Getting started with Consumer Expenditure Survey (CE) Public-Use Microdata (PUMD)*

Division of Consumer Expenditure Survey, Bureau of Labor Statistics†

July 12, 2012

Contents

---

1 Introduction

1.1 Purpose

This document is intended for the novice user of CE PUMD. It will present a brief overview of the PUMD file structure, PUMD file contents, and available user documentation.

1.2 Do I really need CE PUMD for my project?

If you are interested in means or aggregates by groups, then you might be able to find what you are looking for in the tables that the CE publishes.

The online CE tabular data, available at `http://www.bls.gov/cex/tables.htm`, are broken out by various demographic characteristics; including, age of reference person, composition and size of consumer unit, education of reference person, housing tenure, income before taxes, occupation of reference person, and race of reference person. In addition to mean expenditures, there are also tables with expenditure shares, aggregate expenditure shares, and standard errors of the means. The tabular data are constructed from internal CE microdata, and are widely used by researchers, journalists, businessmen, policy makers, academics, non-profits, students, and local, state, federal, and foreign governments.

If you are looking for more detailed expenditures than what are available in the online tables, then you can request disaggregated tables ("prepubs") from our office: `http://www.bls.gov/cex/csxcont.htm`. The disaggregated tables will have more granular expenditures such as "bacon" or "pork chops" instead of "pork." You can see the level of detail available in the tables by looking at the underlying items in the glossary of expenditures terms: `http://www.bls.gov/cex/csxgloss.htm#expn`.

If your research needs cannot be met with the available tables, then the PUMD files may be for you. With the CE PUMD files, you will be able to calculate means and other statistics by demographic groups that you define, as well as conduct more complex analysis such as regression analysis or analysis of variance. The CE PUMD is a large, complex annual dataset that requires an extensive time commitment by users to study the user documentations to learn how to use the data correctly. In addition to the time needed to learn how to use the files, users must also have a good working knowledge of an advanced statistical software package, such as SAS, SPSS, or STATA, to manipulate the files. Because of the large file sizes, it may be difficult to use a package such as Microsoft Excel to conduct analysis.

If you have any questions about whether the CE PUMD is right for you, our staff would be happy to give you guidance: `http://www.bls.gov/cex/csxcont.htm`. We also host an annual CE Microdata Users' Workshop to help our users learn how to use the PUMD files through hands-on practical training. Look for an announcement on the CE website for our next upcoming workshop.

2 PUMD file structure

2.1 Interview survey data

2.1.1 Quarterly files

"Quarter" refers to the calendar quarter in which the interview occurred. For example, any Consumer Unit (CU) interviewed in April, May, or June would have their data stored in the quarter 2 (Q2) datasets. During an interview, the CU is asked to report expenditures for a reference period of three months. So, for a CU interviewed in April, their expenditures in the Q2 file are for January, February, and March. (This is important to remember when calculating calendar year estimates.) Five quarters of interviews are included in each PUMD release, Q1 of (RELEASE YEAR) through Q1 of (RELEASE YEAR + 1). There are five types of data files organized by quarter:

1. FMLY-a file with characteristics, income, weights, and summary level expenditures for the CU.

2. MEMB-a file with characteristics and income for each member in the CU.

3. MTAB-a detailed monthly expenditure file categorized by Universal Classification Code (UCC).

4. ITAB-a household monthly income file categorized by UCC.

5. ITAB_Impute-a household monthly imputed income file categorized by UCC.[1]

2.1.2 "Annual" or "5 quarter" files

The Interview EXPN files contain expenditure and non-expenditure information collected directly from all major sections of the quarterly Interview survey. For each PUMD release, there are roughly 50 annual EXPN datasets, with each dataset corresponding to a specific section of the Interview survey (http://www.bls.gov/cex/csxsurveyforms.htm#interview). Because these datasets correspond directly to the Interview survey sections, there are varying reference periods for the expenditures contained within each EXPN dataset, dependent upon the questions within each section. Each annual EXPN dataset is made up of the five quarters of data within each release.

