

R: Innovating at the Bureau of Labor Statistics

Arcenis Rojas

Economist

Division of Consumer Expenditure Surveys

Federal Committee on Statistical Methodology

March 2018



Overview

- **IPP:** Division of International Prices
- **PPI:** Division of Industrial Prices and Price Indexes
- **CE:** Division of Consumer Expenditure Survey
- **OCWC:** Office of Compensation and Working Conditions
- **OSMR:** Office of Survey Methods and Research



Overview

- Automation (IPP)
- Quality control (PPI)
- Real-time response rates (OCWC)
- Data visualization (CE)
- Other R Shiny applications
- R packages



R Shiny Applications



Sample Refinement Automation

■ International Prices Program

- ▶ Receive data from Census and Customs
- ▶ Must verify Establishment ID Number (EIN), name, and address to provide to field economists
- ▶ 1700 export collections units per sample
- ▶ 2400 import collection units per sample
- ▶ 6 IPP sample team members
- ▶ 16 copies, 20 pastes, and 46 clicks per unit

Data Sources

Access secure data portal U.S. Customs and Border Protection

NOTICE TO ALL USERS REGARDING
You are about to access a Department of Homeland Security computer system. This computer system and data therein are property of the U.S. Government and provided for official U.S. Government information and use. There is no expectation of privacy when you use this computer system. The use of a password or any other security measure does not establish an expectation of privacy. By using this system, you consent to the terms set forth in this notice. You may not possess classified national security information on this computer system. Access to this system is restricted to authorized users only. Unauthorized access, use, or modification of the system or of data contained therein, or in transit to/from this system, may constitute a violation of section 1820 of title 22 of the U.S. Code and other criminal laws. Anyone who accesses a Federal computer system without authorization or exceeds authorized access, in violation of section 1820 of title 22 of the U.S. Code and other criminal laws, or who knowingly provides or attempts to provide unauthorized access to such a system, is subject to penalties, fines or imprisonment. This computer system and any related equipment is subject to searching for administrative oversight, law enforcement, criminal investigative purposes, input for intelligence gathering or analysis, and to ensure proper performance of applicable security, hardware and procedures. Such may conduct searching activities without further notice.

NOTICE TO ALL USERS REGARDING
By logging in to the ACE Portal, you agree to be bound by the language set forth in the [Terms and Conditions](#) document, published on May 16, 2007, last updated on July 7, 2008.
It is mandatory that all ACE users maintain a current email address within their ACE user profile.

Log in Information
Enter your ACE UserID and Password to log in:

UserID:
Password:

[Forgot Your Password?](#)

If you need assistance with the portal, please contact the CBP Technology Support Center at 1-866-530-4177 for Trade and PUA users, or 1-800-827-8770 for CBP personnel.

Please note that all users have a timeout of inactivity. The system will log you off automatically and ask you to log in again.

LDB Extract System

Home | Extract Data | Four-Quarter Extract | Help

Robert Sutton
Logout

[Saved Queries](#)

[View Status of Running Queries](#)

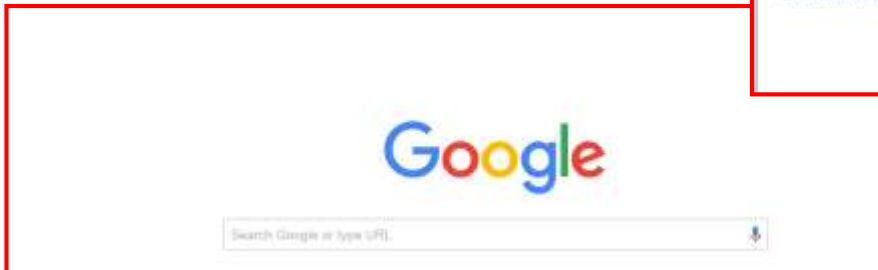
Please let us know if you have questions about the LDB or if we can assist in any way. Help requests should be submitted via the "new topic LANS HelpDesk" button on the BLS Contact Internet ([http://www.bls.gov/helpdesk/](#)). Select "Program and Application Specific" under General Types. Select "Quarterly Census of Employment and Wages (QCEW)" under Program and Application Specific section. Select "Longitudinal Database (LDB)" sub-category. Provide year, Problem Description, Attachments (if any), and User Importance/Impact and click the "Submit" button. Comments, questions, or concerns about the LDB can be submitted directly via e-mail to LDB_Comm@bls.gov.

LDB News

- The Division of Business Establishment Systems has updated the Longitudinal Database (LDB) with data for the third quarter of 2016, and the data are now available to users. The LDB presently contains data from the first quarter of 1990 through the third quarter of 2016. For all states, with the quarter's lead, back-quarter corrections were applied for 2016:1 and 2016:2.

