Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues

Discussion Paper

by Steve Landefeld
Senior Advisor to the United Nations Statistics Division

International Conference on Big Data for Official Statistics
Beijing, China, 28 – 30 Oct 2014

United Nations Global Working Group on Big Data for Official Statistics
Beijing, China, 31 Oct 2014

Abstract

This paper provides an overview of big data and their use in producing official statistics. Although advances in information technology, data sources, and methods have driven interest in the use of "big" sets of business and administrative government data collected and used for non-statistical purposes, use of such data is not new. Nor is it likely to be a panacea for statistical agencies confronting demands for more, better, and faster data with fewer resources. However, with careful attention to incentives, protection of privacy, and integration of these non-statistical data with existing statistical data, big data can play a large role in improving the accuracy, timeliness, and relevance of economic statistics at a lower cost than expanding existing data collections.

Table of Contents

Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues

A. The Growth and Potential Impact of Big Data

1. Uses of "Big Data" were recently heralded in a U.S. White House report as "fundamentally reshaping how Americans and people around the World live, work, and communicate."[1] Examples include saving lives through epidemiological research using big data from neonatal intensive care units; tracking the incidence of flu through geographic analysis of Google searches on the use of the word "flu"; making the economy work better through the analysis of delivery truck GPS data to develop more efficient delivery routes; and saving taxpayer dollars by identifying patterns of fraud in medical care claims.

2. Big data have also been described as a transformative tool for official statistics, The statistical community has recognized the potential for big data in improving accuracy and reducing costs for official statistics. In 2014 the United Nation's established a global working group to:

> "provide a strategic vision, direction, and a global programme on big data for official statistics, to promote practical use of sources of Big data for official statistics, while finding solutions to their challenges, and to promote capacity building and sharing of experiences in this respect."

3. Examples of the use of business and administrative for statistical purposes include the "scraping" of internet data to produce the "billion prices" Consumer Price Index; the use of payroll data from the Automatic Data Processing Company (ADP) for its monthly employment index; the use of international postal data to create an International Letter-Post Index that can be used as a leading index or to improve forecast accuracy, and the use of Google searches for "now-casting" of the state of the economy.

4. Several factors have facilitated these advances in the use of big data. Among the more important of the factors are:

- Advances in information technology that have lowered data collection, storage, and processing costs.
- The development of new sources of data and improved access to existing big data sets, on and off-line,

[1] Big data is defined in this paper as the use of large-scale business and administrative data sets that are being used for secondary purposes other than those for which the data was originally collected.

- The parallel development of creative and powerful new methods to exploit "big data."

- The recognition --- through some of these high profile projects -- that we have massive stores of data collected for such purposes as business, administration, health care, meteorology, and traffic that can be used, alone or in combination with other data, for an array of purposes other than those for which they were originally collected.

B. Uses of Big Data for Official Statistics

5. Although a large share of official statistics are based on official surveys, there is a long history of the use of non-survey data in the area of National Accounts, which have been described as a mosaic of public and private data, with most of it originally collected for purposes other than their use in constructing national accounts statistics.

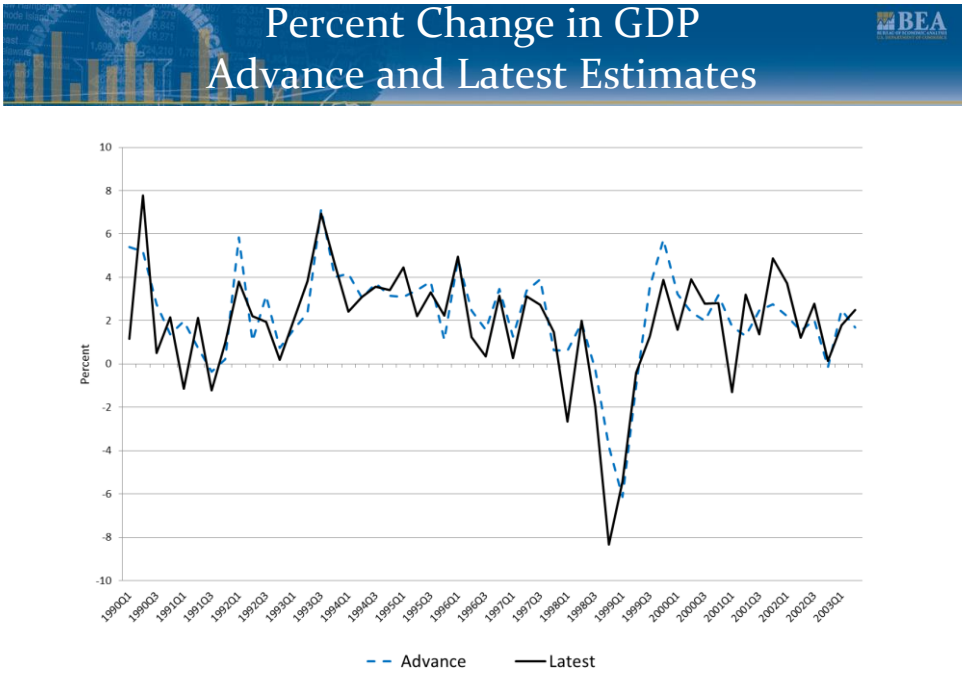
6. Indeed, since their inception, national accounts have used a mix of public and private data to provide a comprehensive picture of overall economic activity that is timely and accurate. The United States and other countries make extensive use of partial data – public and private – as extrapolators for its early estimates. Most of the non-statistical data used have been aggregations of business and government micro-data, although micro-data is used in matching of statistical and non-statistical data to improve official statistics by developing bias adjustments for survey data, improving coverage, and by identifying reporting and other problems.

7. For privacy and other reasons this pattern of the use of big data collected for non-statistical purposes -- as extrapolators and methodological research and improvement tools -- is likely to continue.