2.1.3 Paradata

Beginning in 2009, the CE began releasing paradata, which is data about the interview process. Paradata is available in two datasets:

1. FPAR-a file with data about the survey, including timing for each section and whether the respondent used records.

2. MCHI-a file with data about the contact history between the field representative and the respondent, including reasons for interview refusal and time of contact.

Typically, each paradata file has nine quarters of data, starting with Q1 from the (RELEASE YEAR - 1) to Q1 from the (RELEASE YEAR + 1). Nine quarters are included within each PARADATA file so users can have paradata records across all interviews for a given CU. (A CU completing their

---

[1]Available beginning in 2004

last interview during Q1 of (RELEASE YEAR) has the potential to have had their first (bounding) interview during Q1 of (RELEASE YEAR - 1)).[2]

2.2 Diary Survey Data

2.2.1 Quarterly files

"Quarter" refers to the calendar quarter in which the Diary booklet was placed in the home of the CU by the Census Field Representative. All Diary files are organized as quarterly files. Those include:

1. FMLY-a file with characteristics, income, weights, and summary level expenditures for the CU.

2. MEMB-a file with characteristics and income for each member in the CU.

3. EXPN-a detailed weekly expenditure file categorized by UCC.

4. DTAB-a detailed annual income file categorized by UCC.

5. DTAB_Impute-a household imputed income file categorized by UCC.[3]

2.3 File naming conventions

The naming conventions used for each PUMD release is common to both Interview and Diary data. YY stands for the two digit year and Q stands for the calendar quarter of the Interview or calendar quarter of the placement of the Diary booklet. Below, the example file directories and names are for SAS datasets. The only difference in the naming convention for other dataset types (i.e., SPSS, STATA, etc.) is the file extension.

2.3.1 File names and directories

Interview quarterly files

FMLY \INTRVW YY\FMLI YYQ.sas7bdat

MEMB \INTRVW YY\MEMI YYQ.sas7bdat

MTAB \INTRVW YY\MTBI YYQ.sas7bdat

ITAB \INTRVW YY\ITBI YYQ.sas7bdat

ITAB_Impute \INTRVW YY\ITII YYQ.sas7bdat

Interview annual files
Note: Only selected EXPN files are listed here. All other EXPN files follow the same convention.

CLC \EXPN YY\CLC YY.sas7bdat

---

[2]With the 2009 PUMD release, historical paradata for 2005-2007 were released. The dataset for 2005 contains 3 quarters of data, while the datasets for 2006 and 2007 contains 4 quarters of data each.

[3]Available beginning in 2004.

IHB \EXPN *YY* \IHB *YY*.sas7bdat

VEQ \EXPN *YY* \VEQ *YY*.sas7bdat

Diary quarterly files

FMLY \DIARY *YY* \FMLD *YYQ*.sas7bdat

MEMB \DIARY *YY* \MEMD *YYQ*.sas7bdat

EXPN \DIARY *YY* \EXPD *YYQ*.sas7bdat

DTAB \DIARY *YY* \DTBD *YYQ*.sas7bdat

DTAB_Impute \DIARY *YY* \DTID *YYQ*.sas7bdat

2.3.2 What's the "X" about?

You may have noticed there is an "X" in the name for some of the Interview quarterly files (eg. FMLI *YY* 1x.sas7bdat or MEMI *YY* 1x.sas7bdat). You may have also noticed that this "X" is only in the name of quarter 1 files for the current calendar year. This is because files for the first quarter of any calendar year will appear in two CE PUMD releases:

1. As the "fifth" quarter file in the previous calendar year's release,

2. And as the "first" quarter file in the current calendar year's release.

The "X" signifies that the first quarter file of the current calendar year release is not identical to the fifth quarter file of the previous calendar year release. One reason the files are not identical is because PUMD production processes use the five quarters of data within the release to calculate certain values, such as the critical and topcode values of income and expenditures. During years in which there was a sample design change, there is a bigger difference in the "X" files. For example, the CE changed its sample design in 2005Q1. When this happens, the CE draws an "overlapping sample" for 2005Q1, i.e., some CUs are for the *old* design, some CUs are for the *new* design, and some are for *both* designs.[4] For the 2004 PUMD release, the CUs in the 2005Q1 files are those for the *old* design. For the 2005 PUMD release, the CUs in the 2005Q1 files are those for the *new* design. CUs that were in both the *old* and *new* designs are contained within the 2005Q1 file for both the 2004 and 2005 PUMD releases.

