Big Data and Modernizing Federal Statistics: Update

Ron Prevost Ph.D.
Senior Statistician for Data Integration,
Research and Methodology Directorate

June 12, 2015

Census Bureau "Big Data" Research Agenda

- Methodological
- Computational
- Policy / Legal
- User and Stakeholder Engagement

Goals

- Improve the Utility of Economic and Social Statistics
  - Improve timeliness
  - Improve domain detail (e.g., geographic, industry)
  - Fill data gaps
  - Improve reliability
- Reduce Respondent Burden
- Optimize mix of survey and non-survey source data

Update on Efforts

- New Center in R&M (need name)
  - Hub for Bureau efforts in this area
  - Lead projects
  - Affiliated projects managed in other directorates
  - Developing staffing strategy

Big Data Center Projects

- Innovation Measurement Initiative (IMI)
- MIT Workshops
  - Big Data and Commodity Flows (joint with the Bureau of Transportation Statistics)
  - Big Data and Privacy
  - Big Data and Adaptive Survey Design
- Big Data Class
- Sandbox

Innovation Measurement Initiative (IMI)

- Collaborative research project between Census, University of Michigan, Ohio State, University of Chicago, and New York University
- Integrate university data on federally funded research grants with Census Bureau data assets
- Produce statistics consistent with the Bureau's economic and social measurement mission and directly relevant to the data provider.

IMI Background

- Census Goals:
  - Improve measurement of small but important sector of the economy
  - Address data gaps in the measurement of innovation and relation to economic growth
  - Learn how to collaborate with data providers to deliver data products they value
  - Prototype project that can be scaled and extended to other sectors of the economy

IMI Background

- Innovative Aspects:
  - Collaboration with the University of Michigan's Institute on Research in Innovation and Science (IRIS)
  - Experiment with utilizing "fat pipe" of data for a sector of the economy
    - The University data is complementary to business and household data at Census
  - Makes extensive use of skills our staff learned through the Big Data classes

Establishment of new Institute

- Institute for Research on Innovation and Science (IRIS) founded 01/01/2015
  - Goal – leverage existing data to both serve university data and generate new research
  - Core facility at University of Michigan
  - 3 years seed funding for infrastructure from Sloan & Kauffman
- More efficient mode of data ingest for the Census Bureau
  - One MOU rather than N

(subset of) Preliminary Findings

- So far we've constructed basic indicators on:
  - Worker characteristics
  - Job placements (of students)
  - Vendor characteristics (including geographic patterns)
  - Startups
  - Patents
  - Trade

Job Placements - 1 Year After Leaving Institution

	Individuals on Grants		Proportion by Sector (6+Months)			Proportion by Sector (6+ Months & <50miles)		
Last Year	Overall	6 Months	Industry	Academia	Government	Industry	Academia	Government
2010	11,689	8,041	55.9%	36.0%	7.4%	22.9%	54.1%	19.6%
2011	19,049	13,562	63.1%	29.9%	6.3%	15.6%	49.5%	9.8%
2012	19,722	12,185	58.8%	34.4%	6.1%	20.9%	57.4%	20.0%

- The initial links suggest the main destination of grant recipients is Industry, followed by Academia
- Geographic matches very interesting, but can't be shown for disclosure reasons

Job Placements - 1 Year After Leaving Institution

By Funding Source

Funding Source	Individuals on Grants		Proportion by Sector (6+Months)			Proportion by Sector (6+ Months & <50miles)		
	Overall	6 Months	Industry	Academia	Government	Industry	Academia	Government
NIH	17,336	13,684	61.5%	31.7%	6.1%	16.5%	49.5%	13.9%
NSF	7,118	4,784	56.5%	36.9%	6.0%	19.2%	49.9%	11.5%
Non-Federal	16,082	9,382	63.5%	28.2%	7.7%	18.7%	58.4%	16.7%
Dept of Education	2,852	1,383	49.0%	46.1%	4.3%	32.8%	69.0%	27.1%
Other	7,072	4,555	54.1%	38.7%	6.5%	25.0%	55.1%	21.5%

We can also break out Funding Source and Job Placements Relationship by School and Last Profession

2010 Cohort 2-digit NAICS

NAICS	NAICS Description	LBD	All Universities
11	Forestry, Fishing, Hunting, and Agriculture Support	1.12%	0.77%
21	Mining	0.59%	0.36%
22	Utilities	0.72%	0.32%
23	Construction	4.64%	2.63%
31-33	Manufacturing	9.75%	12.24%
42	Wholesale Trade		
44-45	Retail Trade		
48-49	Transportation and Warehousing		
51	Information		
52	Finance and Insurance		
53	Real Estate and Rental and Leasing		
54	Professional, Scientific, and Technical Services		
55	Management of Companies and Enterprises		
56	Administrative and Support and Waste Management and Remediation Services		
62	Health Care and Social Assistance		
71	Arts, Entertainment, and Recreation		
72	Accommodation and Food Services		
81	Other Services (except Public Administration)		