8. Because these business and administrative data are collected for non-statistical purposes, they usually do not meet statistical standards in terms of representativeness, concepts, definition, collection methods, etc. To use these administrative and business data national accountants must investigate and understand the statistical characteristics of the data and improve the accuracy of these non-statistical extrapolators through weighting, filling in gaps in coverage, bias adjustments, averaging with other extrapolators, and benchmarking and balancing.

9. For most periods, these extrapolators have worked well and at the same time lowered costs relative to a system of ongoing surveys that collected data designed just for national accounts purposes. As can be seen from Chart 1, the early estimates using "mixed data" provide a timely and accurate general picture of economic activity. The early GDP estimates, based on a mix of public and private extrapolators, released roughly 30 days after the end of each quarter track the later estimates, based on benchmark official data, well.

Chart 1



10. Examples of the source data used for the early GDP estimates are official statistics such as U.S. Census Bureau monthly retail sales, shipments, and inventories data and BLS employment data. All are based on early sample results that will subsequently be revised. Where official monthly indicators are not available, other government and private sources are used. Examples of the private source data aggregations included in the accounts are:

- Ward's/JD Powers/Polk (auto sales/price/registrations)
- American Petroleum Institute (oil drilling)
- Airlines for America (airline traffic)
- Variety magazine (motion picture admissions)
- STR (hotels and motels)
- Investment Company Institute (mutual fund sales)

11. In evaluating business and administrative data for use in national accounts, one of the first questions to be asked is how closely do the data fit with national accounts concepts? A leading example of the impact of differences in concepts is found in the differences in profits using the accounting rules for business profits, tax profits, and economic profits. One would expect that use of the profits from what must be two of the largest "big economic data" bases (U.S. corporate reports and tax filings to Securities and Exchange Commission and the Internal Revenue Service) would make corporate profits one of the most reliable and accurate components of the national accounts. As it turns out, they are one of the most volatile, most revised components of national accounts.

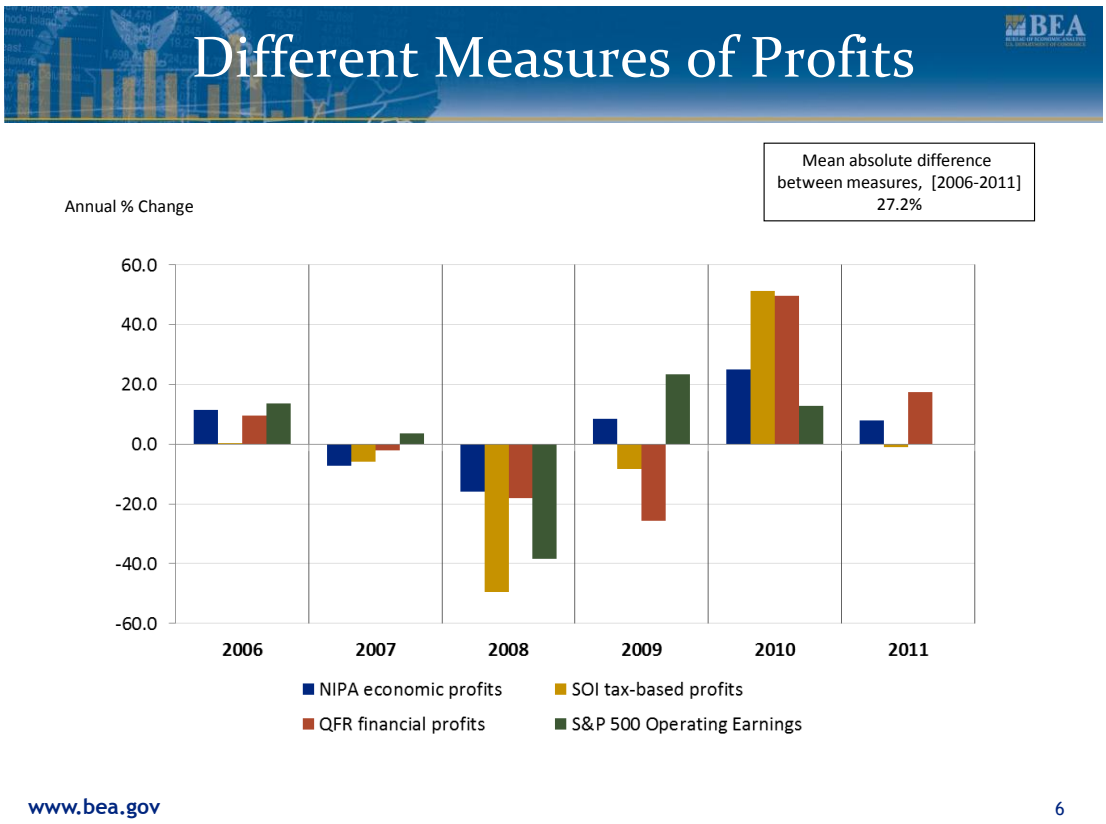
12. Business profits are based on rules such as those promulgated by the U.S. Financial Accounting Standards Board (FASB) and those laid out in the Internationals Financial Accounting Rules (FARS). Profits reported by those same firms to tax authorities use tax accounting rules that include incentives for investment, such as accelerated depreciation, or investment tax credits. Economic profits adjust for inflation and accelerated depreciation to value inventories and depreciation at their "true" replacement cost, and deduct capital gains and losses so that profits reflect the profits earned from production in the current period. Also, the coverage in each of the data sets differs.

13. As can be seen from Chart 2, the differences between growth rates in the three measures can be very large. Although for must years the different measures produce similar changes, periods where there are differences, the differences are quite large. In 2009 the different measures produced estimates that ranged from plus 20 percent to minus 20 percent. Further, tax returns and profit reports can be revised after the initial filing to reflect carry-forward and carry-back provisions for such items as operating losses, research and experimentation credits, as well as revisions based on IRS reviews and audits of their initial filings for up to 10 years after the initial report. As a result of these differences, estimates based on these data can be hard to interpret and are subject, even in the aggregate, to large revisions. Revisions to specific industries can be especially large and have a significant impact on industry profits, output, and supply and use estimates.