We anticipate that data for the next several quarters will be available according to the following tentative schedule:

Data Through Date Available
Q4 2016 week of 12/13



Google

Search Google or type URL



Enter a Collection Unit and Sample

Collection Unit (last 4 digits):

1

0345

Sample:

M43

Company Name (edited):

2

XXXXXX

Corp Div:

33142-000

Street:

YYYYYYYYYYYYYYYYYY

Search in New Tab

5

Open SOS Website

Create Future Note

Sampled Company EIN Spreadsheet ACE LDB Progress Report Help

CORP_DIV	EIN	COMPANY_NAME	ADDRESS_1	CITY	STATE	ZIP_CODE_10	IPP_history
33142-000	0000000000	XXXXXXXXXX	YYYYYYYYYYYYYYYYYY	HUNTSVILLE	AL	350062807	old company

ACE: matching on sampled EIN

3

Entry.Date	Importer.No.	Importer.Name	Mailing.Line.1	M.City	M.St	M.Zip	Physical.Street	P.City	P.St	P.Zip
2016-12-06	00-000000000	XXXXXXXXXX	YYYYYYYYYYYYYYYYYY	HUNTSVILLE	AL	35006-2807	same			

RTS Master Listing: matching on Name, Corp Div, or Street displayed on the left

Show 100 entries

Search:

4

RTS_ID	COMPANY_NAME	DIVISION_NAME	STREET	CITY_STATE_ZIP	SOURCE
All	All	All	All	All	All
33142-000	XXXXXX		YYYYYYYYYYYYYYYYYY	HUNTSVILLE AL 35006	address_init
33142-000	XXXXXX		YYYYYYYYYYYYYYYYYY	HUNTSVILLE AL 35006	sampling_unit
33142-000	XXXXXX		YYYYYYYYYYYYYYYYYY	HUNTSVILLE AL 350062807	address_init
33142-000	XXXXXX		YYYYYYYYYYYYYYYYYY	HUNTSVILLE AL 350062807	sampling_unit
33142-000	XXXXXXXXXX		YYYYYYYYYYYYYYYYYY	HUNTSVILLE AL 35006	sampling_unit
33142-000	XXXXXXXXXX		YYYYYYYYYYYYYYYYYY	HUNTSVILLE AL 35006	old_reporter
33142-000	XXXXXXXXXX		YYYYYYYYYYYYYYYYYY	HUNTSVILLE AL 350062807	address_init
33142-000	XXXXXXXXXX		YYYYYYYYYYYYYYYYYY	HUNTSVILLE AL 350062807	sampling_unit
33142-000	XXXXXXXXXX		Y.YYYYYYYYYYY	HUNTSVILLE AL 350062807	CURRENT_REPORTER
33142-000	XXXXXXXXXX	////////////////////	YYYYYYYYYYYYYYYYYY	HUNTSVILLE AL 35006	address_init

Showing 1 to 10 of 10 entries

Previous 1 Next

Enter a Collection Unit and Sample

Collection Unit (last 4 digits):

Sample:

Company Name (edited):

Corp Div:

Street:

[Search in New Tab](#)

[Open Google Maps](#)

[Open SOS Website](#)

[Create Future Note](#)

Left Side



Right Side

Sampled Company								EIN Spreadsheet	ACE	LDB	Progress Report	Help
CORP_DIV	EIN	COMPANY_NAME	ADDRESS_1	CITY_	STATE_	ZIP_CODE_10	IPP_history					

ACE: matching on sampled EIN

Entry.Date	Importer.No.	Importer.Name	Mailing.Line.1	M.City	M.St	M.Zip	Physical.Street	P.City	P.St	P.Zip
------------	--------------	---------------	----------------	--------	------	-------	-----------------	--------	------	-------

RTS Master Listing: matching on Name, Corp Div, or Street displayed on the left



Search Results

Enter a Collection Unit and Sample

Collection Unit (last 4 digits):

Sample:

Company Name (edited):

Corp Div:

Street:

[G Search in New Tab](#)

[Q Open SOS Website](#)

[+ Create Future Note](#)

Company Name Changes

No company name changes in the reporter or old_co_name_fsn tables for Corp Div: 33142-000

Google Search: company name, city, and state

Search Images Maps Play YouTube News Gmail More

[All](#) [Images](#) [Videos](#) [News](#) [Shopping](#) [Maps](#) [Books](#)

0 results

Any time
Past hour
Past 24 hours
Past week
Past month
Past year

All results
Verbata

Carrabba's Italian Grill in Huntsville, AL
<https://www.carrabbas.com/locations/al/huntsville> +
Looking for a great Italian restaurant? Bring your family and friends to the Carrabba's Italian Grill location in Huntsville today and enjoy classic Italian dishes!

Ralphie May :: Stand Up Live Huntsville
standuplivehuntsville.laughstube.com/event.cfm?id=476375 +
Ralphie May :: Stand Up Live Huntsville ... Stand Up Live Huntsville, Huntsville, AL. Two item minimum ... to show time. Please call the Box Office at xxx.xxx.xxxx ...

Josh Blue :: Stand Up Live Huntsville
standuplivehuntsville.laughstube.com/event.cfm?id=476363 +
Josh Blue :: Stand Up Live Huntsville ... Stand Up Live Huntsville, Huntsville, AL. Two item minimum, 19 & ... to show time. Please call the Box Office at xxx.xxx.xxxx ...