3 PUMD file content overview

NEWID, the unique identifier for each CU, is included within every record on each data file. It can be used to link records of a particular CU across files. For example, linking demographic information in the Interview FMLY file to expenditure information in the MTAB file. Be aware that the samples from the Interview and Diary surveys are independent, so there is no CU contained within both the Interview and Diary files.

---

[4]CUs can be in both the *old* and *new* designs because the designs are based on areas of the United States called Primary Sampling Units (PSU). So if a CU was in a PSU that was in both designs, then they are used in both designs.

3.1 Interview quarterly files

3.1.1 FMLY-CU characteristics, income, and summary expenditure variables

- There exists one record per CU.

- Each CU is uniquely identified by NEWID.

- Variables include demographics for the reference person and spouse of reference person, income at the CU level, sample housing unit information, summary level expenditures, and weighting variables.

Note: Not all the variables within the FMLY file are shown in the sample below.

	NEWID	AGE_REF	FAM_SIZE	ERNKMTHM	FINCBTXM	FDAWAYPQ	FDAWAYCQ	FINLWT21
1	2018195	49	2	14844.50	86000.0	151.6667	553.3333	12108.01
2	2018205	48	1	8344.50	36000.0	650.0000	0.0000	12455.42
3	2018215	74	2	12737.95	69778.0	415.0000	0.0000	16235.13
4	2018245	60	2	6127.25	41285.6	86.6667	173.3333	15614.94
5	2018265	41	1	9236.75	34000.0	108.3333	216.6667	27744.26
6	2018275	52	1	11298.25	60000.0	650.0000	0.0000	14918.52
7	2018285	57	1	6530.85	49526.0	0.0000	0.0000	10575.44
8	2018315	32	7	9582.85	66972.8	130.0000	65.0000	14128.17
9	2018325	33	5	12610.50	57229.0	223.0000	246.0000	26339.37
10	2018335	56	2	15699.65	57495.4	1300.0000	0.0000	20070.33

3.1.2 MEMB-CU member characteristics and income

- There exist multiple records per CU.

- There exists one record per member.

- Unique records are defined by the combination of NEWID and MEMBNO.

- Variables include demographics about CU members, member level income, and member relationship status to the reference person.

Note: Not all the variables within the MEMB file are shown in the sample below.

	NEWID	MEMBNO	AGE	CU_CODE	EARNER	EDUCA	SEX	MEMBRACE	SALARYXM
1	2018195	1	49	1	1	43	2	1	3403
2	2018195	3	17	3	2	9	2	1	1
3	2018205	1	48	1	1	40	1	1	2257
4	2018215	1	74	1	1	39	1	1	3856
5	2018215	2	67	2	1	40	2	1	673
6	2018245	1	60	1	1	39	2	1	1945
7	2018245	2	63	6	2	7	2	1	1
8	2018265	1	41	1	1	40	1	1	2069
9	2018275	1	52	1	1	43	2	1	3206
10	2018285	1	57	1	1	40	2	1	2594

3.1.3 MTAB-Monthly expenditure file classified by UCC

- There exist multiple records per CU.

- There exists one record per expenditure per purchase month/year.

- Prior to the 2010 PUMD release, unique records were not identified. From the 2010 release forward, unique records can be identified using a combination of NEWID, EXPNAME, SEQNO, and ALCNO.

- The monthly expenditure amounts are mapped from corresponding EXPN variables.

Note: Not all the variables within the MTAB file are shown below.

	NEWID	UCC	COST	COST_	GIFT	PUBFLAG	REF_MO	REF_YR
1	2238655	210110	785	1	2	2	3	2010
2	2238655	210110	785	1	2	2	2	2010
3	2238655	210110	785	1	2	2	1	2010
4	2032985	210110	1165	1	2	2	3	2010
5	2032985	210110	1165	1	2	2	2	2010
6	2032985	210110	1165	1	2	2	1	2010
7	2030205	210110	2531	5	2	2	3	2010
8	2030205	210110	2531	5	2	2	2	2010
9	2030205	210110	2531	5	2	2	1	2010
10	2031355	800710	810	1	2	2	3	2010

3.1.4 ITAB-Household monthly income classified by UCC

- There exist multiple records per CU.

