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
Enter a Collection Unit and Sample

Collection Unit (last 4 digits):

Sample:

X43

Company Name (edited):

Corp Div:

countertop(0)

Street:

logit(0)

Search in New Tab

Open Google Maps

Open SOS Website

Create Future Note

Left Side
Right Side

ACE: matching on sampled EIN

RTS Master Listing: matching on Name, Corp Div, or Street displayed on the left
Search Results
Export Addresses at a Glance

<table>
<thead>
<tr>
<th>pct</th>
<th>export_id</th>
<th>eistrata</th>
<th>name</th>
<th>address</th>
<th>city</th>
<th>st</th>
<th>zip</th>
<th>eistrecs</th>
<th>addresses</th>
<th>split</th>
<th>flag1</th>
<th>flag2</th>
<th>problem</th>
<th>Mike.s.Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>33%</td>
<td>2398</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21%</td>
<td>2398</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11%</td>
<td>2398</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7%</td>
<td>2398</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>2398</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4%</td>
<td>2398</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3%</td>
<td>2398</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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: khatcchadourian.ara@bls.gov
- Rob Sutton: sutton.robert@bls.gov
Industrial Prices Visualization Dashboard
Index Comparisons
Index Review and Revision

95% Confidence Intervals of 221122 | Electric power distribution

 percent change


Official Percent Change

EMBARGOED DATA - NOT FOR PUBLIC RELEASE
Visualization Dashboard

- Neil Wagner: wagner.neil@bls.gov
- Steve York: york.stephen@bls.gov
Interactive CE Visualization Tool
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)

<table>
<thead>
<tr>
<th>nwmid</th>
<th>nwmoid</th>
<th>nwmstat</th>
<th>fam_size</th>
<th>faminc</th>
<th>ref_race</th>
<th>region</th>
<th>flncc</th>
<th>high_edu</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>02793005</td>
<td>Owner</td>
<td>1</td>
<td>1,000,000</td>
<td>Black or African American</td>
<td>South</td>
<td>Second 20 Percent</td>
<td>High School Graduate</td>
</tr>
<tr>
<td>2</td>
<td>02793005</td>
<td>Renter</td>
<td>1</td>
<td>1,459.350</td>
<td>White</td>
<td>West</td>
<td>Third 20 Percent</td>
<td>Some College or Less</td>
</tr>
<tr>
<td>3</td>
<td>02793075</td>
<td>Renter</td>
<td>3</td>
<td>1,046.814</td>
<td>White</td>
<td>South</td>
<td>Second 20 Percent</td>
<td>Some College or Less</td>
</tr>
<tr>
<td>4</td>
<td>02793025</td>
<td>Owner</td>
<td>1</td>
<td>2,131.304</td>
<td>White</td>
<td>South</td>
<td>Third 20 Percent</td>
<td>Less than High School</td>
</tr>
<tr>
<td>5</td>
<td>02792115</td>
<td>Renter</td>
<td>4</td>
<td>1,048.001</td>
<td>Asian</td>
<td>Northeast</td>
<td>Third 20 Percent</td>
<td>Bachelor’s Degree</td>
</tr>
<tr>
<td>6</td>
<td>02792125</td>
<td>Renter</td>
<td>3</td>
<td>1,490.459</td>
<td>White</td>
<td>Northeast</td>
<td>Lowest 20 Percent</td>
<td>Some College or Less</td>
</tr>
<tr>
<td>7</td>
<td>02792135</td>
<td>Owner</td>
<td>1</td>
<td>6,143.497</td>
<td>Asian</td>
<td>West</td>
<td>Fourth 20 Percent</td>
<td>Postgraduate Degree</td>
</tr>
<tr>
<td>8</td>
<td>02792225</td>
<td>Owner</td>
<td>1</td>
<td>1,376.790</td>
<td>White</td>
<td>South</td>
<td>Second 20 Percent</td>
<td>Some College or Less</td>
</tr>
<tr>
<td>9</td>
<td>02792245</td>
<td>Renter</td>
<td>4</td>
<td>1,936.831</td>
<td>White</td>
<td>West</td>
<td>Third 20 Percent</td>
<td>Some College or Less</td>
</tr>
<tr>
<td>10</td>
<td>02792265</td>
<td>Renter</td>
<td>2</td>
<td>1,595.534</td>
<td>Asian</td>
<td>West</td>
<td>Second 20 Percent</td>
<td>High School Graduate</td>
</tr>
<tr>
<td>11</td>
<td>02792275</td>
<td>Renter</td>
<td>2</td>
<td>1,079.581</td>
<td>Black or African American</td>
<td>South</td>
<td>Third 20 Percent</td>
<td>Bachelor’s Degree</td>
</tr>
<tr>
<td>12</td>
<td>02792285</td>
<td>Renter</td>
<td>2</td>
<td>1,775.726</td>
<td>White</td>
<td>South</td>
<td>Fourth 20 Percent</td>
<td>Bachelor’s Degree</td>
</tr>
<tr>
<td>13</td>
<td>02792285</td>
<td>Owner</td>
<td>1</td>
<td>1,828.435</td>
<td>White</td>
<td>South</td>
<td>Second 20 Percent</td>
<td>Less than High School</td>
</tr>
<tr>
<td>14</td>
<td>02792285</td>
<td>Renter</td>
<td>4</td>
<td>1,705.913</td>
<td>Black or African American</td>
<td>South</td>
<td>Fourth 20 Percent</td>
<td>Some College or Less</td>
</tr>
<tr>
<td>15</td>
<td>02792345</td>
<td>Owner</td>
<td>2</td>
<td>1,087.147</td>
<td>White</td>
<td>Northeast</td>
<td>Third 20 Percent</td>
<td>Some College or Less</td>
</tr>
<tr>
<td>16</td>
<td>02792345</td>
<td>Owner</td>
<td>1</td>
<td>717.759</td>
<td>Asian</td>
<td>West</td>
<td>Third 20 Percent</td>
<td>Bachelor’s Degree</td>
</tr>
<tr>
<td>17</td>
<td>02792445</td>
<td>Renter</td>
<td>1</td>
<td>1,004.521</td>
<td>White</td>
<td>Northeast</td>
<td>Lowest 20 Percent</td>
<td>High School Graduate</td>
</tr>
<tr>
<td>18</td>
<td>02792555</td>
<td>Renter</td>
<td>2</td>
<td>1,887.781</td>
<td>Asian</td>
<td>West</td>
<td>Highest 20 Percent</td>
<td>Postgraduate Degree</td>
</tr>
<tr>
<td>19</td>
<td>02792595</td>
<td>Renter</td>
<td>2</td>
<td>2,300.029</td>
<td>Black or African American</td>
<td>Midwest</td>
<td>Third 20 Percent</td>
<td>Some College or Less</td>
</tr>
<tr>
<td>20</td>
<td>02792595</td>
<td>Owner</td>
<td>3</td>
<td>1,909.100</td>
<td>Asian</td>
<td>West</td>
<td>Highest 20 Percent</td>
<td>Bachelor’s Degree</td>
</tr>
</tbody>
</table>

