Opportunities and Considerations for the Use of Big Data Techniques in the Consumer Expenditure (CE) Survey

Brett McBride, Economist
Division of Consumer Expenditure Survey

2015 FedCASIC Workshop
March 4, 2015
Outline

1. Background on CE Survey and big data
2. Administrative data as type of big data
3. Respondents’ data source preferences
4. Techniques
   a. Record linkage
   b. Web scraping
   c. Text analysis
5. Summary
1. Background

- CE Survey: National household panel survey that collects information about spending habits of consumers
- CE being redesigned to limit measurement error, reflect new behaviors and technology
- Big data can benefit survey (improve accuracy, reduce respondent burden)
Big Data

Volume
Data size

Big Data

Sources of data
Variety

Velocity
Speed of change
Big Data

- Collection and Analysis Aspects:
  - Data storage (databases),
  - Tools used (APIs, MapReduce)
  - Capabilities (text analysis, record linkage, visualization)
2. Big data and Administrative data
Big data and Administrative data

- Commonalities
  - Origin – data incomplete and not tailored to survey needs
  - Volume – massive data sets of population members

- Differences
  - Variety – admin records usually structured, not arising organically from individuals
  - Database – relational database (SQL) vs. non-relational (NoSQL)
  - Selection – smaller problem of non-coverage
3. Data Source Preferences

- Research question asked of respondents finishing last CE Survey interview:

“If you knew your name and other information would never be singled out and would only be used for statistics, would you prefer that the BLS ask you about the cost of products you buy in a survey or use commercial records, like grocery store loyalty cards? [with ‘neither’ response option]”

- Similar questions for
  - healthcare services: survey, doctor or hospital, neither,
  - income information: survey, IRS, or neither
Similar Preferences for Survey Collection

- **Grocery Products**:
  - Survey: 51%
  - Alternative: 23%
  - Neither: 26%

- **Healthcare**:
  - Survey: 57%
  - Alternative: 19%
  - Neither: 24%

- **Income**:
  - Survey: 53%
  - Alternative: 18%
  - Neither: 28%
4. Techniques: Admin Record Linkage

- Privacy concerns related to linkage can be addressed
- Census can use administrative records without respondent consent if protected from further disclosure, used for statistical purposes (Gates, 2011)
  - Census has given notification when linking data (CPS)
- CE Survey respondents may be receptive to linkage
- Census’ Person Identification Validation System (PVS) links survey responses with records
Linkage Process: PIK Assignment

Reference file: Social Security’s Numerical Identification File

Survey input file: Name & DOB (SSN)
Linkage Process: PIK Assignment

Reference file: Social Security’s Numerical Identification File

Survey input file: Name & DOB (SSN)

Protected Identification Key (PIK)

Decennial Census
ACS
CPS
Admin records (SNAP)
Commercial files
Linkage Process: PIK Assignment

Reference file: Social Security’s Numerical Identification File

Survey input file: Name & DOB (SSN)

Protected Identification Key (PIK)

Decennial Census
ACS
CPS
Admin records (SNAP)
Commercial files

Personally Identifiable Information (PII)
Linkage Process: PIK Assignment

Auxiliary file data  Linked data  Survey file data
Linkage Process: Incomplete Data

- Not all sample units on input file matched
- Having more information on input file (e.g., SSNs) facilitates higher match rates
- Probabilistic method allows setting threshold for failure to match
- Contingencies needed for survey units unmatched to administrative data
Linkage Process: Uses for CE

- BLS would need to negotiate access to Federal tax information
  - Link to IRS income data - substitute to income questions
- SNAP state participation data - improve accuracy/minimize underreporting
- ACS - validate housing, vehicle reports
- Public housing records - rent payment information
Techniques: Web Scraping

- Pursue tool for scraping home value estimations
- Permit burden reduction & accuracy improvement
- Challenges: Sites (e.g., Zillow) have APIs, facilitating the sharing of website information, but prohibit storage of scraped data, other non-API methods of scraping
- Agreements may be needed to allow non-commercial scraping of websites with property data (private or government)
Techniques: Text Analysis

- Named entity extraction – recognition of entities (e.g., Person, Organization)
- CE Diary contains text expenditure descriptions which need converting into predefined codes
- Potential to use classifiers to assign words to certain codes with probability
5. Summary

- New opportunities for sourcing respondent-collected data
- Administrative data structured to permit matching to sample units, but like big data, may be incomplete
- Interact with agency supplying data to ensure data received is the data expected
- Other techniques – scraping, text analysis – improve data quality and process data more efficiently
Contact Information

Brett McBride
Economist
Division of Consumer Expenditure Survey
Office of Prices and Living Conditions
www.bls.gov/cex
202-691-5136
mcbride.brett@bls.gov