- There exists one record per income item per month/year.

- Unique records are defined by NEWID, UCC, REFMO, and REFYR.

- The monthly amounts are mapped from the corresponding CU-level imputed income mean values.

Note: This ITAB sample shows ALL variables that are in the full ITAB file.[5]

	NEWID	REFMO	REFYR	UCC	PUBFLAG	VALUE	VALUE_
1	2018195	3	2010	2010	2	0.0000	NA
2	2018195	4	2010	2010	2	0.0000	NA
3	2018195	5	2010	2010	2	0.0000	NA
4	2018195	3	2010	2020	2	0.0000	NA
5	2018195	4	2010	2020	2	0.0000	NA
6	2018195	5	2010	2020	2	0.0000	NA
7	2018195	3	2010	800910	2	151.6667	NA

---

[5]The dataset views in this document were generated using R, which represents missing values as "NA." Other statistical software packages may represent missing values as a period, blank space, or in another manner.

```
8  2018195     4  2010 800910        2 151.6667    NA
9  2018195     5  2010 800910        2 151.6667    NA
10 2018195     3  2010 800932        2   6.2500    NA
```

3.1.5   ITAB_Impute-Household monthly imputed income file classified by UCC.

- There exist multiple records per CU.

- There exists one record per income item per month/year.

- Unique records are defined by NEWID, UCC, REFMO, REFYR, and IMPNUM.

- The monthly amounts are mapped from the corresponding CU-level imputed income imputes(5 imputes per income item).

- This file is necessary to calculate variances and standard errors for imputed income.

Note: This ITAB_Impute sample shows ALL variables that are in the full ITAB_Impute file.

```
      NEWID REFMO REFYR    UCC PUBFLAG    VALUE VALUE_ IMPNUM
1  2018195     3  2010 800940        2 439.9167     NA      1
2  2018195     3  2010 800940        2 439.9167     NA      2
3  2018195     3  2010 800940        2 439.9167     NA      3
4  2018195     3  2010 800940        2 439.9167     NA      4
5  2018195     3  2010 800940        2 439.9167     NA      5
6  2018195     4  2010 800940        2 439.9167     NA      1
7  2018195     4  2010 800940        2 439.9167     NA      2
8  2018195     4  2010 800940        2 439.9167     NA      3
9  2018195     4  2010 800940        2 439.9167     NA      4
10 2018195     4  2010 800940        2 439.9167     NA      5
```

3.2   Interview annual files–EXPN

- Each EXPN file is different, with some having multiple records per CU and others with one record or no records per CU interviewed each quarter.

- Unique records are defined differently in each file.

- Common variables to all EXPN files: **NEWID** is the unique CU identifier. **QYEAR** defines the year and interview quarter for a given record. **SEQNO** is assigned sequentially during the interview as each expenditure record is recorded into the database. **ALCNO** is assigned sequentially for each record that has been allocated from one expenditure during the process. For example, a CU may report spending $50 on a pair of men's pants and a shirt. The CE will allocate out that record into separate records, one for men's pants and shorts($30) and one for men's shirts($20). **REC_ORIG** describes the origin of the record, i.e., whether the expenditure data was collected during the current interview, expenditure data collected for a current reference period in a previous interview, household inventory data collected during a previous interview that was pulled forward to the current, or data that was created by the CE processing edit system.

- Variables include those directly collected from the Interview survey questionnaire on expenditures and screener questions. Variables that are recalculated by CE are also included.

Examples:

3.2.1 CLC-Sewing materials

- Unique records in CLC are defined by the combination of NEWID, SEQNO, and ALCNO.

- **SEWINGY** is an item code variable for the type of sewing expenditure (materials for making clothes, sewing notions, etc.)

- **SEWINGMO** is an indicator variable for the month in which the purchase occurred.

- **SEWINGX** is the amount of the sewing expense.

