Computational Statistics (2022) 37:2087–2094
https://doi.org/10.1007/s00180-022-01257-w

EDITORIAL

The 2017 *Data Challenge* of the American Statistical Association

Thesia I. Garner[1] · **Wendy Martinez**[2]

Received: 15 June 2022 / Accepted: 15 June 2022 / Published online: 15 July 2022

1 Introduction

There is now an annual *Data Challenge* sponsored by three sections of the American Statistical Association (ASA - https://www.amstat.org/): *Statistical Computing, Statistical Graphics*, and *Government Statistics*. The *Data Challenge* builds upon a long history of ASA *Data Expos* and *Data Challenges* with the first one starting in 1982/1983 (https://community.amstat.org/jointscsg-section/dataexpo). The *Data Challenge* is open to students, professionals, and any person interested in exploring the challenge data set. Many of the challenges utilize publicly available data sets produced by the U.S. federal government.

The data set used in the 2017 *Data Challenge* was provided by the *U.S. Bureau of Labor Statistics (BLS)*. The BLS mission is to "measure labor market activity, working conditions, price changes, and productivity in the U.S. economy to support public and private decision making" (https://www.bls.gov/bls/blsmissn.htm). One of the major principal federal economic indicators produced by the BLS is the Consumer Price Index (CPI). A major source of data used to calculate the CPI is the Consumer Expenditure Survey (CE). In addition, the CE is the only Federal household survey to provide information on the complete range of consumers' expenditures and incomes (https://stats.bls.gov/cex/). The data set used in the 2017 *Data Challenge* was com-

✉ Wendy Martinez
martinezw@verizon.net

Thesia I. Garner
garner.thesia@bls.gov

[1] Chief, Division of Price and Index Number Research Office of Prices and Living Conditions, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E, 20212 Washington, D.C, United States

[2] Director, Mathematical Statistics Research Center, Office of Survey Methods Research U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E, 20212 Washington, D.C, United

posed of public use data files from the CE survey. These data are described in the next section.

There are two award categories in the *Data Challenge* – one for professionals and one for students. Participants are challenged to analyze a data set using data science, statistical and visualization tools, and methods. There was no specific problem statement for the *2017 Data Challenge*, which means that contestants were free to be creative in their analyses.

Contestants are judged based on the analyses and work presented at the Joint Statistical Meetings (JSM) 2017, which was held in Baltimore, Maryland from July 30 through August 3, 2017. We had nine entries in the student category and three in the professional category. The lists of 2017 winners are given here, where authors are listed as they appear in the JSM 2017 program at https://ww2.amstat.org/meetings/jsm/2017/program.cfm.

1.1 Student Level Award

- **1st Place Winner**: Nathan James and Jacquelyn Neal, Vanderbilt University, "Interactive Visualization of Consumer Expenditure Public-Use Microdata."
- **2nd Place Winner**: Joyance Meechai, Pennsylvania State University, "Energy Expenditure Patterns in the United States."
- **3rd Place Winner (tied)**: Mingzhao Hu, University of California Santa Barbara, "Income and Expenditure in US Households: A Multivariate Analysis of Consumer Expenditure CE Dataset."
- **3rd Place Winner (tied)**: Robert Garrett, Thomas Fisher, and Ritu Narahari, Miami University, "An Analysis of Consumer Budgeting and the Great Recession."

1.2 Professional Level Award

- Gaurav Sharma, Bhanu Shandilya, and Aditi Pradeep Sharma, The EMMES Corporation, Mixpanel, and UMBC, "Consumer Spending and Federal Reserve."

Abstracts for all entries presented at the JSM can be found at https://ww2.amstat.org/meetings/JSM/2017/OnlineProgram/ActivityDetails.cfm?sessionid=214577 and https://ww2.amstat.org/meetings/JSM/2017/OnlineProgram/ActivityDetails.cfm?SessionID=214566.

The Guest Editors of this special issue are grateful for the support of Springer and the editors of *Computational Statistics* for giving us the opportunity to publish refereed articles from Data Challenge contestants. This special issue for the 2017 *Data Challenge* is the fifth in a series of special issues in *Computational Statistics* focused on the *ASA Data Expos and Data Challenges*. The first special issue presented papers from the 2006 *Data Expo* with the data set containing geographic and atmospheric measures on a coarse regular grid covering Central America (Murrell 2010). Next

was the 2011 *Data Expo* where the challenge data set was from the *Deepwater Horizon* oil spill. The data came from the monitoring water temperature and salinity, water chemistry, and relevant wildlife counts (Cook 2014). This was followed by the 2013 *Data Expo* with data from the *Knight Foundation* (https://knightfoundation.org/) describing the emotional attachment of residents to their communities (Hofmann et al. 2019). The special issue immediately prior to this current one showcased papers from the 2016 *Data Challenge* with the Department of Transportation's General Estimates System (GES) serving as the challenge data set (Amjadi and Martinez 2021).