Kountry Wayne :: Stand Up Live Huntsville
standuplivehuntsville.laughstube.com/event.cfm?id=480893 +
Kountry Wayne :: Stand Up Live Huntsville ... April 09, 2017 6:30 PM. Stand Up Live Huntsville, Huntsville, AL ... Please call the Box Office at xxx.xxx.xxxx ...

Chingo Bling :: Stand Up Live Huntsville
standuplivehuntsville.laughstube.com/event.cfm?id=481551 +
Chingo Bling :: Stand Up Live Huntsville ... Stand Up Live Huntsville, Huntsville, AL. Two item minimum ... to show time. Please call the Box Office at xxx.xxx.xxxx ...

Searches related to **XXXXXXXX YYYYYYYYYYYYYYYYYY HUNTSVILLE AL 35806**

carrabba's huntsville al coupons	stand up live huntsville menu
carrabba's italian grill huntsville al 35801	italian restaurants huntsville al
carrabba's parkway place mall	outback huntsville al
bonetfish grill huntsville alabama	stand up live huntsville al menu

[Advanced search](#) [Search Help](#) [Send feedback](#)

[Google Home](#) [Advertising Programs](#) [Business Solutions](#) [Privacy](#) [Terms](#)



Export Addresses at a Glance

Enter a Collection Unit and Sample

Collection Unit (last 4 digits):

Sample:

Company Name (edited):

Corp Div:

Street:

Sampled Company EIN Spreadsheet ACE LDB Progress Report Help

EIN Spreadsheet: matching on sampled EIN

pct	export_id	eistrata	name	address	city	st	zip	eistrecs	addrecs	split	flag1	flag2	problem	Mike.s.Comment
33%								2398	800	1	1			
21%								2398	510					
11%								2398	260					
7%								2398	177					
5%								2398	127					
4%								2398	93					
3%								2398	82					



Benefits of Automation

- 80-100 hours per sample of time savings
 - ▶ Much less clicking
 - ▶ Better and more thorough sample review
 - ▶ More time to review more problematic collection units

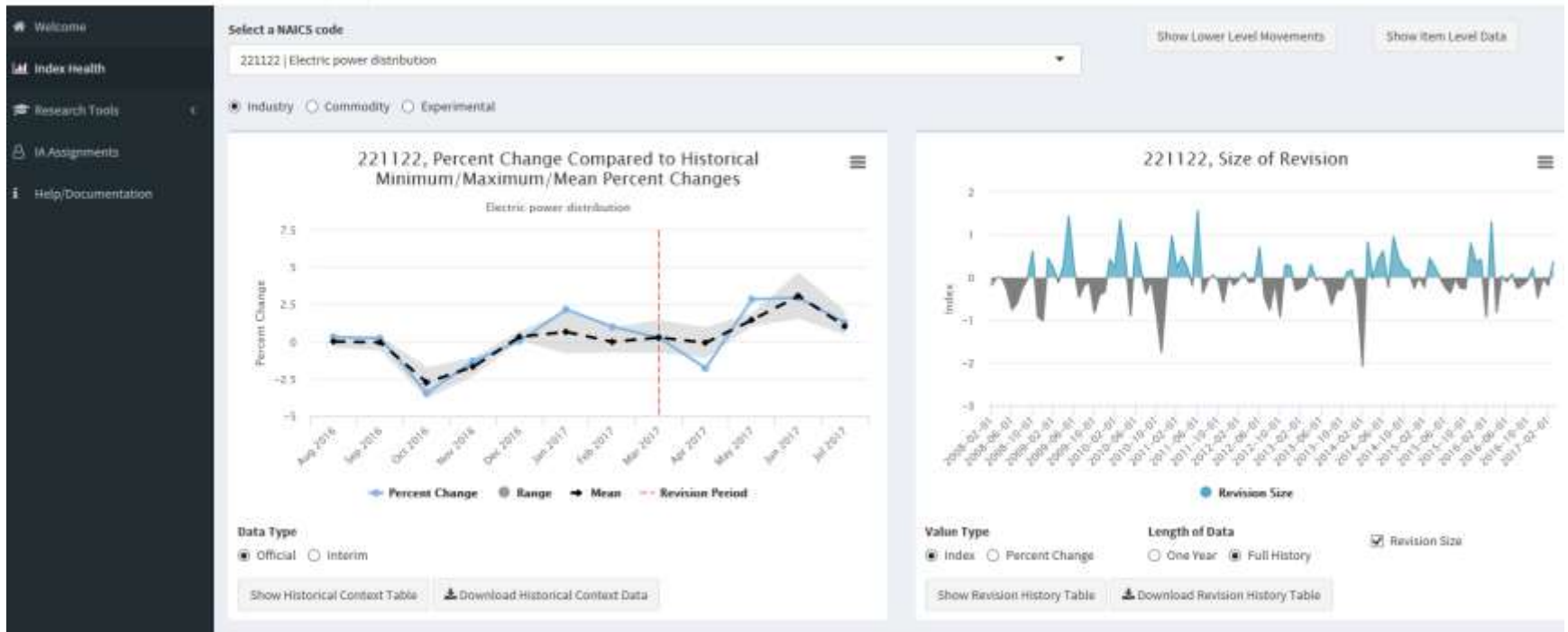
Sample Refinement Automation

- Ara Khatchadourian:
khatchadourian.ara@bls.gov
- Rob Sutton: sutton.robert@bls.gov