14. A second question that must be addressed in the use of these big data is the consistency of the time frame in the source data with the time frame for the national accounts estimate. In another example from the United States, data from a large national monthly payroll survey by the Bureau of Labor Statistics, are used by the Bureau of Economic Analysis to estimate monthly and quarterly compensation. One of the adjustments to the payroll data is for timing, which can be especially important when a major strike occurs during the week covered by the monthly payroll survey (or during a week(s) not covered by the survey). Other examples, include difficult detailed micro-data adjustments to state and local and business data from a fiscal to a calendar-year basis.

15. The third issue relates to the representativeness of the external source data and any selection biases that may be present in the data. For example, in the case of financial reports, they are limited to publicly-held corporations, and exclude important privately-held companies as well as the business income of partnerships and sole proprietors. This is a significant problem because the behavior of the included vs. excluded data can differ markedly over the business cycle and by type of industry (retail or services vs. manufacturing).

Chart 2



16. Unfortunately, some of the largest gaps in coverage for the official economic statistics are difficult to fill using publicly available data. Significant gaps in official output statistics such as those in services and local governments are hard to fill because they are in sectors dominated by a large number of relatively small units using an assortment of concepts, definitions, reporting periods, and accounting rules. Small firms do not file public reports or make available anything other than their industry, services offered, and location (items normally found in business directories), or sales advertised on the internet. Because they are the among the items businesses find most sensitive, small firms generally do not post or file information on their sales, prices, and costs. Filling gaps in local government data is also difficult. While counties, townships, and other small governments provide taxpayers and voters with financial reports few are on a consistent or comparable basis (e.g. reporting period, accounting conventions, etc). Similarly, gaps in income statistics are in hard-to-fill areas like small business income, which is often only reported on individual tax returns.

C. Process for evaluating and using new big data:

17. Incorporating new extrapolators is a multistep process. The first step is evaluating the concepts, definitions, coverage, and performance of extrapolators relative to more comprehensive

and consistent and annual and benchmark data. The second step is developing new methods to use new extrapolators, including benchmarking, weighting and combining with other indicators; bias adjustments. These new extrapolators then need to be evaluated relative to existing extrapolators and benchmarks to access their accuracy. The final step is developing seasonal adjustment factors for the new extrapolated data, which is a difficult process given the normally short time series for the new data.

18. One of the major challenges of this process, is the need to avoid, wherever possible, the use of complex econometric techniques. In general, econometric forecasts and extrapolation add little to accuracy and often do not perform as well in large scale forecasting or as extrapolators as simple trend extrapolations. Complex econometric models also are difficult to understand and assess and result in a loss of transparency to users of the data. Further, if the models are complex or produce only aggregated results, there will be a loss of drill down capacity and links to key indicators.

D. Loss of Control in Using Big Data

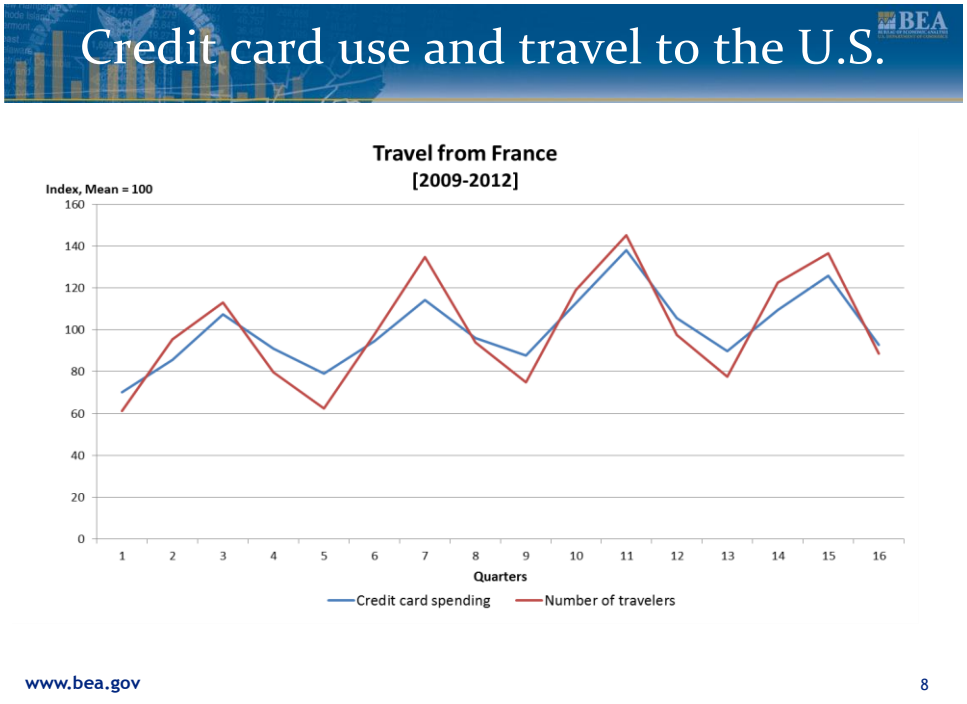
19. In addition to the challenges cited above, the use of big data results in a certain loss of control and dependency on the part of official statisticians. If a company or government agency decides that for business reasons to change definitions, collect different data, or entirely stop their data collections, official statisticians have virtually no leverage to prevent the loss of such data. Although, with increasingly tight budgets, official surveys and statistical sources are subject to discontinuation, official statistical agencies at least have some degree of leverage and control and the ability to reallocate funds to the highest priority statistics.

20. In the United States the challenges associated with depending on business and administrative data has been manageable, but requires having benchmarks and official statistics as the baseline for extrapolations using unofficial data and on the flexibility to change source data and methods relatively quickly.

