, it does not imply it is large for the sub-population of interest; and even if it is, it does not imply that the mean-squared-error metric,

$$E(\theta - \hat{\theta})^2 = (\text{bias})^2 + \text{var}(\hat{\theta})$$

is necessarily small.

- Formulate goals, then formulate priority areas to work on, then formulate problems within areas. Take a look at NSF's solicitation "Critical Techniques and Technologies for Advancing Foundations and Applications of Big Data Science & Engineering":
http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767
- The Center's success will depend on whether it is, or is not, another "unfunded mandate" for those working in it. It should be resourced with serious FTEs.
- Try this:

Center for Socio-Economic Analytics and Data Science

or

CSEADS

- IMI is an agreement (through an MOU) between the University of Michigan (and others) and the Census Bureau.
- The work will be done by UM's institute, IRIS, and Census' nascent center on "Big Data."
- Early research seems to be driven by "can we?" For some questions, it appears that we can. What have we learned, and what do we still need to learn?
- The new center's Chief will need to structure the IMI to give focus and strategic benefit for the Census Bureau.

- This project is still being conceptualized, but the goal for this project is clear: Supplement existing surveys to obtain small area estimates more frequently. How “small” and how “frequent”?
- Inference will almost certainly have to be model-based.
- What methodology exists? What are the expected “Big Data” roadblocks? New and exciting statistical methodology will clearly be needed.

- It is a great initiative!
- The appointment of its Chief should be done tomorrow (or sooner).
- Projects need to be chosen strategically.
- It needs to be staffed with a serious number of FTEs.
- CSAC has a place in it, through the nascent Big Data Working Group.
- Uncertainties still need to be quantified; “Big Data” offers new environments where this (uncertainty quantification) is essential.