#### Expenditures File (1,720,755 Observations)

<table>
<thead>
<tr>
<th>rowid</th>
<th>ref_moid</th>
<th>mce</th>
<th>cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>02793005</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>2</td>
<td>02793025</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>3</td>
<td>02793015</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>4</td>
<td>02792125</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>5</td>
<td>02793225</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>6</td>
<td>02794225</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>7</td>
<td>02792285</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>8</td>
<td>02792335</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>9</td>
<td>02792345</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>10</td>
<td>02792355</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>11</td>
<td>02792385</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>12</td>
<td>02793035</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>13</td>
<td>02793345</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>14</td>
<td>02793355</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>15</td>
<td>02793435</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>16</td>
<td>02793525</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>17</td>
<td>02793545</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>18</td>
<td>02793575</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>19</td>
<td>02793405</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>20</td>
<td>02793415</td>
<td>0</td>
<td>210110</td>
</tr>
<tr>
<td>21</td>
<td>02794175</td>
<td>0</td>
<td>210110</td>
</tr>
</tbody>
</table>
Required Resources / Skills

---

```r
# Compute Annual Mean Estimates
int_df <- left_join(
  fml, expend %>% filter(ucc %in% getUCCs(expenditure, stub)) %>%
  group_by(newid) %>% summarise(cost = sum(cost)),
  by = "newid"
) %>%
  mutate_each_(
    funs(replace(., is.na(., 0)),
      vars = c("cost", psst0("wtrp", str_pad(1:44, 2, "left", 0)))
  )

# Compute an Interview annual mean estimate
int_gm <- int_df %>%
  mutate(wt_cost = cost * finlwt) %>%
  summarise(grand_mean = sum(wt_cost, na.rm = TRUE) / sum(calwt)) %>%
  unlist() %>%

# Merge Diary CU weights and expenditures
dia_df <- left_join(
  fml, expend %>% filter(ucc %in% getUCCs(expenditure, stub)) %>%
  group_by(newid) %>% summarise(cost = sum(cost)),
  by = "newid"
) %>%
  mutate_each_(
    funs(replace(., is.na(., 0)),
      vars = c("cost", psst0("wtrp", str_pad(1:44, 2, "left", 0)))
  )

# Compute a Diary annual mean estimate
dia_gm <- dia_df %>%
  mutate(wt_cost = cost * finlwt) %>%
  summarise(grand_mean = sum(wt_cost, na.rm = TRUE) / sum(popwt)) %>%
  unlist() %>%
```

---
Interactive CE Visualization Tool

Consumer Expenditure Survey

Introduction: Comparisons of reported expenditures

The Consumer Expenditure Survey (CE) program consists of two surveys, the Interview and Diary expenditure, income, and consumer unit (families and single consumers) characteristics. The survey covers 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 approach and 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 rage
     - Lowest 20 Percent
   - Race of the reference person
     - Asian

3. Options
   - Independent scales
   - Download Table

4. Number of households in your sample
   - 28
Interactive CE Visualization Tool

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
Interactive CE Visualization Tool

**Subcategories**

**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

- Independent scales

Download Table

Number of households in your sample: 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

<table>
<thead>
<tr>
<th>Sample</th>
<th>Expenditure</th>
<th>Cost</th>
<th>CV (%)</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subset</td>
<td>Education</td>
<td>30040.00</td>
<td>24.12%</td>
<td>15548.70</td>
<td>44531.30</td>
<td>3</td>
</tr>
<tr>
<td>National</td>
<td>Education</td>
<td>5946.81</td>
<td>0.64%</td>
<td>5870.69</td>
<td>6022.93</td>
<td>4024</td>
</tr>
</tbody>
</table>
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

Response rates by region and/or establishment size

Detailed summaries for each region
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.
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