Computational Statistics (2023) 38:1621–1627
https://doi.org/10.1007/s00180-023-01398-6

EDITORIAL

The 2019 *data challenge expo* of the American Statistical Association

Daniel Goldstein[1] · Elyzabeth Gaumer[1] · Wendy Martinez[2]

Received: 15 July 2023 / Accepted: 15 July 2023 / Published online: 4 September 2023

1 Introduction

The authors are pleased to introduce the special issue of the *Data Challenge Expo* 2019. The *Data Challenge Expo* (https://community.amstat.org/dataexpo/home) is an annual event sponsored by three sections of the American Statistical Association (https://www.amstat.org/): *Statistical Computing, Statistical Graphics*, and *Government Statistics*. ASA data challenges in various forms have been held for several decades. See https://community.amstat.org/jointscsg-section/dataexpo for more information on past challenges, including the data sets that were used with many of them available for download.

The current special issue is the seventh one published in *Computational Statistics*. The first issue included papers from the 2006 *Data Expo* where the challenge data set encompassed geographic and atmospheric measures recorded in Central America (Murrell 2010). The next special issue focused on the 2011 *Data Expo*, where contestants had the chance to explore data from the *Deepwater Horizon* oil spill. Variables in the challenge data set measured water temperature, salinity, water chemistry, and relevant wildlife counts (Cook 2014). The 2013 *Data Expo* focused on social issues and what characteristics attract people to geographic communities (Hofmann, et al. 2019). The following challenges came about three years later with some using government data sets. The 2016 challenge used data from the Department of Transportation's General Estimates System (Amjadi and Martinez 2021); the 2017 challenge used data from the

✉ Wendy Martinez
wendy.l.martinez@census.gov

Daniel Goldstein
goldsted@hpd.nyc.gov

Elyzabeth Gaumer
gaumere@hpd.nyc.gov

[1] Center for Research on Housing Opportunity, Mobility, and Equity (HOME)NYC Department of Housing Preservation & Development, 100 Gold Street, New York, NY 10038, USA

[2] U.S. Census Bureau, Washington, D.C 20233, USA

Bureau of Labor Statistics' Consumer Expenditure Survey (Garner and Martinez 2022); and the 2018 data challenge used observations associated with weather forecasting (Cetinkaya-Rundel and Martinez 2023).

The focus of the recent data challenges has expanded beyond graphics and visualization, which were the objectives of the early data challenges, and asks participants to include statistical computing, modeling, and data science methodologies in their submissions. The advent of open-source software tools has enabled contestants to be innovative, and many participants provide applications written in R as will be discussed later.

The *Data Challenge Expo* is open to students, professionals, and any person interested in exploring the challenge data set. The data set used in the 2019 *Data Challenge Expo* came from the New York City Housing and Vacancy Survey. The story of how the data were collected and the motivation behind the challenge are discussed in the next section. The official *Data Challenge Expo 2019* announcement can be found here https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2019.

Contestants can compete in one of two categories: student or professional. Participants are asked to analyze a data set using tools and techniques from data science and statistics. Typically, a specific problem statement or goal for a data challenge is not given because we want contestants to be creative and innovative in how they tackle the data. The 2019 *Data Challenge Expo* was different in that the agency providing the data also provided research questions, as discussed in Section 2. However, participants were still free to explore and analyze the data in a way that made sense to them and to formulate their own study questions.

Contestants were judged based on the analyses and work presented at the Joint Statistical Meetings (JSM) 2019 that took place in Denver, Colorado from July 27 through August 1, 2019. There were twelve entries in the student category and two in the professional category. The winners are listed below as they appear in the JSM 2019 program available at https://ww2.amstat.org/meetings/jsm/2019/onlineprogram/index.cfm, where presentations for most of the entries are available for download; search for session 28 in the online program.

1.1 Student Level Award

- **1st Place Winner**: Xiang Shen; Shunyan Luo; Mingze Zhang, George Washington University, "An Analysis of Immigrants and House Condition in New York City"
- **2nd Place Winner**: Benjamin Schweitzer; Thomas J Fisher; Karsten Maurer, Miami University., " An Analysis of Rent-Control Policy on Housing Quality"
- **3rd Place Winner (tie)**: Alison Tuiyott; Thomas J Fisher; Karsten Maurer, Miami University, " Immigrant Residency and Happiness in New York City"
- **3rd Place Winner (tie):** Jacob Gertszten; Damian Chambon, University of Virginia, "An Analysis of Housing Quality in New York City"

1.2 Professional Level Award

Robert Montgomery; Quentin Brummet; Nola du Toit; Peter Herman; Edward Mulrow, NORC at the University of Chicago, "Measuring Gentrification Over Time with the NYCHVS"

Abstracts for all entries in the *Data Challenge Expo* can be found on the JSM online program here https://ww2.amstat.org/meetings/jsm/2019/onlineprogram/ActivityDetails.cfm?SessionID=218689.