E. Examples from some preliminary collaborative research and analysis

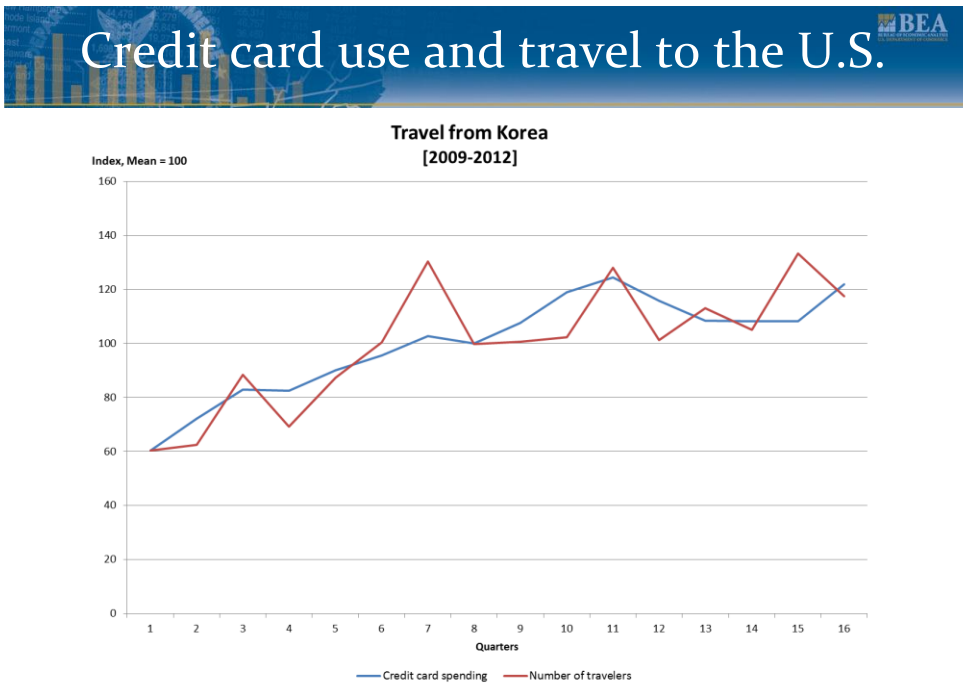
21. The following charts illustrate the challenges in using big data. There are often huge numbers of observations, but they may not be representative observations. Chart 3 compares data on credit card use by U.S. citizens in France with customs data on the number of U.S. citizens travelling to France. Neither is the appropriate measure that one would want to measure spending by all U.S. citizens travelling to France. The credit card measure covers credit card spending by those using credit cards in France and the number of travelers covers all U.S. citizens traveling in France. Although, the two series produce a similar pattern, the credit card pattern produces a more striking seasonal pattern that may be hard to explain. Does spending per traveler have a seasonal pattern, or is the spending pattern by credit card different than the pattern of cash spending?

Chart 3



22. Chart 4 shows credit card data and traveler data for travel from South Korea to the United States and the patterns are quite different. The data may be useful as a long-term trend indicator, but without weighting of credit and cash uses, and development of indicators for cash use, it will be hard to use the credit data for measuring monthly and quarterly patterns.

Chart 4



23. Chart 5 and 6 show data from a popular household budget tracking "app" that covers credit card and all other spending. The data from the tracking service fits quite closely with the representative data from the official U.S. Census Bureau data on retail trade for clothing. However, the same comparison for electronic goods shows that the budget tracking data significantly understate seasonal peaks in spending. Perhaps households that sign up for a budget tracking service are less prone to holiday "binge" shopping for "big-ticket items."

24. Chart 7 shows a measure of small business activity based on a U.S. small business accounting software, labeled alternative net profit indicator in the Chart. The other two indicators labelled NFBI (Non-financial business indicator) are measures used in the U.S. national accounts based on official employment and tax data, as well as other indicators for key small business sectors. (One of the two NFBI excludes capital and inventory gains and losses.) Small business income is extremely hard to track, even with official statistics, and any new data can be a big help, but in this case the trends from the accounting software data for the most recent years of the economic recovery, are so sharply at odds with the official data that it is hard to figure out how to use them. Although, they measure different things it is difficult to square the improving official receipts and employment numbers for small business with the falling net profits from small business coming from the accounting software data.