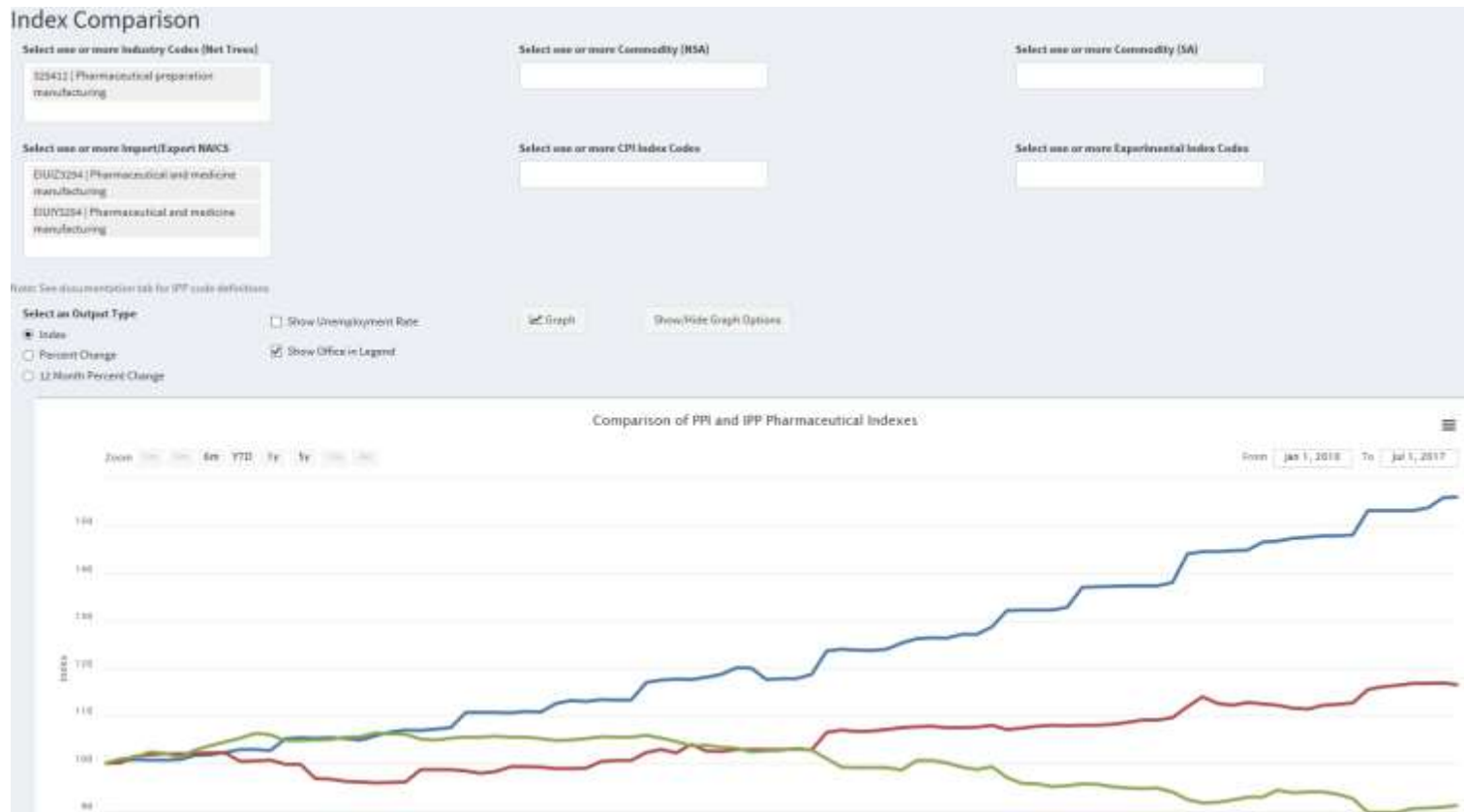


Industrial Prices Visualization Dashboard

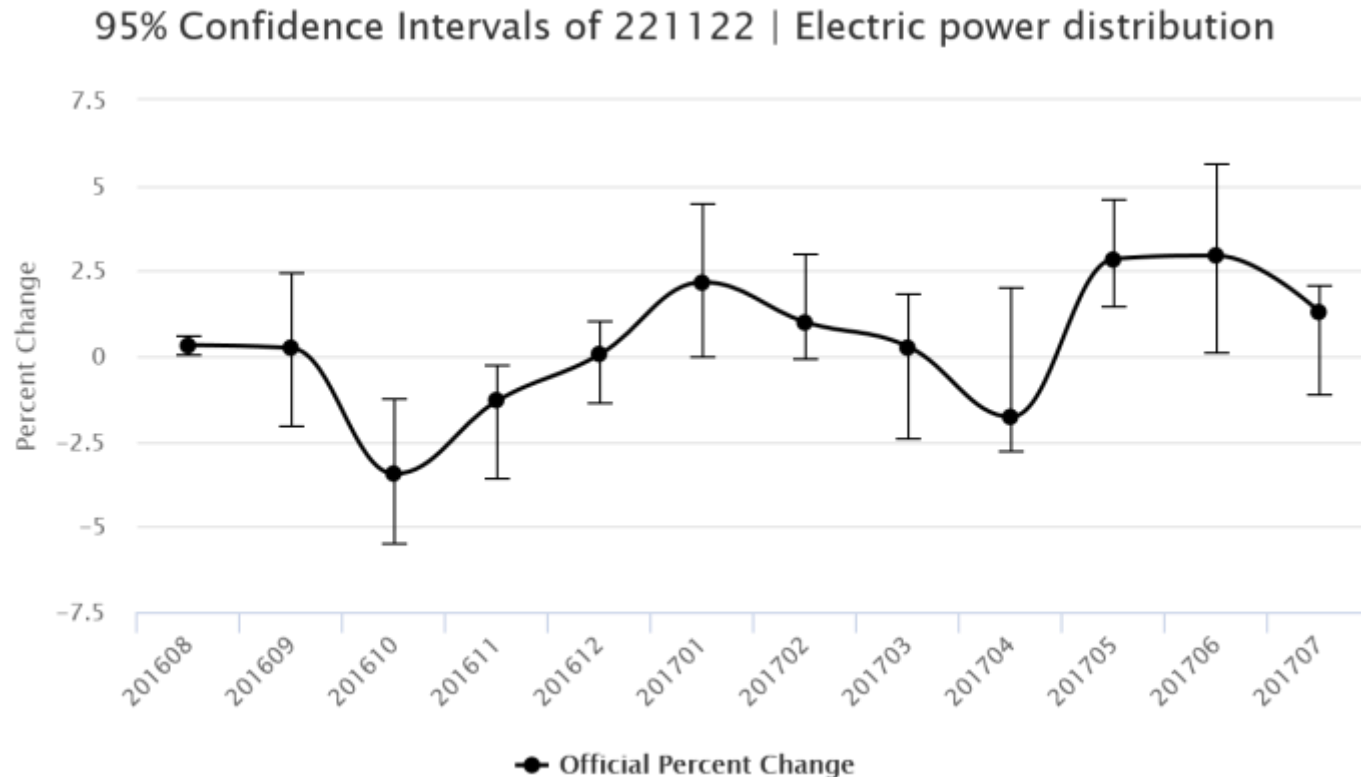
PPI Visualization Dashboard 3.02



Index Comparisons



Index Review and Revision



EMBARGOED DATA - NOT FOR PUBLIC RELEASE

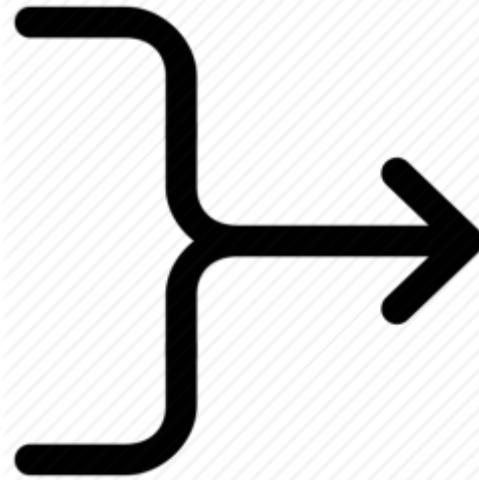


Visualization Dashboard

- Neil Wagner: wagner.neil@bls.gov
- Steve York: york.stephen@bls.gov



Interactive CE Visualization Tool



NewID	Member No.	Income
MEMI		

NewID	Race	Expenditure
FMLI		

NewID	Sequence	Allocation	UCC	Cost
EXPN				

NewID	Sequence	Allocation	UCC	Cost
MTBI				

CE Public-Use Microdata (PUMD)

■ Public-Use Microdata

- ▶ Family-level characteristics
- ▶ Expenditures by Universal Classification Code (UCC)
- ▶ Member-level characteristics
- ▶ Expenditures and their characteristics by type of expenditure (EXPN... > 50 files each year!)
- ▶ And more!

Files Required for Analysis

Family Characteristics File (34,177 Observations)

