Considerations for using the Public Use Microdata (PUMD)

Aaron E. Cobet

Consumer Expenditure Surveys

Microdata Users Workshop

July 18, 2023

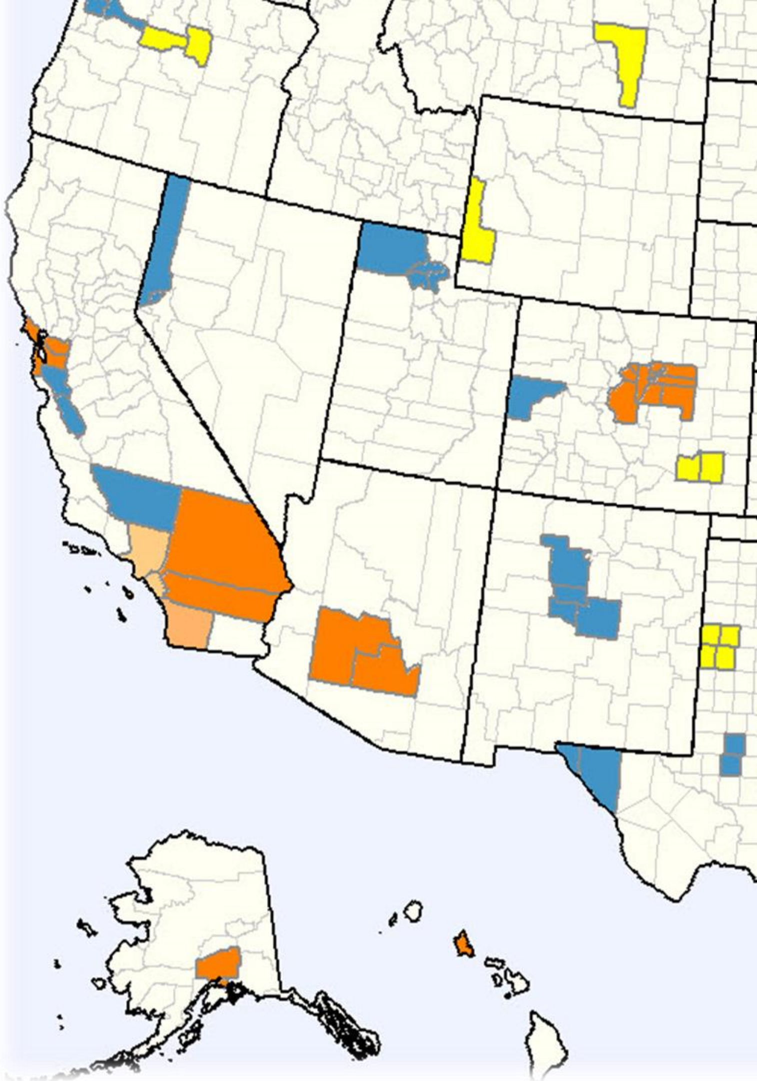
CE’s primary goal:

Provision of annual national and subnational data of U.S. consumer expenditures to the CPI and other users

Presentation outline

- Using CE’s non-expenditure data
- How CE methods affect PUMD
- Analyzing individual Consumer Units (CU)

Geographic data



- Limited geographic data
  - 4 Regions
  - 9 Census divisions
  - 4 selected states
  - 23 selected MSAs
  - Population size of area
- Consider:
  - No or few respondents in some areas
  - No zip codes or Census tracks data published

Available geographic areas

Geographic area	Number areas	PUMD	Tables	Databases
National	1	✓	✓	✓
Census Regions	4	✓	✓	✓
Census Divisions	9	✓	✓	
Selected States	4	✓	✓	
Selected MSA	23	✓	✓	✓
Population Size of Area	NA	✓	✓	✓