Chart 5

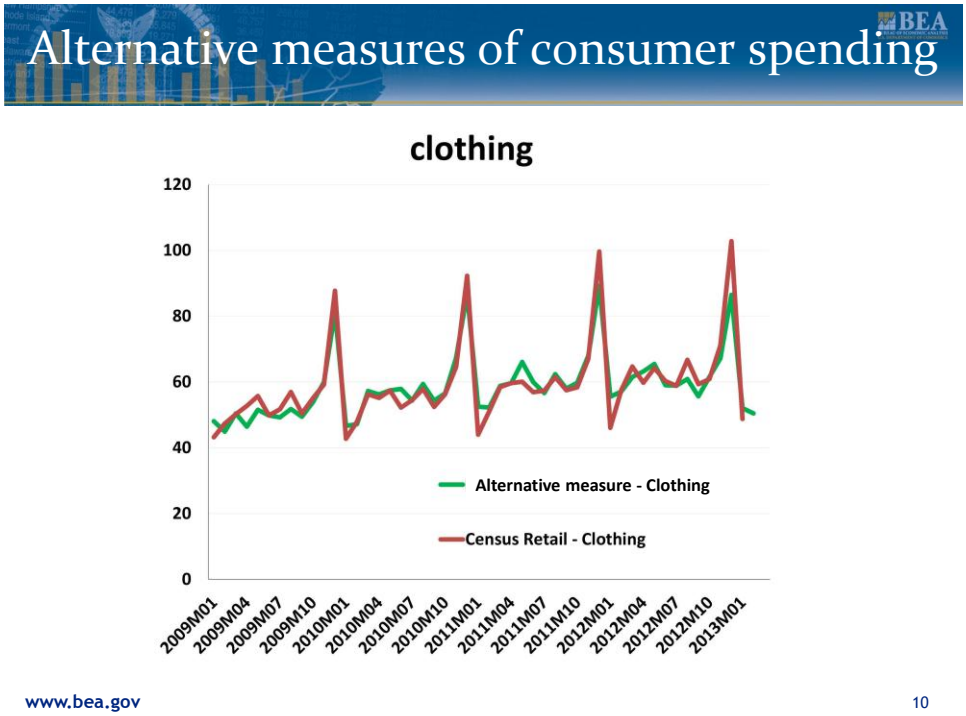


Chart 6

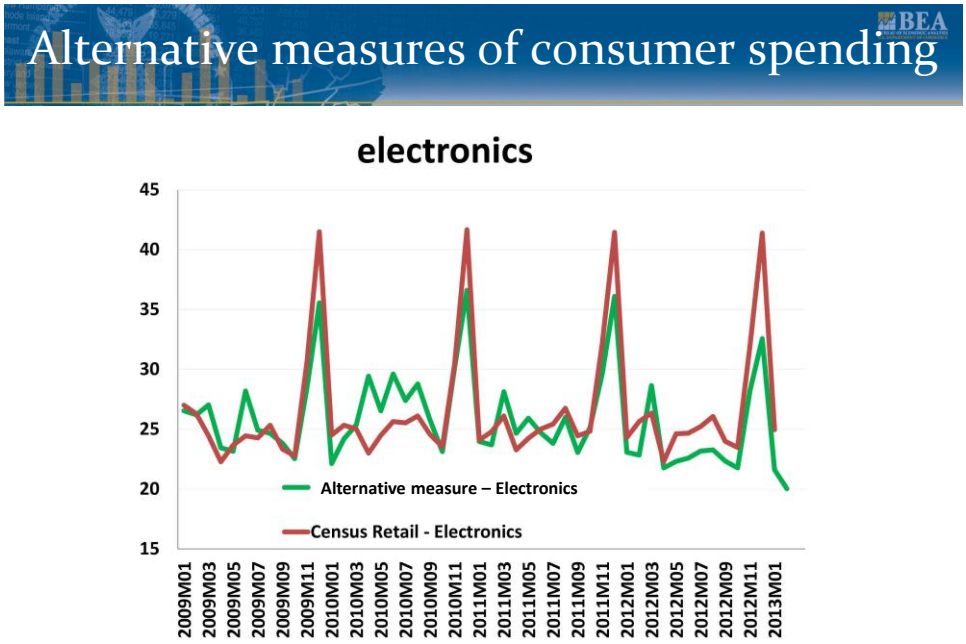
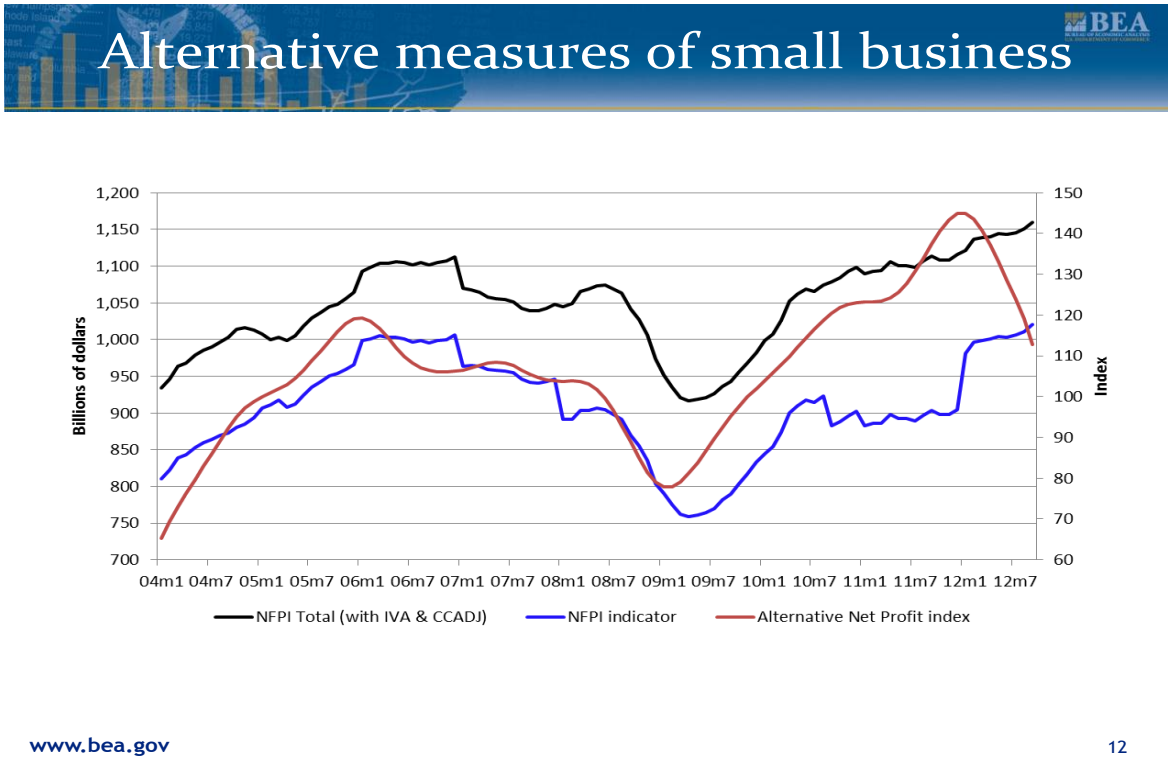
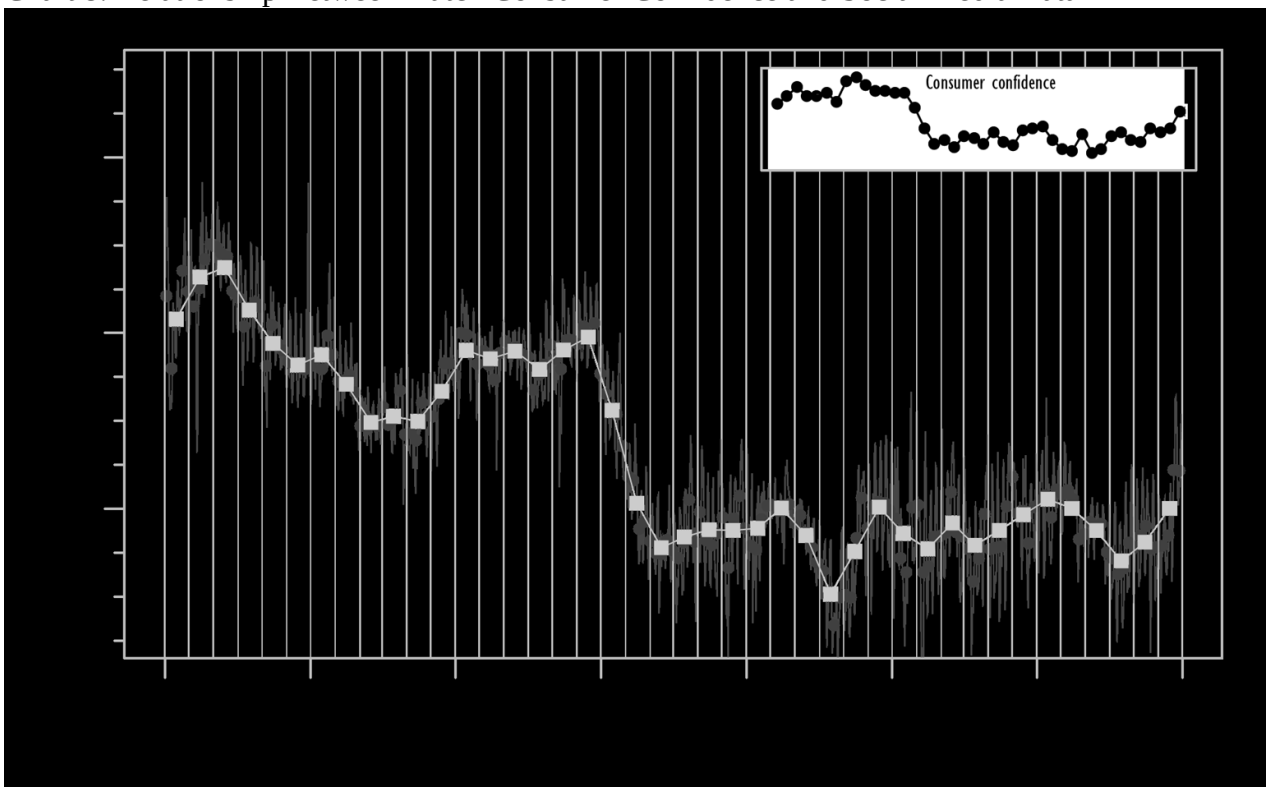