	newid	cutenure	fam_size	finlart21	ref_race	region	finobtxm	high_edu
1	02792005	Owner	1	16991.669	Black or African American	South	Second 20 Percent	High School Graduate
2	02792065	Renter	1	14256.350	White	West	Third 20 Percent	Some College or Assoc
3	02792075	Renter	3	16964.084	White	South	Second 20 Percent	Some College or Assoc
4	02792095	Owner	1	21831.304	White	South	Third 20 Percent	Less than High school
5	02792115	Renter	4	18482.601	Asian	Northeast	Third 20 Percent	Bachelor's Degree
6	02792125	Renter	3	14680.459	White	Northeast	Lowest 20 Percent	Some College or Assoc
7	02792155	Owner	1	6143.497	Asian	West	Fourth 20 Percent	Post-graduate Degree
8	02792225	Owner	1	13793.790	White	South	Lowest 20 Percent	Some College or Assoc
9	02792245	Renter	4	18926.631	White	West	Third 20 Percent	Some College or Assoc
10	02792265	Renter	2	15954.534	Asian	West	Second 20 Percent	High School Graduate
11	02792275	Owner	2	10781.501	Black or African American	South	Third 20 Percent	Bachelor's Degree
12	02792295	Renter	2	17750.726	Asian	South	Fourth 20 Percent	Bachelor's Degree
13	02792335	Owner	5 or More	18636.435	White	South	Second 20 Percent	Less than High school
14	02792385	Renter	1	17653.513	Black or African American	South	Fourth 20 Percent	Some College or Assoc
15	02792415	Owner	2	18687.147	White	Northeast	Third 20 Percent	Some College or Assoc
16	02792435	Owner	1	6171.729	Asian	West	Third 20 Percent	Bachelor's Degree
17	02792455	Renter	1	12064.521	White	Northeast	Lowest 20 Percent	High School Graduate
18	02792525	Owner	2	18876.791	Asian	West	Highest 20 Percent	Post-graduate Degree
19	02792585	Renter	5 or More	23060.929	Black or African American	Midwest	Third 20 Percent	Some College or Assoc
20	02792595	Owner	3	19092.108	Asian	West	Highest 20 Percent	Bachelor's Degree

Expenditures File (1,720,755 Observations)

	newid	ref_mo	ucc	cost
1	02792065	01	210110	650.0000
2	02792075	01	210110	258.3333
3	02792115	01	210110	1370.0000
4	02792125	01	210110	426.0000
5	02792245	01	210110	3189.0000
6	02792265	01	210110	380.0000
7	02792295	01	210110	918.3333
8	02792385	01	210110	675.0000
9	02792455	01	210110	850.0000
10	02792585	01	210110	1500.0000
11	02792815	01	210110	880.0000
12	02793085	01	210110	1050.0000
13	02793345	01	210110	850.0000
14	02793355	01	210110	816.0000
15	02793455	01	210110	419.0000
16	02793525	01	210110	146.0000
17	02793645	01	210110	475.0000
18	02793775	01	210110	800.0000
19	02793805	01	210110	860.0000
20	02793815	01	210110	800.0000
21	02794175	01	210110	196.0000



Required Resources / Skills

```
135 ~ #####
136 #
137 #           Compute Annual Mean Estimates
138 #
139 ~ #####
140
141 # Merge Interview CU weights and expenditures
142 int_df <- left_join(
143   fmli,
144   expend %>% filter(ucc %in% getUCCs(expenditure, stub)) %>%
145     group_by(newid) %>% summarise(cost = sum(cost)),
146   by = "newid"
147 ) %>%
148   mutate_each_(
149     funs(replace(., is.na(.), 0)),
150     vars = c("cost", paste0("wtrep", str_pad(1:44, 2, "left", 0)))
151   )
152
153 # Compute an Interview annual mean estimate
154 int_gm <- int_df %>%
155   mutate(wt_cost = cost * finlwt21) %>%
156   summarise(
157     grand_mean = sum(wt_cost, na.rm = TRUE) / sum(calwt)
158   ) %>% unlist() %>% unname()
159
160 # Merge Diary CU weights and expenditures
161 dia_df <- left_join(
162   fmid,
163   expend %>% filter(ucc %in% getUCCs(expenditure, stub)) %>%
164     group_by(newid) %>% summarise(cost = sum(cost)),
165   by = "newid"
166 ) %>%
167   mutate_each_(
168     funs(replace(., is.na(.), 0)),
169     vars = c("cost", paste0("wtrep", str_pad(1:44, 2, "left", 0)))
170   )
171
172 # Compute a Diary annual mean estimate
173 dia_gm <- dia_df %>%
174   mutate(wt_cost = cost * finlwt21) %>%
175   summarise(grand_mean = sum(wt_cost, na.rm = TRUE) / sum(popwt)) %>%
176   unlist() %>% unname()
177
```



Interactive CE Visualization Tool



Introduction

CE Visualization

Methods

Introduction: Comparisons of reported expenditures

The **Consumer Expenditure Survey (CE)** program consists of two surveys, the Interview and Diary surveys, which collect data on household expenditures, income, and consumer unit (families and single consumers) characteristics. The survey covers the U.S. consumer units (CUs), which we also refer to as households or families.

This application is intended to provide the user an introduction to CE data through an interactive visualization tool and an accompanying table showing comparisons of expenditures between a selected subsample of the (

Click on the **CE Visualization** tab above to use the application



Interactive CE Visualization Tool

Interactive CE Visualization Tool - 2015 Data

1

Demographic Categories

- Region
- Number of people in CU
- Home Owner / Renter
- CU income range
- Highest level of education in the CU
- Race of the reference person