2 The Challenge Data Set

The New York City Housing and Vacancy Survey (NYCHVS) is a representative survey of the housing stock and population in the five boroughs within the city conducted by the US Census Bureau on behalf of the New York City Department of Housing Preservation and Development (HPD). It is the longest running housing survey in the United States and has been conducted about every three years since 1965. The NYCHVS is noteworthy in that it is the only Census product sponsored by a non-federal agency.

Detailed data from the NYCHVS cover many characteristics of New York City's housing units, building, and neighborhoods, including size and composition of the housing inventory, housing costs and affordability, housing quality, rent subsidy, tenure and building characteristics, and vacancy status. Data are also collected on the city's resident population, including demographics, education, income, household composition, and length of residence. Information from the survey is used for planning, program development, policy analysis, academic and applied research, and public information. One of the unique features of the NYCHVS is that it classifies housing units by type, such as public housing, unregulated/market rental, condominiums, cooperatively owned units, and rent controlled housing, among others.

Fieldwork for the NYCHVS is conducted face-to-face by trained Census Bureau Field Representatives throughout the five boroughs of New York City. The field period generally occurs during the first two quarters of the survey cycle year. Occupied interviews are conducted with the current occupant of a unit, who provides information about themselves, the unit, and—by proxy—each member of the household. Vacant interviews are conducted with a key informant who is knowledgeable about the sampled unit, such as the owner or managing agent.

For the *2019 Data Challenge Expo*, HPD provided participants with data from ten cycles of the NYCHVS. This covered the period from 1991 through 2017, including the 8^{th} through 17^{th} surveys. As described below, participants could choose to combine these data in a variety of ways, including as point-in-time representative estimates, or as longitudinal panels. For these three decades, the content of the NYCHVS remained substantially the same, enabling comparisons over time for many measures.

At the beginning of each decade, a representative sample of housing units is selected. All housing units in New York City other than group quarters and special places are eligible for inclusion in the survey, regardless of occupancy status.

In each successive survey cycle within a decade, the NYCHVS gathers information about these housing units and their current occupants. These units are supplemented with a small number of additional units to ensure that a given cycle's data are representative of the citywide housing stock at that point in time. This supplemental sample includes units that were newly constructed or altered or converted to residential use since the prior NYCHVS.

Because the same core sampled units remain in the survey across multiple cycles within a decade, the NYCHVS data are inherently longitudinal within a given decade. Data files from each survey cycle within a decade may be linked across years to observe the same housing unit over time. Participants in the *2019 Data Challenge Expo* could link housing units within two time periods: from 1991, 1993, 1996 and 1999; or from 2002, 2005, and 2008. (Data from 2011, 2014, and 2017 are longitudinal in nature but cannot be linked because of data disclosure avoidance protections.)

For the 8th through 17th survey cycles, NYCHVS data were released in three separate public use files: (1) occupied units (household/unit level), (2) population (person level), and (3) vacant units (unit level). The files can be combined to create a single dataset per the needs of the data user. To facilitate entries to the *2019 Data Challenge Expo*, HPD created and hosted a web page dedicated to the *2019 Data Challenge Expo*. The page included information about the survey and custom harmonized microdata files. The files were prepared to provide consistent variable names over time and made available in Stata, SAS, and CSV formats to lower the bar to entry for students wishing to participate. The web page also included codebooks for each of the ten cycles, a variable crosswalk file to help users understand the original public use file variables and harmonized variables, and additional details to support participants use of the data.

Participants in the *2019 Data Challenge Expo* were invited to consider five research questions tailored to the dataset to encourage their exploration. Contestants were also encouraged to submit admissions based on their own research question.

1. Create a quality index for the NYCHVS. How has the frequency of maintenance deficiencies changed over time?
2. For the last 50 years, part of the NYC rental stock has been subject to price controls. Currently, about half of the City's rental stock fall under these controls. Describe what the NYC rental market would be like if price controls were lifted. If the price control were lifted, what would the NYC rental market look like 10 years from now? Contestants may choose to look at a variety of factors such as quality, housing costs, or population.
3. Create a measure of gentrification to model and/or describe who is most affected. Data are available at a PUMA level. Contestants are welcome to add additional information to the dataset.
4. Describe changes in housing conditions in first and second generation immigrant householders in NYC.
5. What are the costs and benefits of renting and owning in New York City? Is one a better option than the other? Consider the quality of housing, income/debt/benefits, age/stability, etc.

3 Summary of Papers in this Special Issue

Five papers were submitted to this special issue and accepted through the usual journal referee process for *Computational Statistics*. The papers are described below. Supplemental materials for each paper in this issue can be found here https://github.com/asa-stat-computing-and-graphics/COST-DataExpo-2019.

The first article titled "House Quality Index Construction and Rent Prediction in New York City with Interactive Visualization and Product Design" (Shen et al. 2023) covers the analysis conducted by the first-place winners in the student category. Their work examined the housing conditions and price changes in New York City for the past 30 years. The authors define the quality of housing through a composite indicator that encapsulates other housing characteristics, such as floor conditions and availability of elevators. They took an innovative approach using ridge regression to construct the indicator. Spatio-temporal information was added to the housing quality and a predictive model was developed to estimate the rental costs in different areas of New York City. An R Shiny app was developed allowing users to explore historical data on housing quality, immigrant preferences, and expected rent, providing a useful tool for decision makers.