MSA by region

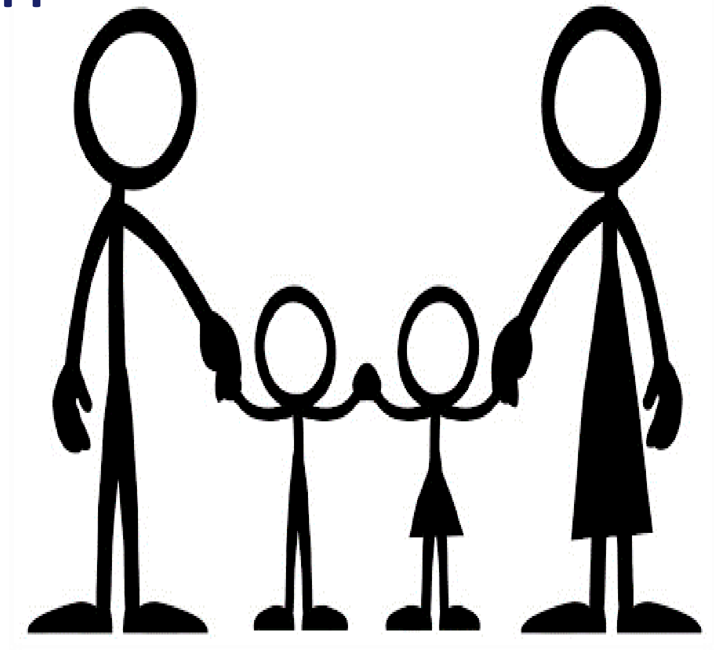
Region	Metropolitan statistical area
Midwest	Chicago, Detroit, Minneapolis St. Paul, St. Louis
Northeastern	New York, Philadelphia, Boston
Southern	Washington DC, Baltimore, Atlanta, Miami, Dallas Fort Worth, Houston, Tampa
Western	Los Angeles, San Francisco, San Diego, Seattle, Phoenix, Denver, Honolulu, Anchorage

Quality and quantity data

- CE mainly collects spending data
- CE selectively collects data on quantity and quality
- Consider:
  - Vehicles: Make and year
  - Houses: Number of houses bought and selected aspects

Purchaser and consumer data

- CE does not collect data on purchasers or consumers
- Users may infer these data sometimes
- Consider:
  - *Single member CU* best case
  - *Multi-member CU* depends on item, e.g. baby food
  - CUs may gift expenditures



Financial data

- CE collects data on income and assets
- Consider:
  - Use imputed income data (ITBI files)
  - Use financial data for categorizing CU and expenditure data and be careful using it for financial estimates
  - Income data may underrepresent affluent CUs
  - Income data may misrepresent economic well-being

See hidden slide for articles on the last two consideration

Financial data

- Financial data may underrepresent affluent CUs: A Nonresponse Bias Study of the Consumer Expenditure Survey for the Ten-Year Period 2010-2019
- CE financial data may misrepresent economic well-being: Interpreting CE income group data

Income tax data

- CE provides income tax liabilities
- Consider:
  - Break in series between 2012 and 2013 due to CE switch from collecting to estimating data
  - TAXSIM estimates tax liabilities that may not match actual taxes filed, especially for States

Income tax background material

- NBER TAXSIM description: http://users.nber.org/~taxsim/feenberg-coutts.pdf
- New estimates of personal taxes in the Consumer Expenditure Surveys: https://www.bls.gov/spotlight/2015/consumer-expenditures-tax-estimates/

Impact of survey design on PUMD

- Interview and Diary Surveys have different coverages
  - Interview Survey does not collect expenditures on nonprescription drugs or detailed food and clothing expenditures.
  - Diary Survey does not collect overnight travel, insurance reimbursements and detailed financial information
- Consider:
  - To identify what each survey covers, see the CE profile.

