Federal Statistical Agency Uses of Private Sector Data: Study Findings to Date

Andrew Reamer
George Washington Institute of Public Policy
George Washington University

Presented to:
Panel on New Vision for Federal Data Infrastructure
Committee on National Statistics
National Academies

December 9, 2021

Project Origins

- 2018: Chief Statistician Nancy Potok asked me to obtain grant to study federal uses of private sector data
- 2019: Sloan Foundation provided $50,000 grant
- Project start postponed due to Nancy's retirement and my work on 2020 Census
- Late 2020: ICSP indicated interest in subject
- March 2021: Mary Bohman facilitated my offer to employ my grant-funded time in support of ICSP-sponsored study. ICSP accepts.
- Project committee: Erich Strassner and Annabel Jouard (BEA), Bill Wiatrowski (BLS), John Staub (EIA)

Project Goal

- Ascertain the:
  - breadth and nature of statistical agency uses of private sector datasets
  - benefits and challenges in the use of such datasets
  - implications of the findings for federal uses of such datasets going forward

Project Tasks

- Survey of ICSP agencies regarding uses of private sector data, benefits and challenges experienced, and adaptations required (Completed July 2021, N = 20)
- Selection of 12 case studies (eight agencies) for further information collection
- Case study survey completed (September)
- Case study analysis and interviews to be done
- Draft report for ICSP review
- Final report to ICSP + paper for publication

Private Sector Data: Definition

- Data purchased or licensed from private-sector entities within the past five years.
- Excludes data:
  - Traditionally collected directly from private entities via survey
  - Extracted or collected solely from public administrative records
  - Scraped from the web and not purchased or licensed from a private entity
  - That is freely accessible from public entities or websites
  - Purchased or licensed more than five years ago

Participating Agencies – Survey 1

- BEA
- BJS
- BLS
- BTS
- Census
- EIA
- ERS
- NASS
- NCES
- NCHS
- NCSES
- SSA
- USAID
- VA Analytics
- 6 unidentified (more blank answers, empty "other" and comment boxes)

Survey 1 -- Preliminary Results

- Presenting quantitative results only
- Agencies provided extensive, detailed comments that we are working through
- Findings to date are preliminary and subject to change on further review

Extent of Private Data Use

- All respondents but one use private data (19 of 20, 95.0%)
- SSA is the exception

Reliance on Private Datasets in Production

Agency	# Private Datasets in Use	% Statistical Programs Using Private Data
BEA	142	100
EIA	Approx 80	100
USAID	50	NA
Census	Approx 20	20
NCES	12	30
BTS	10	50
NCHS	Approx 8	40-50
BLS	Approx 5	10
VA	Approx 4	NA
NSF	Several	NA
NASS	3	NA
BJS	2	6

Reasons for Private Data Use

Survey Responses (check all that apply)	#	%
To supplement or combine with existing agency held data	14	82.4
To better understand other indicators of the economic environment (situational awareness)	13	76.5
To continue current reporting capacity of agency priorities	9	52.9
Verification, quality control or quality assurance for existing data or estimates	9	52.9
For use as or to identify a survey frame	8	47.1
To make available to agency stakeholders and data users (publish data to the web)	8	47.1
Direct data collection is too costly (private data sets are a cost-efficient alternative)	7	41.2
Direct data collection is too burdensome on respondents (resulting in poor response or quality)	4	23.5
Surveys with code(s)	17	

Data Acquisition Challenges

Survey Responses (check all that apply)	#	%
Cost	12	70.6
Legal Hurdles	8	47.1
Other	8	47.1
Difficulty Obtaining Granular Data	6	35.3
Security challenges in accessing data once obtained	4	23.5
Technological challenges in accessing data once obtained	4	23.5
No challenges	3	17.6
Surveys with code(s)	17	

Data Analysis Challenges

Survey Responses (check all that apply)	#	%
Methodology is not clearly documented	14	77.8
Data require cleaning or additional manipulation to make useful	12	66.7
Data are of poor quality or incomplete	11	61.1
Data do not fully capture/match agency measurement objectives (construct validity)	11	61.1
Data are not fully representative of the measurement universe (agency frame)	10	55.6
Data are not easily matched to existing agency data	9	50.0
Other	5	27.8
No challenges	3	16.7
Surveys with code(s)	18	

Concerns about Sustainability of Private Dataset Access

Survey Responses	#	%
Concerned for data sustainability	10	52.6
NOT concerned about data sustainability	9	47.4
Surveys with code(s)	19	

Data Management Approaches

Survey Responses (check all that apply)	#	%
Varies by project	10	62.5
Decentralized (at sub-agency level, such as individual program)	8	50.0
Centrally (at the agency level)	8	50.0
Other	5	31.2
Limited to a specified dollar amount	3	18.8
Surveys with code(s)	16	

Private Data Sharing Practices Intra-Agency

Survey Responses (check all that apply)	#	%
As needed by other internal groups or persons	10	55.6
Only if there is a work-related need	10	55.6
Standardized process to share data within contractual/MOU limits	7	38.9
Dependent on any additional costs to share data internally	7	38.9
The project team managing the private data determines	7	38.9
Other	7	38.9
Data sharing is not permitted	1	5.6
Surveys with code(s)	18	

Adoption of Data Principles or Criteria

Survey Responses	#	%
Yes	14	73.7
No	5	26.3
Surveys with code(s)	19	

Specified Use of Private Data in Public Agency Plan, CBJ, Evaluation Documents

Survey Responses	#	%
Yes	11	57.9
No	8	42.1
Surveys with code(s)	19	

Ad Hoc/ Opportunistic Efforts to Obtain Private Data

Survey Responses	#	%
Yes	9	47.4
No	10	52.6
Surveys with code(s)	19	

Greater Reliance on Written Guidance Expected

Survey Responses	#	%
Yes	10	52.6
No	9	47.4
Surveys with code(s)	19	

Survey 2 – Case List

Respondents provided 30 cases. From these, we selected 12:

- BEA – Fiserv card transaction data for ML prediction
- BLS – J.D. Power motor vehicle data for CPI
- BLS/BEA/Census -- Opportunity Insights
- BTS -- Cell location data to track travel during COVID
- Census
  - Housing vacancy status during COVID (ACS)
  - Real Time 2020 Administrative Record Census
  - Simulation
- NASS -- Satellite data to assess impact of extreme weather events and anomalies
- NCES -- SAT/ACT data linked to NPSAS data
- NCHS -- National Hospital Care Survey data supplement
- VA Analytics Office
  - Acxiom data on vets not using VA services – for comprehensive vets database
  - Appriss data to identify incarcerated vets
  - Woods & Poole data

Survey 2 – Initial Observations

- Pandemic catalyzes expanded uses of private sector data
- Private sources can:
  - Provide data that are more granular and timely than survey-collected data
  - Expand coverage of individuals to those not in government administrative datasets
- Innovative uses of commercial location data and imagery
- Concerns expressed:
  - Data quality
  - Lack of technical documentation (impedes willingness to purchase, make best use of)
  - Data representativeness
  - Fewer years in time series than desired
  - Continued availability
  - Availability of staff skills to make optimal use of data
  - Capacity to store data (need for cloud storage)
- Costs vary widely – some are expensive

Next Steps

- Systematically analyze text insertions in survey responses
- Conduct interviews with select case studies
- Prepare draft report for ICSP review
- Provide final report to ICSP
- Prepare article for publication

Discussion

We welcome your questions