2 The challenge

The Consumer Expenditure Survey is a nationwide household survey conducted by the U.S. Bureau of Labor Statistics to collect data on the spending behavior of consumers living in the United States. It is the only federal government survey that provides information on expenditures for all goods and services as well as sources and levels of income, demographic and economic characteristics, and assets and liabilities. CE data are used by the BLS, other government agencies, and the private sector to provide insights regarding the economic well-being of consumers. Policymakers use the data to assess the impact of economic policy on consumer spending behavior. BLS uses the data to create the weights for the Consumer Price Index (https://stats.bls.gov/cpi/) and to produce the Supplemental Poverty Thresholds (https://stats.bls.gov/pir/spmhome.htm). See the BLS website https://www.bls.gov/cex/ and the BLS (2018) *Handbook of Methods* for detailed information on the CE survey and data.

The Consumer Expenditure Survey data series is one of the oldest among those produced by the U.S. Bureau of Labor Statistics. The first survey was initiated in 1888 and continued through 1891; the purpose of the survey was to study workers' spending patterns as elements of production costs. This was followed by nine additional iterations of the survey. A significant design change from earlier surveys was introduced in 1972–1973 that included both an *Interview* and a *Diary* for data collection, with each instrument having its own sample. Previous survey data collection had been periodic at approximate 10-year intervals. The ninth and most recent iteration in 1980 included the introduction of the continuing survey which is still used today. In addition, the *Diary-Interview* instrument and separate samples design, introduced in 1972-73, has been followed since 1980 and continues through today (Reed, 2014; Henderson and Safir 2018; Jacobs and Shipp, 1990).

As noted, the CE has two components, an *Interview* and a *Diary*. The *Interview* Survey is designed to collect data on large and recurring expenditures that consumers can be expected to recall for a period of three months or longer, such as rent and utilities. The *Diary* Survey is designed to collect data on small, frequently purchased items, including most food and clothing. For the *Interview* component of the CE, sampled consumers are interviewed every three months over a maximum of four calendar quarters. At the end of the fourth interview, the sample unit is replaced by a new sample unit. Approximately 25% of the *Interview* sample each quarter are new to the survey, replacing consumers who have finished their participation. The separate *Diary* Survey is completed by a separate sample of consumers for two consecu-

tive one-week periods, and then these consumers rotate out of the sample. The *Diary* is designed such that consumers self-record detailed descriptions of all spending, with this recordkeeping starting any day of the week. Together, data from the *Interview* and *Diary* cover all consumers' expenditures. However, data collected in the *Interview* can be used alone to represent approximately 95% of expenditures using detailed data on an estimated 60 to 70% of household expenditures, in combination with global estimates obtained for food and other selected items such as alcoholic beverages and tobacco products. These global estimates account for an additional 20 to 25% of total expenditures (BLS, 2018, *Handbook of Methods*).

CE data are collected from a set of individuals referred to as a Consumer Unit (CU). A CU is defined as one of the following: (1) all members of a housing unit who are related by blood, marriage, adoption, or other legal arrangement, such as foster children; (2) a person living alone or sharing a household with others, or living as a roomer (someone who rents a room in a home) in a private home, lodging house, or in permanent living quarters in a hotel or motel, but who is financially independent; or (3) two or more unrelated persons living together who use their income and other resources to make joint expenditure decisions. Students living in university-sponsored housing are also included in the sample as separate CUs. Information on members (e.g., age, gender, education, individually earned income) in the CU is identified by their relationship to the reference person. The reference person is the first member mentioned by the respondent when asked to "Start with the name of the person or one of the persons who owns or rents the home" (BLS, 2018, *Handbook of Methods*).

The BLS releases CE data as published tabulations and at the CU level. Annual estimates of CU expenditures by select sociodemographic variables are published twice per year. Public use micro data are released once per year; separate files are available for the *Interview* and the *Diary* from https://stats.bls.gov/cex/.

3 Summary of Papers in this special issue

Winners of the 2017 *Data Challenge* were invited to contribute an article to this special issue. All other contestants were invited to submit a paper based on the quality of papers submitted to the JSM proceedings. Only two papers were accepted for publication, both of which are by student winners of the challenge. We briefly introduce these papers here.

Second place winner in the student category Joyance Meechai (Meechai 2017; Meechai & Wijesinha, 2022) expanded the work presented at JSM by examining patterns in energy expenditures and consumption in the U.S. (from the Energy Information Administration). Joyance was seeking a relationship between energy usage in a household and socio-demographic characteristics. The quantitative socio-demographic variables included family size, number of cars, number of rooms in the housing unit, highest education level, and income. Categorical socio-demographic variables included urban/rural classification, geographic region, building type, occupancy type, and family situation. The findings indicated patterns of higher energy demands with rural residents and lower energy demands in urban blue-collar CUs

(consumer units). Not surprisingly, the number of people in a CU contributed to its direct energy consumption, along with geography and income.