Chart 7



25. Chart 8 shows the relationship between Dutch social media sentiments on the state of the economy (Facebook supplemented by twitter messages) and monthly consumer confidence. The high correlation over the three-and- a-half years studied (2010-2013) provide a good example of how such data could supplement, and or be integrated with existing indexes.

Chart 8: Relationship Between Dutch Consumer Confidence and Social Media Data



Source: Piet J.H. Daas and Marco J.H. Puts, "Social Media and Consumer Confidence, " European Central Bank, Statistics Paper Series, No. 5 September 2014

26. Other examples include an Indonesian study of Facebook and twitter data on expressions regarding food prices and food price inflation and the potential for the use of social media for official statistics including the need for validation, methods for filtering out noise in the data gathered from social media, and developing robust estimates for their use with official statistics.[2]

27. Another study by Eurostat to assess the feasibility of the use of mobile positioning data for measuring tourism cited their large potential but pointed to the need to resolve a number of issues including: privacy and regulatory issues related to privacy; public opinion relating to the use of mobile data; financial and business related barriers to access; technical issues regarding use of and access to mobile data; and methodological issues relating to effectively using mobile data.[3]

[2] Pulse Lab Jakarta, "Mining Indonesian Tweets to Understand Food Price Crises," UN Global Pulse, Methods Paper, February 2014.

[3] EUROSTAT, "Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics, Consolidated Report Eurostat Contract No 30501.2012.001-2012.452, 30 June, 2014.

28. The final example is the use of postal big data. Despite declining use of the postal service around the world, Anson and Heble show how postal data can be used to develop an International Letter-Post Weight Index that seems to work well as a leading indicator catches changes in direction before a commercial leading index over the two year period 2010-2012 as well as differences in performance across countries in Europe. They also suggest that such real-time data could improve forecast accuracy. As with the other studies, use in leading indicators or forecasts would require testing over a longer time frame, developing methods that would explore not only direction of change, but relative magnitude of changes, and testing across countries.

F. Incentives: Exploiting Public and Private Benefits of New Extrapolators

29. The potential benefits to official statistical agencies and owners of public and private administrative data are large. For official statistical agencies, if the challenges of concepts, definitions, and representativeness can be resolved, the use of big data has the promise of more timely and detailed data at a significantly lower cost than new or expanded survey collections.

30. For example, in the United Sates, mounting a new survey collection may cost well over $20 million, whereas a micro-data matching exercise using existing data that produces more timely and detailed BEA may cost less than 1/5 of the cost of a new survey. Indeed, a project that matched existing firm level data with existing plant level data expanded the industry detail on U.S. foreign direct investment data from roughly 100 industries to over 600 industries and from national level data to data for all 50 U.S. states at a cost of $3 million.

31. EUROSTAT has matched existing business register (public and private) and trade data at the micro and aggregate level to create detailed data on the characteristics of importing and exporting firms that are extremely helpful to designing trade and investment policies.

