Journal of Data Science 21 (2), 173–176
April 2023

DOI: 10.6339/23-JDS212EDI
Editorial

Editorial: Symposium on Data Science and Statistics 2022

Claire McKay Bowen[1],* and Michael J. Grosskopf[2]

[1]*Labor, Human Services, and Population and Technology and Data Science at the Urban Institute, USA*
[2]*Statistical Science Group at the Los Alamos National Laboratory, USA*

The COVID-19 pandemic continues to impact global society, exacerbating existing social inequalities, housing insecurity, disruptions to the education system, job losses, and demand for reformation of healthcare and education systems. In light of these challenges, the *Symposium on Data Science and Statistics (SDSS) 2022* focused on the theme, "Influencing Science, Technology, and Society." Through this theme, we wanted to showcase the data science and statistics community's incredible work in advancing society through these turbulent times.

This special issue highlights 15 articles that demonstrate the diverse ways data science and statistics contributes to social progress. All articles in this special issue were peer-reviewed.

Data science in action

Ankerst and Neumair (2023) proposes four data science-driven strategies to improve clinical risk prediction: actively designing prospective data collection, making risk tools available online, dynamically updating risk tools, and accommodating missing data at both the training and end-user stage. Their work is motivated from clinical risk prediction models being developed in post-hoc and passive ways and often end at their publication instead of helping patients. Bradford and VanderPlas (2023) explores whether interactive dashboards can be used to empower novice analysts to make data-driven decisions, using Iowa data as an example with their primary audience being leaders of small towns in Iowa. They provide suggestions for future work to better support small and rural places from shrinking using an interactive dashboard design, implementation, and use. Cheng and Liao (2023) examines 11 different classification models to determine the best-performing model for predicting and diagnosing mesothelioma (a rare but aggressive cancer). Their goal is to provide a comprehensive empirical comparison that will help with early diagnosis since mesothelioma is usually identified in late stages. They found that random forest performed the best based on sensitivity, where age and duration of asbestos exposure ranked as the most important features affecting diagnosis. Daas et al. (2023) presents an interesting application of machine learning to text data sets across different languages – specifically Spanish, English and Italian. They compare a suite of supervised classification approaches to identify the websites of European drone companies, identifying logistic regression as the highest performing model. The insight into the use of a translation list of key terms in relation to their model and its performance on novel languages was of particular interest. Youngs et al. (2023) present the visualization and data communication methods developed for the 2020 Census Assessment Tool. This tool, developed by the authors, uses advanced approaches to visual communication of complex data from the 2020 Census and provides a way for individuals to assess any unexpected results. This is particularly critical work due to the upheaval of 2020 and the extreme circumstances under which the Census was conducted. Zhou et al. (2023) summarizes fifteen

*Corresponding author. Email: cbowen@urban.org.

boundaries that consist of five error spending functions that allow early termination for futility, difference, or both, as well as a fixed sample size design without interim monitoring. They implemented a rigorous simulation study to determine if general guidance could be given for A/B experimentation.

Computing in data science

Peterson et al. (2023) proposed a new approach to causal discovery, particularly for small sample size data. Leveraging convolutional neural networks, the proposed model is demonstrated to be competitive with current state-of-the-art methods in causal discovery like the PC and greedy equivalence search algorithms. Their approach shows particular value on small samples and with conservative identification of causal effects. Robinson, Howard, and VanderPlas (2023) adapts the New York Times interactive feature, "You Draw It," for graphical testing and evaluation when using in modern web-applications. They present an empirical evaluation of this testing method for linear regression, and briefly discuss an extension of this method to non-linear applications. Schmid and Hunter (2023) proposes to estimate maximum pseudolikelihood estimator standard errors for exponential random graph models using an estimated Godambe matrix. Their results provide empirical evidence for the asymptotic normality of the maximum pseudolikelihood estimator under certain conditions. Wu and Zhang (2023) propose penalized discriminant analysis as able to efficiently handle high-dimension, low-sample-size classification tasks. They highlight the improved performance and robustness compared against linear discriminant analysis and support vector machine classification. Sartore et al. (2023) propose a neural network approach to long time-series modeling of crop-specific land cover. The authors compare multiple neural network approaches, including dense networks, recurrent networks, and quantum-inspired networks to the conventional high-order Markov chain approach to prediction. Despite the theoretically desirable properties of quantum-inspired neural networks, the dense networks provided optimal performance across geographic regions.