Note: Not all the variables within the CLC file are shown below.

```
    QYEAR    NEWID SEQNO ALCNO SEWINGY SEWINGMO SEWINGX
1  20101 1995545    24     0     400        2       42
2  20101 1996005    20     0     430        2        3
3  20101 1996265    20     0     400       12      180
4  20101 1996305    18     0     420        1       42
5  20101 1996385    13     0     400       11      104
6  20101 1997045    25     0     420       12      142
7  20101 1997095    21     0     430       12      123
8  20101 1997345    33     0     420        1      123
9  20101 1997505    32     0     410       11       12
10 20101 1997505    33     0     410        1      171
```

3.2.2 IHB-Hospitalization and health insurance detailed questions

- Unique records in IHB defined by NEWID, SEQNO, ALCNO, and HHIPDLIB.

- **HHIPDLIB** is the policy number for the health insurance plan. This number is derived by the order in which the respondent lists their health insurance plans.

- **HHICOVQ** is the number of people covered by health insurance plan.

- **QHI3MCX** is the amount paid in health insurance premiums during the reference period (derived by CE).

- **HHICODE** is an indicator variable for the type of insurance plan (HMO, FFS, Medicare supplement, or other special purpose plan).

Note: Not all the variables within the IHB file are shown below.

```
     QYEAR    NEWID SEQNO ALCNO HHIPDLIB HHICOVQ QHI3MCX HHICODE
1   20101 1995545    38     0        2        1      37        2
2   20101 1995545    39     0        3        1     949        4
3   20101 1995545    40     0        4        1     156        4
4   20101 1995595    17     0        1        2     367        2
5   20101 1995655    21     0        1        2    1119        1
6   20101 1995685    24     0        1        1    1117        1
7   20101 1995705    27     0        1        2    1343        2
8   20101 1995705    28     0        2        2     821        4
9   20101 1995725    23     0        1        1     157        2
10  20101 1995745    21     0        1        1    1248        1
```

3.2.3   VEQ-Vehicle maintenance and repair

- Unique records in VEQ defined by NEWID, SEQNO, and ALCNO.

- **VOPSERVY** is an indicator variable that decribes the type of maintenance or repair untaken.

- **VOPMOA** is an indicator variable for the month in which the expense occurred.

- **VOPEXPX** is the total cost of the maintenance or repair expense.

Note: Not all the variables within the VEQ file are shown below.

```
     QYEAR    NEWID SEQNO ALCNO VOPSERVY VOPMOA VOPEXPX
1   20101 1995545    72     0      365      12     844
2   20101 1995545    73     0      365       1     844
3   20101 1995545    74     0      365       2    1024
4   20101 1995585    58     0      500      12     792
5   20101 1995585    58     1      160      12       1
6   20101 1995585    58     2      150      12       1
7   20101 1995615    29     0      100       1     553
8   20101 1995745    40     0      320       1     735
9   20101 1995785    50     0      100      12     508
10  20101 1995805    50     0      340       2      56
```

3.3   Diary files

- Most Diary files are analogous to Interview files.

- There are slightly different variable names in the Diary files.

- The summary level expenditures in the Diary FMLY file are not as extensive as those in the Interview file.

3.3.1 FMLY-CU characteristics, income, weights, and summary level expenditures

- Analogous to FMLY file in Interview.

Note: Not all the variables within the FMLY file are shown below.

```
      NEWID STRTDAY STRTMNTH WEEKI CUTENURE REF_RACE FINCBEFM   FOODHOME FINLWT21
1  1071891       9        5     1        1        4  94462.6   99.99996 33205.63
2  1071892       9        5     2        1        4  75835.4  100.00000 33205.63
3  1071901      18        4     1        1        1  73674.0   62.28000 46815.81
4  1071902      27        4     2        1        1  73674.0   78.03000 46815.81
5  1071911      16        4     1        2        1  33114.2   33.10000 65162.89
6  1071912      24        4     2        2        1  34057.4   69.17000 65162.89
7  1071941      27        4     1        4        1  64000.0   99.50000 20470.70
8  1071942      27        4     2        4        1  64000.0    0.00000 20470.70
9  1071961      20        4     1        2        1  69618.0   64.23000 34713.36
10 1071962      28        4     2        2        1  62062.4   97.31001 34713.36
```

3.3.2 MEMB-CU member characteristics and income

- Analogous to MEMB file in Interview.