32. The owners of public and private data want to exploit the data they collected for business and administrative purposes for other purposes. For businesses, these other purposes include marketing, geographic location plans, short and long-term investment plans, and the benchmarking of company to industry performance. For the public sector, these other purposes in include geographic planning for infrastructure and the provision of services; for understanding behavioral responses and characteristics of the population for designing health and unemployment insurance, for tax and other policies, for tracking the incidence of disease for public health purposes; and for evaluating the effectiveness of government programs.

33. For big data to be useful for all these secondary purposes the administrative and business data must meet the measurement purpose for which it is to be used. For example, is the number of Facebook uses of the word flu a reliable indicator for the incidence of flu? Or are the health records of the members of a large health insurance plan in an urban area representative of the general population, insured and uninsured? Or in the area of economic statistics, are on-line price

quotes representative enough of all prices to predict the official CPI with enough accuracy to be useful to traders in financial markets?

34. Official statisticians and big data users and producers, therefore, all benefit from harmonization of the various sets of statistics for secondary uses and have strong incentives to collaborate in doing so. As shown from the experience of national accountants, after examining the characteristics -- definition, population, methods, and performance -- of the source data they can often be benchmarked, bias-adjusted, and weighted (for use with other data) to successfully project official statistics.

35. There are also benefits from collaboration between official statistical agencies and information firms such as Google, Baidu, and Yahoo. Search engines may have problems identifying the most timely and authoritative data and may send users to dated secondary sources. By collaborating with official statistical agencies they can provide their customers with more accurate, timely, and relevant data searches. By linking to the original source data that customers are seeking they help their customers by raising their "statistical literacy" in understanding the implications of matches of what appear to be similar data from searches and "mash-ups." They also help official statistical agencies by providing improved access to, and understanding of, official statistics through broader internet dissemination and "branding."

36. Some of the most successful examples of the use "big data" have come from joint projects between the owners of private data and official statistical agencies. Such joint projects provide significant mutual benefits not realized by unilateral use of private source data. Examples include collaboration between the U.S. Bureau of Economic Analysis and IBM in the development of hedonic computer prices, with Chrysler and other motor vehicle manufacturers in developing new auto and truck pricing data, and with Google in the development of regional search engines.

G. Successful public-private collaboration requires

- A recognition of the mutual benefits of such a collaboration
- Transparency between the official statistical agency and owner of the business or administrative data regarding data collection and estimation methods used to produce the data; and
- Clear and strong rules for protecting the confidentiality of the data and of the proprietary methods used to produce the business or administrative data

37. Although seemingly obvious, the infrequency of public-private collaboration in some countries may inhibit such collaboration and require that statistical agencies initiate and outline the benefits of such collaboration.

38. Further, although transparency and understanding the data is critical to effective benchmarking of the private data to the fully representative official statistics, providing transparency on source data and methods may be difficult for proprietary data suppliers. Protection of the source data and methods are essential to the firm's ability to sell the data product to customers. Without protection of their intellectual property -- which is usually done through trade secrets rather than patents or copyrights -- competitors may be able to replicate the private data product without the development costs borne by the original owner, and undercut their sales.

39. Finally, protection of the private data, much of which is confidential customer data, is essential. Unfortunately, long-standing lack of trust in government in the area of privacy is a problem to such collaboration that has been exacerbated by recent highly public examples of governments' accessing private "big data" for National Security reasons (with or without the private owner's consent). While none of the recent transgressions, or indeed most past data breaches, have involved official statistical agencies, businesses and the public are inclined to mistrust government in general. Further, cases where official statistical agencies have made inappropriate use of confidential data have been quite public and likely tarnished the general reputation of statistical agencies around the world.

40. In addition to the use of collaborative projects to access confidential public and private data, patent and copyright protection may require their use even in the use what appear to be publicly available private data. Even if legal protections do not bar the use of public data through "data scrapping," collaboration that helps in understanding the characteristics and proprietary methods used to producing the internet and other data may be critical to the successful use of that data by official statisticians.

H. Privacy Concerns about New Uses for Big Data

41. As noted above, advances in the use of "big data", including high-profile political and national security access to big data, have raised significant privacy concerns. For individuals the concerns relate to the disclosure of detailed personal medical, financial, legal and other sensitive data; uses that would lead to discriminatory outcomes; and uses for tax, investigatory, legal, and other governmental purposes.

42. For businesses, the concerns are disclosure and release to the public of: commercially valuable marketing and other data sets; proprietary information on the methods and sources used to produce those data; disclosure to competitors of important strategic information on pricing, costs, profits, and markets; and the use of such information for tax, regulatory, investigatory, legal, and other purposes.

43. These concerns are not new concerns, nor is the use of data collected for business and governments for non-statistical purposes in the production of official statistics. In addition, many of the same confidentiality and privacy concerns have arisen in the course of centuries of government surveys of households and businesses. The mechanisms for addressing those concerns may be able to be carried over to rules, or protocols, for protecting data in today's big data world.

44. In developing such rules, protecting privacy and confidentiality are key. For business data - the following types of information must be protected:

- The data itself as an information product (micro and macro). This intellectual property of firm has economic value and can be sold. Government must make sure there is no disclosure that would give it away free.
- Data on details of businesses, such as prices, costs, and market share, that would be useful to competitors.
- Personal information on customers. Loss of such data through security breaches or hacking undermines the reputation of the firm, and discourages use of electronic transactions that can result in the loss of business.
- Proprietary information on the methods and sources used to produce the data

45. Such protocols to protect privacy are essential, and have been used for years-- to promote trust and address concerns that government may use of micro-data for regulatory, tax, and other policies. An erosion of public trust can reduce response rates on official surveys, reduce honesty in reporting, and reduce the overall accuracy of the official statistics collected from the business community.

46. The public also need to have their concerns addressed and their data protected. They need to be sure that there will be no disclosure of:

- Name and address or other identifying information that could be used for marketing and other business purposes.
- Intimate personal details, including such information as marital and health status, and income.
- Any information whose use could lead to discriminatory outcomes in such areas as employment, eligibility for loans, or eligibility for government programs.
- Any disclosure to non-statistical agencies that alters the balance of power of power between individuals and government, including the use of data collected for statistical purposes for tax, regulatory, investigatory and other non-statistical purposes.

