Big Data and Official Statistics: Opportunities and Obstacles

San Cannon

Kansas City Fed

(standard disclaimer applies)

Outline

- Introduction
- Data lifecycle and the effect of “big data”
  - Acquisition
  - Preparation
  - Analysis
  - Storage
  - Publication
- Examples from the Federal Reserve Bank of Kansas City
- Lessons learned

Introduction

- FRBKC: Covers the 10th District in the US (Colorado, Kansas, Nebraska, Oklahoma, Wyoming, parts of Missouri and New Mexico). Very technology oriented.
- CADRE: Formed from a research technology group, provides services to support analytics across the Federal Reserve System. The "big data" folks at the Fed.
- ME: Economist/data scientist/secret librarian who thinks about how to make data as useful as possible to the largest number of users.

The new data world

- Past: surveys are carefully crafted using proven statistical methods to collect data that become official government statistics.
- Present: many private companies sell millions of transactional or administrative records or create alternative measures.
- Sometimes they look the same and measure the same thing; sometimes they don't.
- But more data can't be bad, can it?

Data lifecycle

Get data: data acquisition

- We used to have to look for data and sometimes design and implement a data collection. Now we acquire data in many different ways.
- Comparing carefully designed data collection with data captured without a design concept (administrative records, sensor data, internet "exhaust", text, etc.) can be difficult.

Get data: data compilation

- What about combining data sets?
- Traditional (small) data often couldn't be efficiently combined because of differences in study design.
- More often data are now being combined with other data with potential for unintended consequences.
- The "mosaic effect": big picture from smaller pieces may not have been seen previously.

The KC experience

- We host both acquired and collected data. For example, we have government data on mortgages as well as loan level mortgage data on performance that we purchase from private vendor.
- Together they are an invaluable resource for researchers and bank supervisors but…
- When you combine them, it is possible to determine details about a particular mortgage servicer which violates the terms of the contract with vendor.

Data lifecycle

Prepare data: data management

- Traditional (small) data are relatively easy: limited number of variables or observations, uses traditional data validation or quality measures, variety of storage options.
- New (big) data are different: many variables or observations, data scrubbing often with limited quality checks employed, storage options more limited.
- Unstructured data (of any size) has additional challenges – how does it fit in your schema?

The KC experience

- Curated data from official sources generally have better governance and more concern about quality.
- We manage data from large financial institutions and have strict validation rules. Errors require the banks to resubmit the entire collection.
- We also manage data from private sources and if we find errors, they may or may not be corrected.

Data lifecycle

Data analysis

- Data now have many applications but do different uses require different data?
- Traditional policy work relied on hypothesis testing to understand underlying causes.
- Predictive analytics and data mining focus on the what will happen without concern for why it happens.
- Correlation vs. causation – where is one more relevant than the other? When is it okay to only answer the "what" without the "why"?

Process also changes

- Traditional research approach means formulating the question THEN gathering the data. Right?
- There are questions that get asked based on theory or preconceived hypothesis.
- Now there are questions that get asked because they can be answered.
- And there is a belief that statistical training is less relevant in a big data world. (!)

The KC Experience

- We provide a full range of statistical and econometric tools to support traditional policy work.
- We have begun to add more tools to support data science and machine learning.
- And we are teaching statistics and data science to all parts of the Bank to increase our analytical skills in all areas.

Data lifecycle

Store data to use

- How do we make it easier for people to use more data than they have ever had before?
- Does everyone need to create a local sample?
- Does it even make sense to pull the data out of the database?
- How can you manage the performance on your networks when users play with “big data”?

Store data to preserve

- What about preservation? If we use a LOT of data to create on official output, what do we need to keep?
- The tension between archiving for records purposes and preserving for research increases with data size.
- Curation becomes more challenging with more than one of the Vs of "big data" (volume, velocity, variety).

The KC experience

- We create specific views and samples that work well with traditional analytic tools.
- We have found that our Massively Parallel Processing (MPP) storage has made a huge difference in performance.
- We co-locate our data and analytics tools to minimize network effects.
- We are encouraging users to use in-database analytics instead of smaller extracts.

Governing data storage

- Or can you store it? Or use it in a particular manner?
- Many data providers have restrictions on how data may be preserved or stored.
- “Free” data often comes with restrictions that no one notices.
- It’s not clear how valid these “browse through” contracts are legally but….

Store data: fine print

"You agree not to **copy, reproduce, modify, display, perform, publish, distribute, transmit, broadcast, circulate, create derivative works from, store**, or link to the web site or any Content without the express prior written consent of S&P Dow Jones Indices (which may be in the form of an email). "

Retrieved from http://us.spindices.com/terms-of-use/ on 29 August 2017. Emphasis is the author's.

The KC experience

- We have a data librarian and data curator position to provide full time support for data services.
- We are pioneering archiving standards and practices for large data sources.
- We specifically address licensing and governance issues as part of data contracts or agreements.

Data lifecycle

Publish data: sharing

- The commoditization of data means more data are available but they cost more money. The private data market place can cater to users specific needs... for a price.
- The "open data" movement is changing expectations. Government agencies are expected to provide fancy applications and interfaces to data... for free.

The KC Experience

- We are new entrants to the data dissemination world - we don't have official statistics.
- We have created some indicators that we publish using our central developers and design team.
- For larger data, we are piloting some data access platforms for the public to help improve the way that some larger datasets are shared.

What have we learned?

- Support in the "big data" world is as important as technology.
- Data librarians can be incredibly helpful in finding, evaluating, organizing and describing data. They understand the uses and can help with acquisition negotiations.
- New tools are helpful for experienced data users but new users require more help.

What else have we learned?

- Data storage may be inexpensive but it's only part of the story. Data need to be managed and data quality assessed regardless of size or source.
- Sharing data is an important public good with monetary costs and a non-monetary return. The budget pressures should be addressed from the beginning.
- Expectations need to be managed: "big data" doesn't solve everything.

Moral of the story

- Lots of data is the new normal.
- It is "big data" if it's bigger than you are used to.
- It is "too much" if you don't know what to do with it.
- With consideration, careful planning, and creativity, the opportunities for big data can overcome the obstacles.

Last slide

Thanks for listening!

sandra.cannon@kc.frb.org

+1 816 881 2596