2010 Cohort 3-digit NAICS (Manufacturing)

NAICS	NAICS Description	LBD	All Universities
330	Primary Metal Manufacturing	0.00%	0.01%
331	Primary Metal Manufacturing	0.33%	0.28%
332	Fabricated Metal Product Manufacturing	1.18%	1.01%
333	Machinery Manufacturing	0.85%	1.38%
334	Computer and Electronic Product Manufacturing	0.78%	1.73%
335	Electrical Equipment, Appliance, and Component Manufacturing		
336	Transportation Equipment Manufacturing		
337	Furniture and Related Product Manufacturing		
339	Miscellaneous Manufacturing		
541	Professional, Scientific, and Technical Services		
621	Ambulatory Health Care Services		
622	Hospitals		
623	Nursing and Residential Care Facilities		
624	Social Assistance		

2010 Cohort 4-digit NAICS (Computer & Electronics Manufacturing)

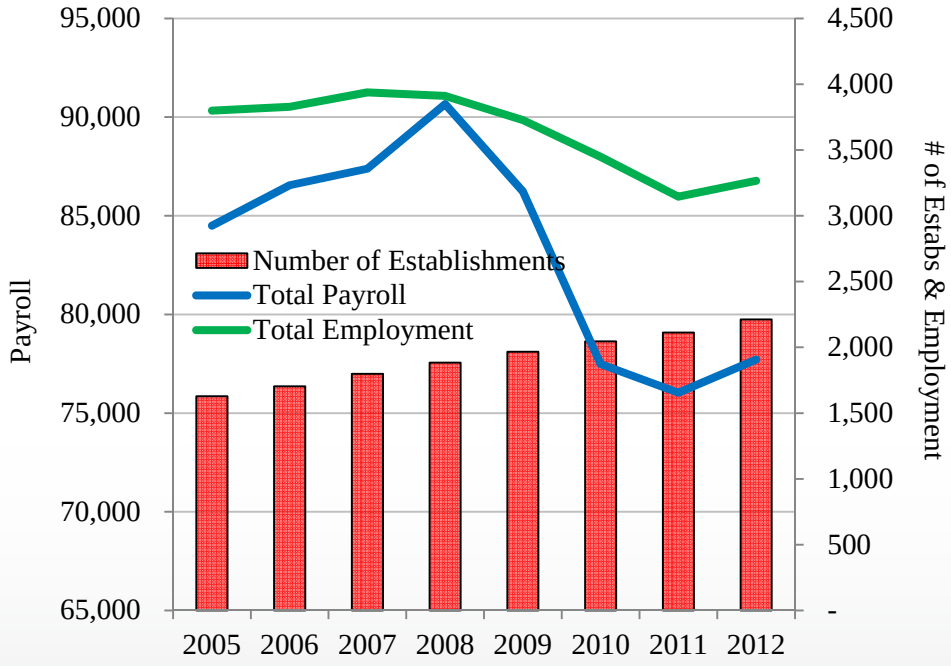
NAICS	NAICS Description	LBD	All Universities
3341	Computer and Peripheral Equipment Manufacturing	0.06%	0.26%
3342	Communications Equipment Manufacturing	0.10%	0.17%
3343	Audio and Video Equipment Manufacturing	0.01%	0.02%
3344	Semiconductor and Other Electronic Component Manufacturing	0.25%	0.54%
3345	Navigational, Measuring, Electromedical, and Control Instruments Manufacturing	0.34%	0.74%
3346	Manufacturing and Reproducing Magnetic and Optical Media	0.01%	0.00%
5411	Legal Services	1.02%	1.23%
5412	Accounting, Tax Preparation, Bookkeeping, and Payroll Services	1.15%	1.29%
5413	Architectural, Engineering, and Related Services	1.13%	1.92%
5414	Specialized Design Services	0.09%	0.04%
5415	Computer Systems Design and Related Services	1.30%	1.99%
5416	Management, Scientific, and Technical Consulting Services	0.86%	1.67%
5417	Scientific Research and Development Services	0.63%	0.00%