The second-place student team from Miami University led by Benjamin Schweitzer performed "An Analysis of the Impact of Rent Control on New York City Housing" (Schweitzer et al. 2023). Their goal was to assess the results of rent control policies on the housing market. They modelled the effects of the city's rent controls, along with rent stabilization, on housing quality over the previous 30 years. Various characteristics were used to assess the housing condition or damage as they relate to rent control. The authors found that poorer housing conditions were positively associated with rent-controlled housing.

The next paper in the issue is authored by one of the third place teams, also from Miami University. The title of their paper is "Immigrant Residency and Happiness in New York City" (Tuiyott et al. 2023). This team examined the quality of life immigrants in New York City experienced with respect to conditions in their neighborhoods, as well as individual housing. The student team combined the challenge data set with data from various New York City agencies, such as the Police Department, Department of Education, the Department of Health and Mental Hygiene, and the Happy City Index developed by the New Economics Foundation. Through combining, exploring, and modeling these data, Tuiyott et al. (2023) found that Manhattan and Staten Island had higher happiness scores and fewer immigrants, while the opposite was the case for the Bronx.

The second third-place student team first conducted and presented their analysis as a project for their class at the University of Virginia. The prize for winning their class competition was a trip to the Joint Statistical Meetings where they presented their work on "A Statistical Framework for Analyzing Housing Quality: A Case Study of New York City" (Chambon and Gerszten 2023). They developed a principal component based standardized index measuring housing quality that incorporated demographic, spatial, and economic characteristics. When applied

to the New York City housing data, they found that renters encountered poorer housing quality than homeowners did.

The final paper in this special issue is also by a student team, which was led by Jhonatan Medri (Medri et al. 2023). Their paper titled "Housing Variables and Immigration: An Exploratory Analysis in New York City" explored the challenge data at the borough and sub-borough levels. In particular, they looked at how immigration status related to home ownership, renting, and housing costs. They expanded their work beyond a visual exploration of the data and conducted hypothesis tests to assess spatial autocorrelation among the housing variables. An R Shiny app was developed to give others the opportunity to explore and analyze the data.

4 Supplementary Materials

The data challenge special issues published in *Computational Statistics* have followed the principles of reproducibility, starting with the 2013 Data Expo (Hofmann, et al. 2019). To facilitate the re-use of the software, data, and methods described in the papers, we asked authors to upload supplementary materials to a repository located here: https://github.com/asa-stat-computing-and-graphics/COST-DataExpo-2019. These materials include any files created as part of their analysis and article, such as figures, tables, additional data sets used, and any computer code (readme files, macros, SAS PROCS, R files, Shiny apps, etc.). All code and materials were reviewed as part of the referee process.

Acknowledgments The Guest Editors of this special issue are grateful for the support of Springer and the editors of *Computational Statistics* (COS) for giving us the opportunity to publish refereed articles written by authors competing in the *Data Challenge Expo* and for their help and patience as we prepared this special issue. In particular, we thank COS Editor, Usha Govindarajulu, for her patience and guidance and Danny Yang from the Bureau of Labor Statistics for reviewing the supplemental materials.

References

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

Cetinkaya-Rundel M, Martinez W (2023) The 2018 *data challenge expo* of the American statistical association. Comput Stat 38:1117–1122. https://doi.org/10.1007/s00180-023-01363-3

Chambon D, Gerszten J (2023) A statistical framework for analyzing housing quality: a case study of New York City. Comput Stat. https://doi.org/10.1007/s00180-023-01394-w

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

Garner TI, Martinez W (2022) The 2017 *data challenge* of the American Statistical Association. Comput Stat 37:2087–2094. https://doi.org/10.1007/s00180-022-01257-w

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

Medri J, Probst BD, Symanzik J (2023) Housing variables and immigration: an exploratory analysis in New York City. Comput Stat. https://doi.org/10.1007/s00180-023-01412-x

Murrell P (2010) The 2006 data expo of the American statistical association. Comput Stat 25:551–554. https://doi.org/10.1007/s00180-010-0207-3

Schweitzer BW, Garrett R, Carter L, Tuiyott A, Maurer K, Fisher TJ (2023) An analysis of the impact of rent control on New York City housing. Comput Stat. https://doi.org/10.1007/s00180-023-01397-7

Shen X, Luo S, Zhang M (2023) House quality index construction and rent prediction in New York City with interactive visualization and product design. Comput Stat. https://doi.org/10.1007/s00180-023-01391-z

Tuiyott A, Garrett RC, Carter L, Schweitzer B, Maurer K, Fisher TJ (2023) Immigrant residency and happiness in New York City. Comput Stat. https://doi.org/10.1007/s00180-023-01392-y