2

Subcategories

Region
Midwest

CU income range
Lowest 20 Percent

Race of the reference person
Asian

submit

3

Options

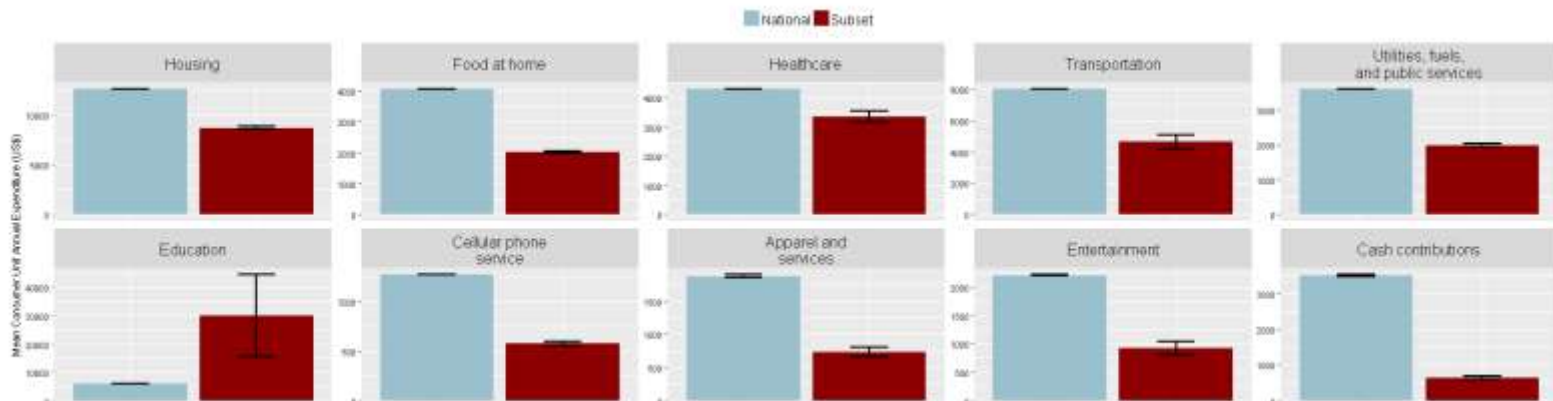
Independent scales

Download Table

4

Number of households in your sample

28



Interactive CE Visualization Tool

Demographic Categories

1

- Region
- Number of people in CU
- Home Owner / Renter
- CU income range
- Highest level of education in the CU
- Race of the reference person

Interactive CE Visualization Tool

Subcategories 2

Region

Midwest ▼

CU income rage

Lowest 20 Percent ▼

Race of the reference person

Asian ▲

White

Black or African American

American Indian or Alaskan Native

Asian

Native Hawaiian or Other Pacific Islander


Multi-race



Interactive CE Visualization Tool

Options **3**

Independent scales

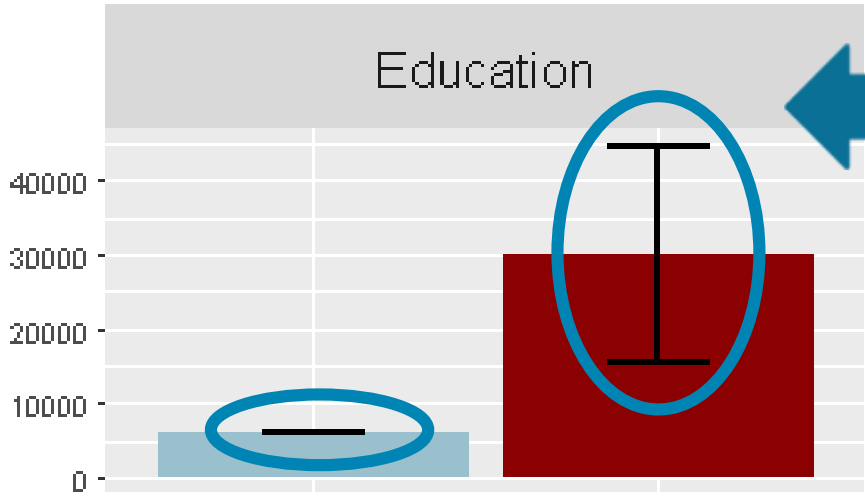
 Download Table

Number of households in your sample: **4**

28



Interactive CE Visualization Tool



Error Bars

Mean = \$30,040.00

CV = 24.12%

Sample Size = 3

Lower Bound = \$15,548.70

Upper Bound = \$44,531.30

Sample	Expenditure	Cost	CV (%)	Lower Bound	Upper Bound	Sample Size
Subset	Education	30040.00	24.12	15548.70	44531.30	3
National	Education	5946.81	0.64	5870.69	6022.93	4024



Benefits to the user

- Accessibility: The user can access the app for **free** as long as they have internet access on a device with a web browser
- Usability: The user operates only the clean, user-friendly UI to get data, results, and visualizations