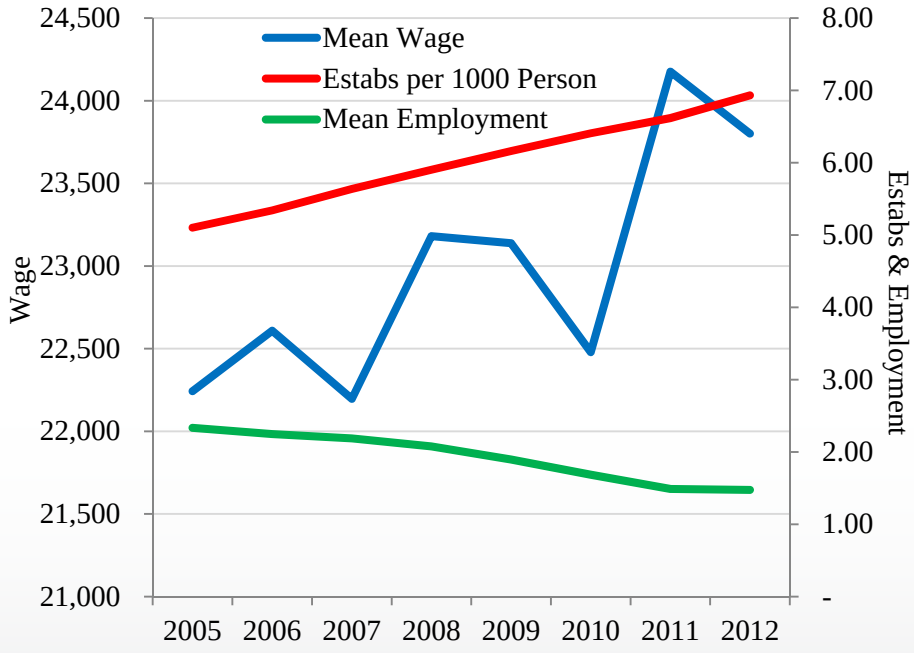
Over/Under-Represented Industries

Most Overrepresented 4-digit NAICS

	NAICS	NAICS Description	U.S.	Univs.	Dif
1	5413	Architectural, Engineering, and Related Services	1.13%	4.34%	3.21%
2	5415	Computer Systems Design and Related Services	1.30%	3.97%	2.68%
3	5613	Employment Services	3.87%	6.26%	2.39%
4	5416	Management, Scientific, and Technical Consulting Services	0.86%	2.67%	1.82%
5	6221	General Medical and Surgical Hospitals	4.63%	5.96%	1.33%
6	4236	Electrical and Electronic Goods Merchant Wholesalers	0.43%	1.72%	1.28%
7	6214	Outpatient Care Centers	0.69%	1.82%	1.12%
8	8132	Grantmaking and Giving Services	0.17%	1.25%	1.08%
9	5112	Software Publishers	0.32%	1.35%	1.03%
10	5191	Other Information Services	0.23%	1.25%	1.02%
Most Underrepresented 4-digit NAICS					
1	7222	Limited-Service Eating Places	3.63%	1.84%	-1.79%
2	4451	Grocery Stores	2.26%	0.69%	-1.58%
3	2382	Building Equipment Contractors	1.39%	0.27%	-1.12%
4	5221	Depository Credit Intermediation	1.80%	0.71%	-1.09%
5	4529	Other General Merchandise Stores	1.51%	0.44%	-1.07%
6	7211	Traveler Accommodation	1.66%	0.66%	-1.00%
7	7221	Full-Service Restaurants	4.03%	3.12%	-0.91%
8	8131	Religious Organizations	1.47%	0.56%	-0.90%
9	5617	Services to Buildings and Dwellings	1.46%	0.56%	-0.89%
10	6231	Nursing Care Facilities	1.46%	0.64%	-0.82%

Startup Business Dynamics (Matched through SS-4)





- Number of startups has been steadily increasing, although the cumulative size of these firms has been somewhat flat

Affiliated Projects

- 2020
  - Reengineering Address Canvassing
  - Planning and executing Non-Response Follow-UP
  - Cost Reduction to Field Reengineering
- Retail Statistics

Retail Big Data Project Goal

To explore the use of “Big Data” to **supplement** existing monthly/annual retail surveys to fill in data gaps and increase relevance. Primary focus is to try to generate geographic level estimates more frequently than once every five years through the Economic Census.

Retail Big Data Team Goal

To evaluate the data obtained from the NPD Group
to determine its usefulness in meeting the goal
of supplementing our retail statistics
with more frequent geographic level estimates.

The “Big” Data

- Completed RFI and RFP processes
- Awarded contract to NPD for two off-the-shelf datasets on 9/19/14
  - Automotive parts
  - Jewelry & watches
- Final datasets (2012-2014) received on 2/6/15

About NPD

- NPD has agreements with approximately 900 retailers worldwide covering approximately 150,000 locations/stores and $400 billion in annual sales.
- Smaller businesses are generally not included
- Retailers provide aggregated (SKU-level) transaction data to NPD generally using a weekly feed (Sunday through Saturday) following the National Retail Federation reporting calendar
  - Store identifier/location
  - Item/Product code (e.g., SKU)
  - Dollar volume of sales
  - Units sold
  - Average price (calculated)
  - Flag distinguishing on-line from in-store sales
- Estimates of non-food and drug categories

Evaluation Plan

- Analyze incoming NPD estimates to identify potential errors prior to use
  - Errors in geographic coding
  - Missing data

- Evaluate comparability of NPD estimates against Census Bureau estimates
  - Aggregate levels
  - Month-to-month trends
  - Year-to-year trends
  - Coverage and representativeness

- Census Bureau Estimates
  - Monthly Retail Trade Survey
  - Annual Retail Trade Survey
  - 2012 Economic Census

Other Possibilities

- Exploring Feasibility of Obtaining Company Feeds
  - Agreements with individual companies
  - Access through 3rd party such as NPD, Nielsen, etc.
  - Store level data from credit card transactions (Mastercard, 3rd party processors, banks, etc)
- Benefits
  - Reduces reporting burden on companies
  - Obtain data quicker
  - Leveraging a 3rd party could help with standardized formats
- Test with select number of companies in 2017 Economic Census