Impact of survey design on PUMD

- Survey design affects your ability to interpret the data
  - Bundle individual items into broad categories, i.e., tomatoes into fresh vegetables
  - Allocate bundled items to individual items, i.e., internet, TV, and telephone, i.e., UTA file has 40.8 % allocated and imputed in 2021
  - Impute or estimate missing items, i.e., income and income tax liabilities

Impact of allocation and imputation on EXPN files, 2021

File	Percent	File	Percent	File	Percent
UTA	40.85	TRF	17.14	MDC	08.91
UTC	29.20	HHM	17.90	CRB	08.77
OPI	28.60	FRA	13.99	OVC	06.78
VEQ	22.47	OPD	13.86	CRA	06.53
OPH	19.56	EDA	13.72	LSD	06.11
INB	18.92	TRB	12.40	OVB	05.24
VLR	18.71	APA	11.92	APB	05.24
OPB	18.38	MDB	11.29		

List of EXPN files

File	Description
APA	Appliances, Household Equipment, and Other Selected Items, major appliances
APB	Appliances, Household Equipment, and Other Selected Items, minor appliances
CRA	Construction, Repairs, Alterations, and Maintenance of Property - Screening Questions
CRB	Construction, Repairs, Alterations, and Maintenance of Owned and Rented Property - Job Description
EDA	Educational Expenses
FRA	Home Furnishings and Related Household Items - Purchases
HHM	Hospitalization and Health Insurance
INB	Insurance Other Than Health – Detailed Questions
LSD	Rented and Leased Vehicles
MDB	Medical and Health Expenditures – Payments For Medical Expenses
MDC	Medical and Health Expenditures – Reimbursements For Medical Expenses
OVB	Owned Vehicles – Detailed Questions
OVC	Owned Vehicles – Disposal of Vehicles

List of EXPN files

File	Description
OPB/D/H/I	Owned Living Quarters and Other Owned Real Estate
TRB	Trips and Vacations – Detailed Questions
TRF	Trips and Vacations – Local Overnight Stays
UTA	Utilities and Fuels for Owned and Rented Properties – Telephone Expenses
UTC	Utilities and Fuels for Owned and Rented Properties – Additional Telephone Expenses
VEQ	Vehicle Operating Expenses – Vehicle Maintenance and Repair
VLR	Vehicle Operating Expenses – Licensing, Registration, and Inspection of Vehicles

Impact of topcoding on PUMD

- Topcoding affects your ability to interpret the data
  - Adjust data to protect respondents' privacy, i.e., geographic data and high income
- Impact ranges widely by data type
  - Geographic data: PSUs over 50% topcoded
  - Income data: 90th percentile over 50% topcoded

Impact of topcoding on geographic data, 2021

Interview Survey

Geography	Topcoded
PSU	56%
State	9%
Division	7%
Region	2%

Diary Survey

Geography	Topcoded
PSU	53%
State	8%
Division	6%
Region	2%

Impact of top-coding on income data, 2021

Interview Survey

Decile	Topcoded
10th	50%
9th	8%
8th	5%
7th	2%
6th	2%
5th	0%
4th	1%
3rd	0%
2nd	0%
1st	1%

Diary Survey

Decile	Topcoded
10th	53%
9th	10%
8th	6%
7th	3%
6th	2%
5th	1%
4th	2%
3rd	1%
2nd	0%
1st	0%

Survey materials

- **Questionnaires and forms for Interview and Diary Survey** ask respondents about expenditures and income. Both surveys use a computer assisted personal interviewing (CAPI) instrument. Diary Survey also uses a paper form.
- **Information Booklets** provide response options to survey questions, conceptual definitions, examples and privacy statements, and administrative assistance.

Effects of data variability

- Survey data are subject to variability, which may reduce the data's usefulness
- Consider:
  - Whether your data have sufficient precision for your research needs
  - How does variability of Consumer Expenditure data impact your analysis?

