Comments on
“Methodological Issues and Challenges in the Production of Official Statistics”

Lawrence D. Brown

Statistics Department, University of Pennsylvania

The author’s work on these comments was supported in part by a grant from the National Science Foundation.

General Discussion:

Danny Pfeffermann has had a notable statistical career that spans (what he calls) the ivory tower of academia and the day-to-day real world organizations that produce official national statistics. This paper provides interesting and useful perspectives on the production of official statistics from both these points of view. Within these perspectives is a common theme of adjusting to and coping with the contemporary demands and challenges of 'big data'. Here, "big data" (="massive data") or is a catchphrase for what is elsewhere, and more accurately, called "organic data" (see Groves 2011) or opportunistic data or diffuse data.

From my understanding of the realm of Official Statistics, Danny's emphasis on these new types of data is perfectly placed. Official Statistics needs to learn what these data are, or will be in future, how they can be usefully obtained and what to do with them. Many authors have recognized the positive possibilities as well as the dangers in such data. Challenges A through F, treated in Sections 2 through 8 of Danny's paper, add useful discussion and should help along the difficult path.

One important challenge in coping with these new data sources is to classify them according to characteristics that are helpful in also managing their use and analysis. Section 2.1 describes some of the varied types of such data. Table 5.1 of Citro (2014) provides a more extensive typology that is consistent with what Danny has included here.

Another important challenge is measurement of bias, an issue briefly discussed in Section 2.4 and 2.5. Indeed, for many newly emerging data sources and consequent analyses, the classical definitions of bias and variance, while still useful, need to be re-thought and probably re-defined. Official Statistics should strive to produce reliable information that meets the needs of society. "Bias" in a technical sense involves a consistent tendency to lean in a particular direction from what should be the ideal, accurate measurement relevant to that societal need. As Danny notes, one way to try to judge this would be to benchmark estimates from new sources of data against more traditional estimates; as would be the case if the BPP were used to predict the classically produced CPI. This presumes that the classically produced CPI is itself an unbiased indicator of what it should be measuring. But extensive investigation and experience may be needed to know whether this is so. Perhaps the new measure might perhaps ultimately fill the societal need better than the classical one even if in certain circumstances it has a consistent tilt in some direction or the other.

Both Groves (2011) and Citro (2014) as well as many others have noted the evolutionary nature of societal statistical analyses. Concerns about big and yet bigger data organically obtained and processed are not new. The statistical world has met such challenges previously. We can realistically hope that it will continue to find ways to meet the new challenges and take advantage of the new opportunities. Danny's paper is a useful step in this direction.

I will conclude this brief discussion with an example from my own limited experience. This example relates to the theme that "official" data must be relevant and unbiased for the societal purposes for which it is intended to be used. It is also an example that shows how 'big' data and statistical analysis beyond adjusted means and standard deviations have been used for some time. (It also demonstrates that this big data need not be officially produced or analyzed in order to serve official purposes, although this is not a theme Danny has pursued, or that I would like to feature.)

House-Price Indices:

In the U. S. the most widely used index of house prices is the S&P Case-Shiller Index. (Actually, there are several related indices – a national index, a 20-city index, etc.). It's based on a large and reasonably timely data collection that involves most house sales in many selected areas. This data collection is nevertheless some sort of sample, not a census. Also, houses sold are not a random sample of houses in the area or even of houses for sale in the area; but this is the data that's available. The data is then organized and processed through a mathematical algorithm and prepared for monthly release (The monthly figures involve a 3-month moving average adjustment.) It's not a direct area-adjusted sample mean. Broad aspects of the algorithm are publicly available, though so far as I know smaller, though important, details are not publicly accessible. (In fact, this is a repeat-sales index; only houses that have previously been sold are included in the direct processing. First time sales are not included in the index, though they are included in the data-base. Nagaraja, Brown and Zhao (2009) suggests some modifications to this basic algorithm and allows for inclusion on a suitable basis of first time sales.)

There is another U. S. house price index that is an Official Statistic. The U. S. Census Bureau issues it on a regular basis. It's based on a traditional type of sample survey of house prices. This involves a much smaller data collection than the 'big-data' Case-Shiller data base. The Census Bureau prepares and issues statistical output via a version of the usual sample-mean paradigm. (Other competitors to S&P Case-Shiller, such as Zillow also collect and process real-estate data including sale prices, and produce informative output. According to a personal claim to me from a Zillow executive, the Zillow data-base is much larger than that involved in the S&P Case-Shiller indices.)

From 1975 to 2000 the two indices told a very similar story. They tracked each other fairly faithfully. See Figure 1. (C-S is generally a little lower on this plot than the Census index. But note that the base year here is 2000. If the base year had been 1975, then the C-S index would have tracked a little above the Census index throughout this period.)

But from 2000 to 2015 the two indices more noticeably diverged. The Case-Shiller index seems to be telling a much more useful story of the housing bubble that peaked in 2006-2007 and crashed thereafter. House prices are better understood from this big data collection, even though it is not a random sample and involves more sophisticated, indirect statistical analyses than the more traditional Census Bureau survey.

