Journal of Data Science 22 (2), 173–175
April 2024
DOI: 10.6339/24-JDS222EDI
Editorial

Editorial: Inquire, Investigate, Implement, Innovate – Symposium on Data Science and Statistics 2023

Emily Dodwell[1,*] and Amanda A. Koepke[2]

[1]*Chief Data Office, AT&T, Bedminster, NJ, USA*
[2]*Statistical Engineering Division, National Institute of Standards and Technology, Boulder, CO, USA*

In this special issue, we are pleased to present a selection of 10 peer-reviewed papers that capture the spirit of the theme of the 2023 *Symposium on Data Science and Statistics (SDSS)*: "Inquire, Investigate, Implement, Innovate." Contributors introduce new modeling methodologies and assess the performance of established methods on new applications; they motivate the development of new data science software and discuss the practicalities of implementing experimental procedures. These papers and their authors represent the diversity of topics and fields that make up the community of data science and statistics. They demonstrate the requisite curiosity, creativity, and iteration that progress our field with new techniques and valuable breakthroughs.

Data Science in Action Wiederich and VanderPlas (2024) present the results of an initial experiment they conducted to determine whether the decrease in accuracy often associated with interpretation of information from a 3D chart still applies for 3D virtual renderings and 3D-printed bar charts. The authors extend a 1984 study by Cleveland and McGill to evaluate perception accuracy of 2D, 3D virtual renderings (which include the ability of users to interact with and rotate the charts), and 3D-printed bar charts. By asking participants to estimate the ratio between two bars (appearing in some assignments adjacent to each other and separated in others) across these three mediums, the authors present initial results that suggest that errors in interpretation of these 3D bar charts are no greater than that of their 2D counterparts. They identify opportunities to improve and adjust subsequent experiments with additional participants, inclusion of a fixed 3D perspective chart, and manipulations of the data collection app's slider measurement methodology. Yavuz Ozdemir et al. (2024) explore the application of the Sparse Wrapper AlGorithm (SWAG) to the detection of Attention Deficit Hyperactivity Disorder (ADHD) using data collected by the ADHD-200 Consortium. This iterative multi-model selection method searches low dimensional combinations of features for a chosen learning mechanism and selects those that deliver the highest prediction accuracy. The authors demonstrate that SWAG delivers predictive performance similar to previously published results on ADHD detection, and the selected features enable identification of important brain regions and the connections between them, thereby providing insight into the potential for neuroimaging data as a clinical diagnostic tool.

Computing in Data Science Cairns et al. (2024) introduce *largescaler*, a proof of concept platform for distributed modeling that enables definition of novel statistical algorithms in a syntax that is familiar to `R` users. This platform provides different levels of abstraction for interaction with distributed data for different types of users – analyst, researcher, developer,

*Corresponding author. Email: ed720d@att.com.

© 2024 The Author(s). Published by the School of Statistics and the Center for Applied Statistics, Renmin University of China. Open access article under the CC BY license.

and architect – via four packages intended to serve their respective needs. The authors illustrate the use and contributions of the *largescaler* platform with an example of distributed LASSO regression.

Statistical Data Science Fagnant et al. (2024) present a point-to-area random effects (PARE) model which addresses a change-of-support problem for spatio-temporal extreme value modeling, enabling inferences and predictions at the areal level with data observed at the point level. The effectiveness of this modeling approach is showcased using both simulated data and real rainfall data from the Houston, TX area. Haycock et al. (2024) introduce the `stressor` package, offering an `R` interface to `Python`'s `PyCaret` package which enables automatic tuning and training of multiple machine learning (ML) models for accuracy comparisons. Demonstrated on agricultural datasets for crop suitability and yield prediction, `stressor` streamlines full ML benchmarking workflows into just a few lines of `R` code, making accessible various ML models with minimal programming effort. Nzekwe et al. (2024) study the predictive performance of two approaches for interaction selection in high-dimensional data: penalty-driven algorithms and tree-based ensemble algorithms. Comparative analysis reveals that results vary depending on data dimension and structure and clarifies the strengths and limitations of each algorithm. Rice et al. (2024) use a large, nationally-representative, probability-based panel of survey respondents to test perception in stacked bar charts, exploring different design structures and viewer behavior. The authors offer actionable insights for data visualization professionals, such as the importance of lining up comparisons of interest and tips for interactive bar plot visualizations. Yang et al. (2024) develop an interaction survival tree approach to identify heterogeneous treatment effects within subgroups of subjects experiencing recurrent events. The new method is evaluated through a simulation study and then used to identify subgroups in a randomized phase III clinical trial for colorectal cancer.