Analysis of individual CUs over time

- Interview Survey:
  - Expenditure data for up to 4 quarters
  - Income data covers the entire year
- Diary Survey for up to 2 consecutive weeks

- Consider:
  - Limit research to available timeframes
  - Analyze data by groups to expand the time frame (To identify which survey CE sources data from, see source selection file)

Analysis of individual CUs over time

- For more information on the effect of non-response and the bias it creates, see A Nonresponse Bias Study of the Consumer Expenditure Survey for the Ten-Year Period 2010-2019

Analysis of group data over time

- CE survey changes
  - Census draws the sample addresses from the Master Address File (MAF) every 10 years
  - BLS refreshes the sample set of primary sample units (PSU) every year
  - CE may add, drop, or merge questions or items
- Consider:
  - Use large aggregates: Fresh fruit instead of apples
  - Review underlying UCCs over time for changes

Hierarchical grouping files

- Hierarchical groupings (zip) includes a description of each UCC with its hierarchical standing within each expenditure or income category for a given year. For available 1996 forward, these three hierarchical groupings are available:
  - **Integrated groupings** lists UCCs that the CE tables use, and identifies the survey source for the UCCs. These files use this naming convention: CE-HG-Integ-2017
  - **Interview groupings** lists the UCCs from the Interview Survey. These files use this naming convention: CE-HG-Inter-2017. Not available for 1996
  - **Diary groupings** list the UCCs from the Diary Survey. These files use this naming convention: CE-HG-Diary-2017. Not available for 1996
- See PUMD documentation page

Supplement CE PUMD with non-CE datasets

- Consider:
  - Ensure the concepts match before using two datasets as one
    - Variables: Name of data and definition
    - Methodology: Collection and aggregation
    - Presentation: Time frame, unit, and geography
  - Adjust the data to increase compatibility
  - Review the CE Data Comparisons

Income tax concepts comparison

(Differences in red)

Concept	CE (BLS)	IRS	CPS (Census)	CBO
Name of series	Average income tax	Average income tax return	Household income tax	Average personal tax
Concepts included	Individual income taxes minus tax credits	Individual income revenue minus tax credits plus trust accumulation distribution	Individual income taxes minus tax credits	Individual income taxes plus payroll taxes, corporate income taxes, and excise taxes
Basic unit	Household	Tax return	Household	Household
Federal	Federal	Federal	Federal	Federal
Method	Model	Sample of actual tax returns	Model	Model

Income tax format comparison (Differences in red)

Format	CE (BLS)	IRS	CPS (Census)	CBO
People counted	People living financially independent	People filing one tax return	People living in one housing unit	People living in one housing unit
Rank by	Population weighted income	Adjusted gross income	No ranking	Size-adjusted income
Presen-tation	Income quintile	Income segment	No rank (mircodata)	Income quintile
Currency	U.S. dollars	U.S. dollars	U.S. dollars	U.S. dollars
Year	2013	2012	2013	2011

Use the CE tables to familiarize yourself with CE data

- CE tables and LABSTAT database provide means and aggregates for CE data by demographic characteristics
- Consider:
  - Use tables as reference for your estimates
  - Use tables for sense of trends

CE tables

- Topline tables provide 12 months data for all CUs for multiple years or in greater detail
- Calendar year and midyear means tables by demographic characteristics provide 12 months means, shares across all items, and variances for two time periods
- Calendar year aggregate shares tables by demographic characteristics provide 12 months aggregate expenditures and shares across demographic groups
- Geographic tables provide 24 months means data by state, region, Census Divisions, and population size of area of residence
- Cross-tabulated tables provide 24 months means data by two socio-economic characteristics

Want more information?

- Survey materials: How did we ask for the data?
- Data flags: How we adjust the data?
- Protection of respondent confidentiality page: How does CE topcode data?
- PUMD Getting Started Guide: What are PUMD methods?
- Tables Getting Started Guide: Considerations for tables

Thank you!

Aaron Cobet

Senior Economist, Consumer Expenditure Surveys

(202)-691-5018

Cobet.Aaron@bls.gov