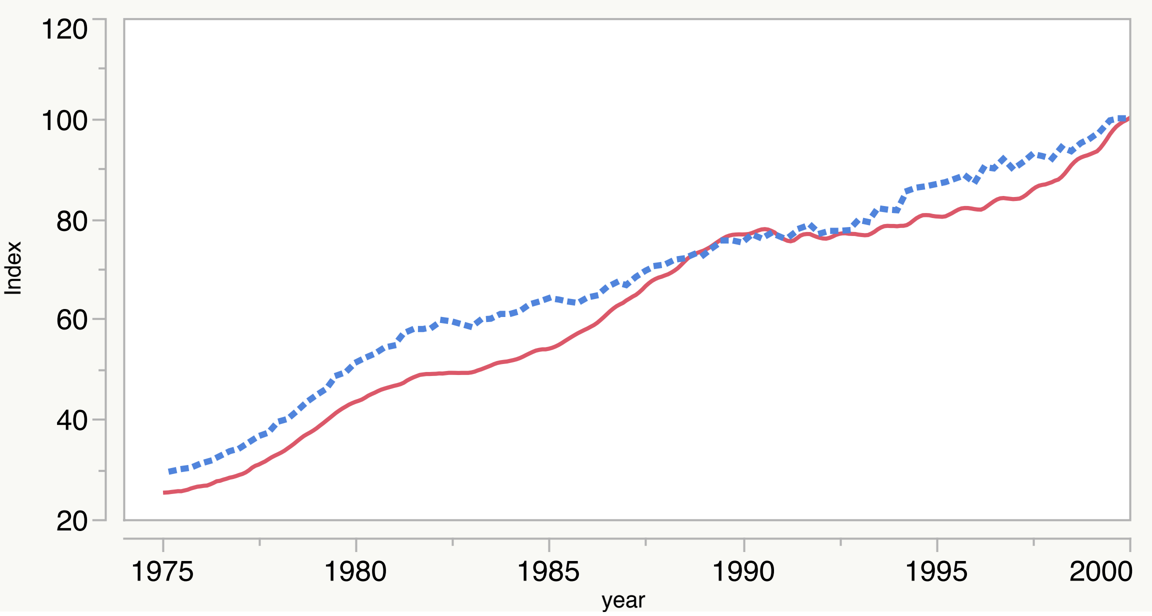

Fig 1: **Case-Shiller** (solid curve) and **Census** (dotted curve) Indices from 1975 – 2000. Base here is 100, for the year 2000.

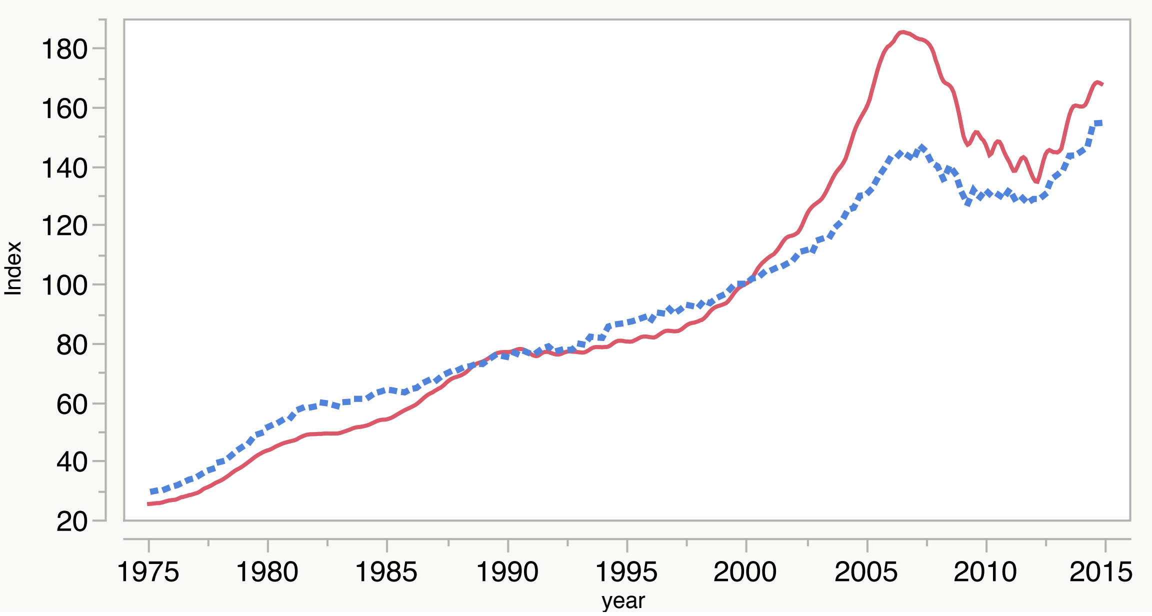

Fig 2: **Case-Shiller** (solid curve) and **Census** (dotted curve) Indices from 1975 – 2015. Base here is 100, for the year 2000.

References:

Groves, R. M. (2011), Three eras of survey research. *Public Opinion Quarterly* **5**, 861-871.

Citro, C. F. (2014), From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology* **40**, 137-161.

Nagaraja, C., Brown, L. D. and Zhao, L. (2009), An autoregressive approach to house price modeling. *Annals of Applied Statistics* **5**, 124-149.