Education in Data Science Rogers and VanderPlas (2024) describe methodology for and analyze results of their initial experiment to assess jury understanding and conclusions when a bullet matching algorithm and demonstrative evidence are introduced and explained as part of expert testimony in a criminal trial. Participants were asked to rate their impression of credibility, reliability, and scientificity when the bullet matching algorithm and demonstrative evidence were included or not. The authors discover that their ability to draw statistical conclusions from responses were limited by scale compression – that is, the majority of participants selected high ratings regardless of experimental condition – and they propose adjustments that may alleviate this effect in future experiments. Thatcher et al. (2024) introduce three interdisciplinary courses offered as part of Truman State University's Masters in Data Science and Interdisciplinary Storytelling, and they analyze survey responses to assess the experiences of their first cohort of students who completed them. Narrative, Argument, and Persuasion in Data Science; Principles of Design in Data Visualization; and Big Data Ethics and Security were designed in collaboration with faculty across disciplines. Graduate students found coursework engaging and applicable to their careers, and the authors share intended course outcomes and Signature Assessment materials for the benefit of other educators.

We would like to express our sincere appreciation for the four associate editors from the SDSS 2023 Program Committee – Xiaoyue Cheng, Owais Gilani, Ritwik Mitra, and Justin Strait – for their time and care in managing the peer-review procedure for all submissions. Thank you to the anonymous referees with expertise in the topics represented who ensured a thorough, well-rounded review of each paper.

References

Cairns J, Urbanek S, Murrell P (2024). A platform for large scale statistical modelling in R. *Journal of Data Science*. 22(2): 208–220.

Fagnant C, Schedler JC, Ensor KB (2024). Spatial-temporal extreme modeling for point-to-area random effects (PARE). *Journal of Data Science*. 22(2): 221–238.

Haycock S, Bean B, Burchfield E (2024). Producing fast and convenient machine learning benchmarks in R with the stressor package. *Journal of Data Science*. 22(2): 239–258.

Nzekwe CJ, Kim S, Sayed M (2024). Interaction Selection and Prediction Performance in High-Dimensional Data: A Comparative Study of Statistical and Tree-Based Methods. *Journal of Data Science*. 22(2): 259–279.

Rice K, Hofmann H, du Toit N, Mulrow E (2024). Testing perceptual accuracy in a U.S. general population survey using stacked bar charts. *Journal of Data Science*. 22(2): 280–297.

Rogers R, VanderPlas S (2024). Demonstrative Evidence and the Use of Algorithms in Jury Trials. *Journal of Data Science*. 22(2): 314–332.

Thatcher S, Alberts S, Beregovska T (2024). Interdisciplinary Approaches to Teaching Data Science. *Journal of Data Science*. 22(2): 333–351.

Wiederich T, VanderPlas S (2024). Evaluating Perceptual Judgements on 3D Printed Bar Charts. *Journal of Data Science*. 22(2): 176–190.

Yang Y, Perera C, Miller JP, Su X, Liu L (2024). Precision Medicine: Interaction Survival Tree for Recurrent Event Data. *Journal of Data Science*. 22(2): 298–313.

Yavuz Ozdemir Y, Nukala NCP, Molinari R, Deshpande G (2024). A Multi-Model Framework to Explore ADHD Diagnosis from Neuroimaging Data. *Journal of Data Science*. 22(2): 191–207.