I. Data Protocols for Public-Public and Public-Private Collaboration

47. Uses of Public Administrative and Business Data include both publicly available and confidential data. Data protocols focus on confidential data, although, as noted above they may be useful in fostering collaboration, understanding, and effective use of publicly-available administrative and business data.

48. A protocol for data access should begin by describing the benefits of such a collaboration. For example, a recent U.S. Executive order providing guidance in the use of protocols that would promote the use of Administrative data to leverage and improve statistical data, notes that, "...high-quality and reliable statistics provide the foundation for the research, evaluation, and analysis that help the Federal Government understand how public needs are changing, how well Federal policy and programs are addressing those needs, and where greater progress can be achieved."[4]

49. The U.S. memo then goes on to emphasize the importance of developing strong data stewardship policies and practices for the use of administrative data collected for non-statistical purposes (protection of privacy and confidentiality); documenting the key statistical attributes and quality controls for use of data; and developing memorandums of understanding (contracts or protocols) that cite relevant laws, regulations, policies, practices, responsible parties, and penalties associated with use and misuses of the data.

50. A protocol for access to private data should first cite the purpose of the agreement. It should cite the specific benefits (see incentives cited above) accruing to the public and private partners and the specific data products that will be produced by the agreement.

51. Second, the protocol should address the uses of the data and the quality of the data. Such an understanding helps prevent difficulties later in the project. It also provides a better understanding of the baseline data's strengths and weaknesses that will need to be addressed by benchmarking, weighting, and for seasonal, bias, and other adjustments.

52. Third, the protocol should cover the roles and responsibilities for the protection of the data. Elements that should be covered include confidentiality and privacy, data security; data transfer, media, and methods for transmission of data. It should also set out the specific penalties for unauthorized disclosure of information, including any applicable privacy laws and their penalties, up to, and including imprisonment.

[4] MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES, FROM Sylvia M. Burwell, Director of the U.S. Office of Management and Budget, "Guidance for Providing and Using Administrative Data for Statistical Purposes," February 14, 2014.

53. Fourth, the Protocol should cover such key details as the parties to the agreement; duration of the agreement; legal and programmatic authorities; estimated costs and payment (or cost sharing agreement, including in-kind services); administrative and programmatic contacts; procedures for the resolution of conflicts; procedures for reviewing, updating, modifying, cancelling, and renewing the agreement.

54. A copy of the elements included in a model agreement developed by the U.S. Office of Management and Budget that is the basis for the above recommendations is included below.

J. Other Roles for Official statisticians in the Use of Big Data

55. Some have suggested that official statisticians might play a role in auditing and certifying the accuracy of "big data" for official and private uses similar to the role played by consumer testing and ratings services such as Consumer Reports. Such a proposal was put forth in the United States and encountered stiff resistance from data users and the business community.

56. There were several reasons for this resistance. First, information products are an important source of revenue for firms and they do not wish to risk disclosure of the proprietary methods and source data they use to produce the data that they sell (see customers' privacy discussion above). Second, at the time and more so now, firms and their customers simply don't fully trust governments and are concerned that the data will be used for non-statistical uses (tax, regulatory, or investigatory); Third, there is a general resistance by industry to expanded government oversight and a strong preference for voluntary oversight. Fourth, such a role for official statistical agencies may well produce an adversarial environment with the very businesses that official statisticians must rely on for their regular survey data. Such an environment could well end up weakening, rather than strengthening official statistics by lowering response rates, and reducing the accuracy of responses.

K. Don't under-estimate the value of existing statistical and administrative data!

57. Despite the high level of excitement surrounding the use of on-line search data, internet prices, and other business data, one of the most promising areas for the use of big data is the use of existing Statistical and Administrative data. Data matching, at the micro or even sub-aggregate level, across statistical agencies and with non-statistical agencies can produce large gains at a relatively low price. Example include the use of medical and health insurance records for the construction of medical care price indexes, the matching of business registers, establishment, and enterprise data to produce for more detailed data on the characteristics of firms engaging in trade and foreign direct investment, and the use of motor vehicle registrations data to estimate

depreciation schedules. In general, expanded access to tax and administrative data for statistical purposes has the potential for quickly producing large benefits in the accuracy, level of detail, and efficiency of official statistics.

L. Conclusion

58. This paper has provided an overview of big data and their use in producing official statistics. Although advances in information technology, data sources, and methods have driven interest in the use of business and administrative government data collected and used for non-statistical purposes, use of such data is not new. Nor is it likely to be a panacea for statistical agencies confronting demands for more, better, and faster data with fewer resources. However, with careful attention to incentives; protection of privacy through data protocols and collaborative agreements; and integration of these non-statistical with existing statistical data, big data can play an important role in improving the accuracy, timeliness, and relevance of economic statistics at a lower cost than expanding existing data collections.

Appendix B: Standard Elements of A Model Agreement for the Provision of Administrative Records for Statistical Purposes (US OMB M-14-06, February 14, 2014)

1. Parties to the Agreement
2. Legal and Programmatic Authority
3. Duration or Period of Agreement
4. Purpose
5. Use of Data
6. Data Quality
7. Roles and Responsibilities for Data Protection
    a. Confidentiality and Privacy
    b. Data Security
    c. Data Transfer, Media and Methods for Transmission of Data
    d. Record Keeping, Retention, and Disposition of Records
8. Specific Penalties for Unauthorized Disclosure of Information
9. Potential Work Constraints
10. Breach
11. Disclaimers
12. Reporting
13. Administrative Points of Contact
14. Funding Information
15. Estimated Costs and Payment
16. Resolution of Conflicts
17. Modification/Amendment of Agreement
18. Cancellation of Agreement
19. Periodic Review of Agreement
20. Concurrence and Agency Signatory