Forthcoming, Annals of the American Academy of Political and Social Science

First Draft: Standards and Guidelines for Combined Statistical Data

Nancy Potok

(May 2, 2017)

A general consensus has formed within the federal statistical data community thata proactive approach is needed to make use of new sources of data to supplement data collection through random sample surveys and censuses . The role of the Office of Management and Budget is central to ensuring that the high quality and value of federal statistical data is maintained.

It is clear that radical change is needed. Data collection from surveys is challenged by less cooperation from survey respondents, resulting in both lower response rates and greater expends. Data users want more complex data faster and at lower levels of geography without sacrificing quality. However, federal agency budgets are not keeping pace with the increased costs of addressing these issues. As a result, major, technology-enabled methodological changes, requiring additional research and development, need to be incorporated quickly into ongoing data collection and production efforts.

Many other sources of data could be used. These include administrative records (federal and state records collected for program administration), commercial data (such as aerial photos, aggregated credit card records, etc.) and open source data (such as consumer goods prices scraped from web sites). These data may be combined with survey data, used in lieu of survey data, or used to create new products. Although some statistical data have a long history of coming from multiple sources, such use of combined data is still being developed.

The U.S. Office of Management and Budget (OMB) has the statutory responsibility under the Paperwork Reduction Act to assure the integrity, objectivity, impartiality, utility, and confidentiality of information collected for statistical purposes. This is done through the development of policies, principles, standards, and guidelines concerning statistical collection procedures and methods, among other things. To this end, OMB has issues four statistical directives, including Statistical Policy Directive 2: Standards and Guidelines for Statistical Surveys. However, no standards or guidelines have been issued addressing emerging combined statistical data sets, leaving producers and users with little consistent information about assessing the quality of future combined federal statistical data. In order for OMB to issue such guidelines, additional research is needed to examine the different dimensions of quality around combined data. These are briefly described below:

1. Transparency: How were the data collected and for what purposes? What are consistent ways of creating and providing such documentation for the user?
2. Fitness for various purposes: How can a determination be made about whether various data sets are appropriate for specific uses? Data may be

good enough for some purposes but not for others (examples?) Should there be a quality rating system developed to guide the user?

3. Privacy: Data may have been collected for particular program purposes. Are additional permissions needed from providers in order to combine that information with other data for different purposes? When data are available commercially does permission for use come from the vendor?
4. Disclosure avoidance: When data are combined and then used for multiple purposes, what new techniques are needed for protecting privacy and confidentiality, especially when researchers want to replicate research results?
5. Microdata: What type of access can be granted to microdata coming from multiple sources, some of which may be proprietary? How are privacy and legal agreements protected?
6. Ownership: What rights do original owners, including statistical agencies conducting surveys and governments providing administrative records, retain for future uses? What happens when commercial data are procured under a licensing agreement?
7. Quality: How does one measure traditional aspects of quality such as accuracy, coherence, comparability, reproducibility, bias, and coverage when combining data from multiple sources with varied collection methods?
8. Break in series: What is the responsibility and appropriate methodology for the statistical agency to bridge a break in a longitudinal data series when the sources of data are dramatically changing?
9. Risk: What are the mitigation procedures needed for an agency to reduce the risk of discontinued availability of commercial or other data that it is acquiring but not responsible for collecting?
10. Post-collection processing: What changes in methodology are needed in production activities such as editing, imputation, weighting, and modeling when data are coming from multiple sources? To what extent do these methods need to be consistently applied?

As research in these areas proceeds, more issues will certainly be identified. They must be addressed if the federal statistical system is to continue as the Gold Standard for providing high quality statistical data. As more statistical data are derived from combined data sources, new research-driven OMB standards and guidance will help assure that data continue to be available for businesses, people, civic purposes, and evidence-driven policy making arising out of federal and state program evaluation and academic research. This research effort will require extensive collaboration between the federal statistical and program agencies, academia, states, and other stakeholders in order to expeditiously advance our learning on combined statistical data.

Such collaboration should take many forms. Intergovernmental projects that bring together city, state and federal partners can be especially valuable, particularly when universities can assist with training, data analytics and meaningful insights.

Many such pilot projects are currently underway, funded by foundations interested in evidence-based policy at all levels of government.

Private sector data providers can work with federal partners in statistical agencies to increase the transparency of commercial data, as well as identify ways to standardize legal and operational approaches to incorporating commercial data into official statistics.

In addition, the Federal Committee on Statistical Methodology (FCSM) should hold workshops on some of these research topics and invite members of the National Academy of Sciences Committee on National Statistics (CNSTAT) and other academics to participate

These are just a few of the opportunities available for collaboration, and OMB can play a critical role in coordinating and corralling the learning in order to develop new, much needed standards governing combined data sets. Although much work is already underway, much remains to be done. Exciting times indeed.