Third place student winner Mingzhao Hu (Hu 2017; Hu, 2022) used multivariate analysis methods to analyze income and expenditure patterns for consumer units living in the U.S. Mingzhao used thirty-five variables from the CE public-use data that encompassed categories corresponding to demographics, income, and expenditures. Exploratory data analysis methods, such as principal component analysis and canonical correlation analysis were used to understand the relationships between the three categories. Interesting findings indicate that income (after-tax and wage/salary), quarterly total expenditures, food expenditures, and housing expenditures are the most important variables.

4 Other submissions to the *2017 Data Challenge*

As mentioned previously, not every participant in the 2017 *Data Challenge* submitted a paper to this special issue. However, all contestants had the opportunity to submit a paper to the JSM conference proceedings. This is a benefit of presenting at the *Joint Statistical Meetings* (https://www.amstat.org/ASA/Meetings/Joint-Statistical-Meetings.aspx). We briefly describe some of the analyses that were published in the JSM proceedings.

Fig. 1 Contestant Mike Jadoo answering questions regarding his analysis

Garrett et al. (2017) used the data from the CE survey to study how the Great Recession of 2007–2009 affected the economic behavior of individuals and consumer units in the U.S. They examined the change in buying habits of consumers before and after the recession and compared these to macro-economic trends. They also created a model using the CE data to predict how the recession might continue to affect consumers in the future.

Michael Jadoo (Jadoo 2017) explored multi-year trends in the CE survey data to better understand the buying habits of consumers. Using data from 2010 to 2015, along with tables and graphs, he found that an economic phenomenon happened in 2013, which was caused by fiscal policy, impacted consumer spending. The data and graphics indicated that consumers opted to buy less expensive goods during a time when taxes and health care premiums were higher. See Fig. 1 for a picture of Mike presenting his work at the Joint Statistical Meetings in Baltimore, MD.

Using data from the CE survey from 1996 to 2015, Zhao and Li (2017) explored ways to model the relationship between household age and consumer expenditures for dairy and beef; these commodities exhibited a humped shape or quadratic trend over the time period. Single-year cohorts were formed based on household birth year ranging from 1931 to 1980. Three models were fitted and compared: a simple regression assuming all parameters are constant across cohorts, a cohort-based regression model where the correlation between cohorts is ignored, and a varying coefficient

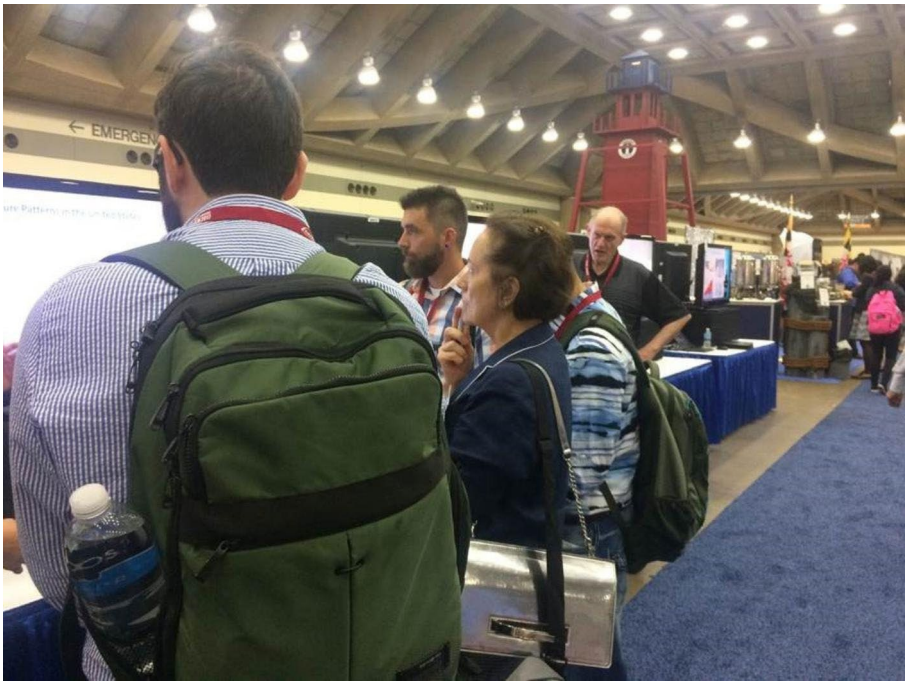

Fig. 2 Judges learning more about one of the entries. Going from left to right, we have Jonathan Auerbach, Thesia Garner, and Jeffrey Gonzalez

model where model coefficients are smoothed functions of the cohorts. The simple regression model provided stable estimates but does not allow for parameters to change over time. The cohort-based model allows for parameters to vary over time, but it was not robust and was highly affected by extreme values. The varying coefficient model allows for changes in parameters through time and accounts for possible dependence between cohorts. The authors found it provided a smoother and more robust approach, compared to the other two models.