Interactive CE Visualization Tool

- Arcenis Rojas: rojas.arcenis@bls.gov

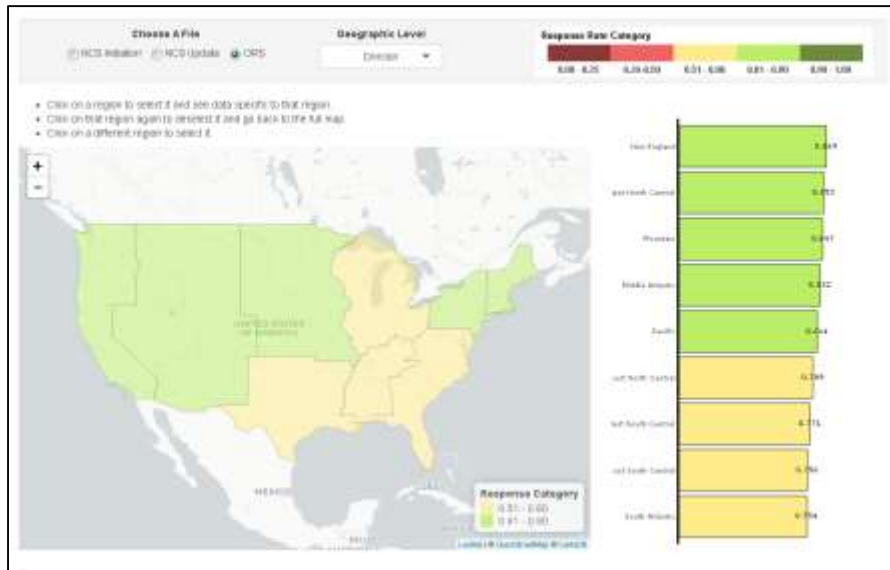


Real-time Response Rate Tool

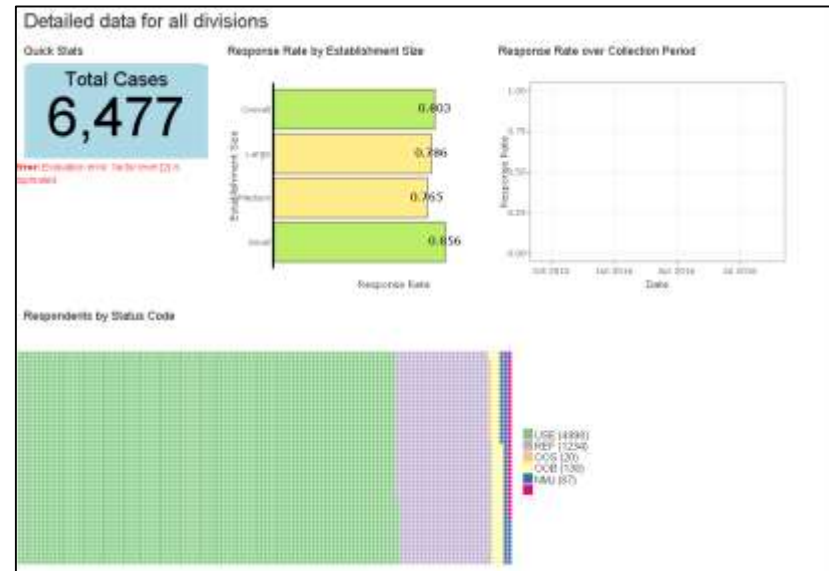
- Office of Compensation and Working Conditions
- Provide real-time response rates to field offices
 - ▶ Focus on problem collection areas
 - ▶ Improved sample representativity



Real-time Response Rate Tool



Detailed summaries for each region



Response rates by region and/or establishment size



Real-time Response Rate Tool

- Brandon Kopp (OSMR):
kopp.brandon@bls.gov
- Randall Powers (OSMR):
powers.randall@bls.gov
- Arcenis Rojas (CE):
rojas.arcenis@bls.gov



Other Shiny Applications

- Choropleth maps of unemployment data (OSMR)
- Energy Information Administration analyzer (PPI)
- Text analysis Shiny App (Survey Methods)



R Packages



R Packages

- **rpms**: Recursive Partitioning for Modeling Survey Data package (Survey Methods)
- **growfunctions**: Bayesian Non-Parametric Dependent Models for Time-Indexed Functional Data package (Survey Methods)



rpms

- Fits a linear model to survey data in each node obtained by recursively partitioning the data.
- Adjusts for complex sample design features used to obtain the data.
- Produces design-consistent coefficients to the least squares linear model between the dependent and independent variables.



rpms

- The main function returns the resulting binary tree with the linear model fit at every end-node.
- Daniell Toth (OSMR): toth.daniell@bls.gov

growfunctions

- Bayesian Non-Parametric Dependent Models for Time-Indexed Functional Data package (Survey Methods)
- Estimates a collection of time-indexed functions under either of Gaussian process (GP) or intrinsic Gaussian Markov random field (iGMRF) prior formulations

growfunctions

- Dirichlet process mixture allows sub-groupings of the functions to share the same covariance or precision parameters
- The GP and iGMRF formulations both support any number of additive covariance or precision terms, respectively, expressing either or both of multiple trend and seasonality.

growfunctions

- Terrance Savitsky (OSMR):
savitsky.terrance@bls.gov



Challenges



Challenges

- Data confidentiality
- Need for an R server to make apps/programs public
- Can only put Shiny apps on a webpage via iFrames or setting up an account on a cloud server (i.e., Digital Ocean, R Studio)



Contact Information

Arcenis Rojas

Economist

Division of Consumer Expenditure Surveys

www.bls.gov/cex

202-691-6884

rojas.arcenis@bls.gov