Note: Not all the variables within the MEMB file are shown below.

```
      NEWID MEMBNO AGE EDUCA MARITAL EMPLTYPE WAGEXM
1  1071891      1  46    43       1        1   1862
2  1071891      2  45    39       1        1    423
3  1071891      3  11    NA       5       NA      1
4  1071892      1  46    43       1        1   1581
5  1071892      2  45    39       1        1    474
6  1071892      3  11    NA       5       NA      1
7  1071901      1  74    44       1       NA      1
8  1071901      2  74    39       1       NA      1
9  1071902      1  74    44       1       NA      1
10 1071902      2  74    39       1       NA      1
```

3.3.3 EXPN-Detailed weekly expenditure file categorized by UCC

- Analogous to MTAB files in Interview.

- There are multiple records per CU.

- There is one record per expenditure per purchase month/year.

- Unique records are not defined.

- The weekly amounts are recorded directly from listings in the completed Diary.

Note: This EXPN example shows ALL variables that are in the full EXPN file.

```
    NEWID ALLOC    COST GIFT PUB_FLAG     QREDATE QREDATE_    UCC
1  1071891    1 3.09857    2        2 1105092010        3 10210
2  1071891    1 2.38117    2        2 1105092010        3 20210
3  1071891    1 3.68334    2        2 1105092010        3 20510
4  1071891    1 4.35908    2        2 1105092010        3 30110
5  1071891    1 4.63953    2        2 1105092010        3 40510
6  1071891    1 4.88788    2        2 1105092010        3 50110
7  1071891    1 3.83796    2        2 1105092010        3 60210
8  1071891    1 5.17891    2        2 1105092010        3 60310
9  1071891    1 2.16478    2        2 1105092010        3 90110
10 1071891    1 4.18612    2        2 1105092010        3 90210
```

3.3.4 DTAB-Detailed annual income file categorized by UCC

- Analogous to ITAB file in Interview.

Note: This DTAB example shows ALL variables that are in the full DTAB file.

```
    NEWID    UCC AMOUNT AMOUNT_ PUB_FLAG
1  1071891 800940   6998      NA        1
2  1071891 900000  94463      NA        1
3  1071891 980000  94463      NA        1
4  1071891 980010      3      NA        1
5  1071891 980020     46      NA        1
6  1071891 980030      2      NA        1
7  1071891 980040      2      NA        1
8  1071891 980050      1      NA        1
9  1071891 980070  94463      NA        1
10 1071891 980090    100      NA        1
```

3.3.5 DTAB_Impute-household imputed income file categorized by UCC.

- Analogous to the ITAB_Impute file.

Note: This DTAB_Impute example shows ALL variables that are in the full DTAB_Impute file.

```
    NEWID    UCC PUB_FLAG AMOUNT AMOUNT_ IMPNUM
1  1076521 800940        1   4570      NA      1
2  1076521 800940        1   6056      NA      2
3  1076521 800940        1   5004      NA      3
4  1076521 800940        1   5163      NA      4
5  1076521 800940        1   4910      NA      5
6  1076521 900000        1  59735      NA      1
7  1076521 900000        1  79163      NA      2
8  1076521 900000        1  65411      NA      3
9  1076521 900000        1  67485      NA      4
10 1076521 900000        1  64179      NA      5
```

4   Documentation

To aid in using the PUMD, the CE produces annual user guides and data dictionaries. Since the 2007 PUMD release, the user guides and data dictionaries, for both Interview and Diary data, have been parsed out into separate documents. Prior to 2007, the user guides and data dictionaries were combined in one document. The user documentations include:

- A summary of changes from the previous release of CE PUMD.

- More in depth information of the PUMD files.

- Information on CE PUMD nondisclosure requirements.

- Formulas and procedures for estimating aggregate statistics from the data.

- Data reliability statements.

- And other information, including information on sampling design and cooperation levels.

The data dictionaries include a listing of all the CE PUMD variables and descriptions.

User guides and data dictionaries dating back to the 1996 release, along with other miscellaneous documentations, can be accessed online at the CE website, http://www.bls.gov/cex/csxmicrodoc.htm.