Statistical data science

Bang and Oh (2023) propose a Bayesian network learning algorithm based on sparse Cholesky decomposition. Their FROSTY approach is shown to give strong performance at recovery of directed acyclic graphs when compared to other current state-of-the-art while having greatly reduced computational complexity. This leads to a scalable, accurate graph learning approach. Thorp et al. (2023) proposes an extension to existing random forest of interaction trees methods to account for ordinal treatments. These methods aim to estimate the individualized treatment effect to account for heterogeneity that is not accounted for in the average treatment effect. The existing methods have been developed for binary and categorical treatments and incorporate propensity scores in the construction of the random forest. The proposed method allows for ordinal treatments allowing for more accurate estimation of the individualized effect when the treatment is ordered and is demonstrated on simulation studies and educational data. Pomeyie et al. (2023) present a comparison of methods for extreme value modeling applied to snow depth data. The authors compare six approaches to estimation of the 50-year mean recurrence interval and assess the relative performance of each. No single estimator dominates, with the conditions of best performance of each estimator characterized and discussed.

Education in data science

Seo and Dogucu (2023) argues that the focus on visualization in data science courses creates barriers for individuals with visual impairments or learning disabilities and that instructors should teach multiple data representation methods to make data products more accessible. They recommend that accessibility be taught early on in data science curricula and share specific examples from two institutions where accessibility is already taught in lower-division courses.

We extend our gratitude to the four associate editors from the original program committee (Elizabeth Chase, Emily Griffith, Alicia Lamere, and Maria Tackett) for their dedication and diligence in overseeing the peer-review process for all the articles. We also thank the anonymous reviewers for their insightful reviews of the papers.

Our hope is that these articles inspire readers to engage in collaborative efforts to influence and improve our society through the powerful tools of data science and statistics.

References

Bang J, Oh S-Y (2023). FROSTY: A High-dimensional Scale-free Bayesian Network Learning Method. *Journal of Data Science*, 21(2): 354–367. https://doi.org/10.6339/23-JDS1097

Bradford D, VanderPlas S (2023). Exploring Rural Shrink Smart Through Guided Discovery Dashboards. *Journal of Data Science*, 21(2): 193–204. https://doi.org/10.6339/22-JDS1080

Cheng TSY, Liao X (2023). Binary Classification of Malignant Mesothelioma: A Comparative Study. *Journal of Data Science*, 21(2): 205–224. https://doi.org/10.6339/23-JDS1090

Daas P, de Miguel B, de Miguel M (2023). Identifying Drone Web Sites in Multiple Countries and Languages with a Single Model. *Journal of Data Science*, 21(2): 225–238. https://doi.org/10.6339/23-JDS1087

Pauler Ankerst D, Neumair M, (2023). Active Data Science for Improving Clinical Risk Prediction. *Journal of Data Science*, 21(2): 177–192. https://doi.org/10.6339/22-JDS1078

Petersen AH, Ramsey J, Ekstrøm CT, Spirtes P (2023). Causal discovery for observational sciences using supervised machine learning. *Journal of Data Science*, 21(2): 255–280. https://doi.org/10.6339/23-JDS1088

Pomeyie KK, Bean B, Sun Y (2023). Comparing Extreme Value Estimation Techniques for Short-Term Snow Accumulations. *Journal of Data Science*, 21(2): 368–390. https://doi.org/10.6339/23-JDS1086

Robinson EA, Howard R, VanderPlas S (2023). 'You Draw It': Implementation of visually fitted trends with r2d3. *Journal of Data Science*, 21(2): 281–294. https://doi.org/10.6339/22-JDS1083

Sartore L, Boryan C, Dau A, Willis P (2023). An Assessment of Crop-Specific Land Cover Predictions Using High-Order Markov Chains and Deep Neural Networks. *Journal of Data Science*, 21(2): 333–353. https://doi.org/10.6339/23-JDS1098

Schmid CS, Hunter DR (2023). Computing Pseudolikelihood Estimators for Exponential-Family Random Graph Models. *Journal of Data Science*, 21(2): 295–309. https://doi.org/10.6339/23-JDS1094

Seo J, Dogucu M (2023). Teaching Visual Accessibility in Introductory Data Science Classes with Multi-Modal Data Representations. *Journal of Data Science*, 21(2): 428–441. https://doi.org/10.6339/23-JDS1095

Thorp J, Levine RA, Fan J (2023). Random forest of interaction trees for estimating individualized treatment regimes with ordered treatment levels in observational studies. *Journal of Data Science*, 21(2): 391–411. https://doi.org/10.6339/23-JDS1084

Youngs I, Dick C, Prevost R (2023). Creating a Census County Assessment Tool for Visualizing Census Data. *Journal of Data Science*, 21(2): 239–254. https://doi.org/10.6339/22-JDS1082

Zhang C, Wu Z (2023). Assessment of projection pursuit index for classifying high dimension low sample size data in R. *Journal of Data Science*, 21(2): 310–332. https://doi.org/10.6339/23-JDS1096

Zhou W, Kroehl M, Meier M, Kaizer A (2023). Building a Foundation for More Flexible A/B Testing: Applications of Interim Monitoring to Large Scale Data. *Journal of Data Science*, 21(2): 412–427. https://doi.org/10.6339/23-JDS1099