5 Supplementary Materials

The special issue for the 2013 *Data Expo* (Hofmann et al. 2019) set the precedent for publishing a set of papers following the principles of reproducibility, and this was continued for the *2016 GSS Data Challenge* (Amjadi and Martinez 2021) special issue. We followed this important concept for this special issue, and we asked authors to upload supplementary materials to a Github site: https://github.com/asa-stat-computing-and-graphics/COST-DataExpo-2017. These were any files for the paper (Tex, bib, figures), data sets, and computer code (project files, macros, SAS PROCS, R files, Shiny apps, etc.) the authors created as part of their analysis and the published article. All code and materials were reviewed as part of the referee process.

Acknowledgements The Guest Editors and authors would like to thank the Editors, the referees, and the journal management staff of *Computational Statistics* and Springer for their help and patience as we worked through the process of preparing this special issue. A special thank you goes to Jürgen Symanzik (Past Editor-in-Chief) as he obtained the approval for this issue and helped guide us through the process. We also thank Lucy D'Agostino and Samantha Tyner for their help setting up the Github site. Danny Yang from the BLS and a CE program expert reviewed the code and supplemental materials for the papers published in this special issue, and we thank him for his efforts. We are grateful for the support of the sponsoring sections of the ASA and our judges Jonathan Auerbach, MoonJung Cho, Thesia I. Garner, Harold Gomes, Jeffrey Gonzalez, and Jürgen Symanzik. See Fig. 2 for some of the judges examining one of the entries. Finally, a special thank you to Spyros Bakas and Tatiana Plotnikova from Springer who supported color printing at no charge and enabled this fully reproducible special issue.

References

Amjadi R, Martinez W (2021) Comput Stat 36:1553–1590. https://doi.org/10.1007/s00180-021-01076-5. The 2016 Data Challenge of the American Statistical Association

Bureau of Labor Statistics (BLS) (2018) "Consumer Expenditures and Income." *Handbook of Methods*, Last Modified Date: March 28, 2018. https://www.bls.gov/opub/hom/cex/home.htm

Cook D (2014) Comput Stat 29:117–119. https://doi.org/10.1007/s00180-013-0474-x. The 2011 Data Expo of the American Statistical Association

Garrett R, Narahari R, Fisher TJ(2017) An analysis of consumer budgeting and the Great Recession, 2017 JSM Proceedings, American Statistical Association, Alexandria, VA, pp 2207–2211

Henderson S, Safir A(2018) "130 Years of the Consumer Expenditure Surveys (CE) 1888–2018." Presentation June 7, 2018. https://www.bls.gov/cex/ce-130-presentation-safir-henderson.pdf

Hofmann H, Wickham H, Cook D (2019) The 2013 Data Expo of the American Statistical Association. Comput Stat 34:1443–1447. https://doi.org/10.1007/s00180-019-00923-w

Hu M (2017) Income and expenditure in US households: A multivariate analysis of Consumer Expenditure fmli161 dataset, 2017 JSM Proceedings, American Statistical Association, Alexandria, VA, pp 1086–1145

Hu M (2022) *Multivariate understanding of income and expenditure in United States households with statistical learning*, https://doi.org/10.1007/s00180-022-01251-2
Jacobs E, Shipp S, (March(1990) How family spending has changed in the U.S.,Monthly Labor Review, pp20–27. https://www.bls.gov/opub/mlr/1990/03/art3full.pdf
Jadoo M (2017) Tracking expenditures, 2017 JSM Proceedings, American Statistical Association, Alexandria, VA, pp 3633–3638
Meechai J (2017) Household expenditure and consumption patterns in the United States, 2017 JSM Proceedings, American Statistical Association, Alexandria, VA, pp 3520–3534
Meechai J, Wijesinha M (2022) Household Energy Expenditure and Consumption Patterns in the United States, https://doi.org/10.1007/s00180-022-01255-y
Murrell P (2010) Comput Stat 25:551–554. https://doi.org/10.1007/s00180-010-0207-3. The 2006 Data Expo of the American Statistical Association
Reed SB (April 2014) One hundred years of price change: The Consumer Price Index and the American inflation experience. Mon Labor Rev. https://doi.org/10.21916/mlr.2014.14
Zhao Z, Li F(2017) CE-based Consumer Expenditure behaviors study, 2017 JSM Proceedings, American Statistical Association, Alexandria, VA, pp 3405–3415