Journal of Data Science 23 (2), 265–268
April 2025
DOI: 10.6339/25-JDS232EDI
Editorial

Editorial: Beyond Big Data: Bridging the Gap Between Theory and Practice—Symposium on Data Science and Statistics 2024

Kristine Gierz[1,*], Owais Gilani[2], and Julia Schedler[3]

[1]*Statistical Engineering Division, National Institute of Standard and Technology, Gaithersburg, MD, USA*
[2]*Associate Professor (Biostatistics), Tufts University School of Medicine, Boston, MA, USA*
[3]*Assistant Professor (Statistics), California Polytechnic State University, Bailey College of Science and Mathematics, San Luis Obispo, CA, 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 2024 Symposium on Data Science and Statistics (SDSS): "Beyond Big Data: Bridging the Gap Between Theory and Practice." Contributors highlight the richness and diversity of research in statistics and data science, introducing new modeling methodologies, advancing computational techniques, and assessing the impact of emerging technologies. Their work spans foundational theoretical developments, innovative applications, and pressing societal challenges. These papers and their authors reflect the breadth of inquiry that defines the data science and statistics community. They exemplify the curiosity, creativity, and iterative refinement that drive the field forward, bridging the gap between theory and practice to generate meaningful insights and real-world impact.

Education in Data Science McGee and Sadler (2025) examine the accuracy and textual characteristics of different versions of OpenAI's ChatGPT—GPT-3.5, GPT-4, and GPT-4o-mini—when answering a graduate-level statistics exam. With the increasing use of generative AI as a tutoring tool, the authors investigate whether free and paid AI versions provide equitable learning experiences. While GPT-4 and GPT-4o-mini produce more statistically relevant and structured answers, GPT-3.5 often provides incomplete or tangential explanations. The study highlights the growing digital divide in AI-assisted education, emphasizing the need for institutions to address disparities in access to high-performing AI tools.

Statistical Data Science Maboudou-Tchao, Agbemade, and Chung (2025) introduce the Rescale Hinge Loss Support Vector Data Description (RSVDD), an improved extension of the SVDD model designed to enhance robustness against anomalies and outliers in one-class classification. Using the half-quadratic optimization method, the authors develop an efficient dynamic optimization algorithm to improve classification accuracy. Experimental results on synthetic and breast cancer datasets demonstrate that RSVDD outperforms standard SVDD, Density-Weighted SVDD (DW-SVDD), and Stahel-Donoho SVDD (SD-SVDD), particularly in maintaining lower Type II error rates.

Johnson and Mostafa (2025) explore the impact of statistical disclosure control methods on machine learning performance, focusing on General Additive Data Perturbation (GADP) and Copula-based General Additive Data Perturbation (CGADP). These techniques aim to protect sensitive data while preserving statistical utility, yet their effects on predictive modeling remain

*Corresponding author. Email: kristine.gierz@nist.gov or gierz.kristine@gmail.com.

understudied. The authors find that parametric models such as logistic regression and LASSO perform well under both GADP and CGADP, while nonparametric models, including tree-based methods and neural networks, are more sensitive to perturbation and prone to overfitting.

Jeremy Flood and Sayed A. Mostafa (2025) investigate the integration of data from probability and nonprobability survey samples in their paper, introducing the matched mass imputation (MMI) approach as a solution for improving estimations in such settings. The MMI method leverages statistical matching and mass imputation to address nonignorable bias in nonprobability samples by using shared covariates between the two sample types. The authors demonstrate that MMI outperforms other imputation estimators in the presence of nonignorable bias, especially when employing nearest-neighbor matching. They also explore variance estimation using bootstrapping, highlighting the need for further research to address potential underestimation of sampling variance when the ignorability assumption is violated.

Chen, Habans, Douthat, Losh, Dekharghani, and Lin (2025) study migration patterns driven by extreme environmental events. Motivated by a real dataset about human migrations, this paper develops a transformed varying coefficient model for origin and destination (OD) regression to elucidate the complex associations of migration patterns with spatio-temporal dependencies and socioeconomic factors. Existing studies often overlook the dynamic effects of these factors in OD regression. They address the challenge by proposing a new Bayesian interpretation for the proposed OD models, leveraging sufficient statistics for efficient big data computation.

Abeykoon et al. (2025) investigate the double descent phenomenon in machine learning, a surprising pattern where test error initially decreases, then increases, and finally decreases again as model complexity grows. Using a two-layer neural network with a ReLU activation function for binary classification, they analyze the mathematical foundation of this behavior and introduce a ratio-based perspective on over- and under-parameterization. The authors' findings confirm that double descent occurs when regularization is weak, but appropriate regularization mitigates the effect.

Computing in Data Science Urbanek (2025) addresses one of the most pressing challenges in modern data science: how to scale not only computation but also collaboration. As datasets grow in size and complexity, analysis has shifted from lone practitioners on personal machines to teams working on distributed clusters. RCloud, an open-source, web-based platform, decouples users from data locations while preserving security, interactivity, and rich visualization. Its built-in collaborative tools allow data scientists in both industry and academia to explore, share, and co-author analyses seamlessly. The authors detail the key design decisions that make RCloud a robust solution for large, distributed data-science teams.

DePratti and Singh (2025) investigate computational improvements to the R package `lcpm`, which implements the Log Cumulative Probability Model (LCPM) for ordinal regression. Unlike traditional logit-link models, LCPM estimates probabilities directly using a log-link, requiring constrained maximum likelihood estimation (cMLE). While this approach improves interpretability, it poses computational challenges, particularly for large datasets. The authors identify three key runtime bottlenecks—marshaling, the objective function, and the optimizer—and test alternative implementations to improve efficiency. Their most significant enhancement comes from replacing the `constrOptim` optimizer with `auglag`, which reduces runtime from over two hours to under six minutes for large datasets with multiple predictors. These optimizations enable the `lcpm` package to handle big data more effectively, making it a more practical tool for large-scale ordinal regression analysis.

Data Science in Action Tezbasaran and Ricci (2025) explore the transformative role of ChatGPT in academic data science consultation services at North Carolina State University. Their program, which assists students, faculty, and staff with diverse data-related inquiries, has integrated ChatGPT to enhance efficiency and broaden support capabilities. The authors emphasize how generative AI accelerates workflows by assisting with data visualization, statistical analysis, and code generation, reducing the time consultants spend on routine tasks. Additionally, they discuss best practices for AI-assisted consultations, including prompt engineering and mitigating response variability.

Brian Dumbacher, Daniel Whitehead, Jiseok Jeong, and Sarah Pfeiff (2025) investigate the Business Establishment Automated Classification of NAICS (BEACON) tool in their paper, which is designed to help respondents to the U.S. Census Bureau's economic surveys self-classify their business activity in real time. BEACON uses natural language processing, machine learning, and information retrieval techniques to classify businesses into the appropriate North American Industry Classification System (NAICS) codes. The authors highlight BEACON's success in handling a wide vocabulary, returning relevant results, and reducing the clerical workload associated with manual coding. Additionally, BEACON's implementation is discussed in the context of reducing costs and errors compared to traditional methods, with ongoing work to improve its training data and extend its functionality, such as better handling of Spanish-language write-ins.

We would like to express our sincere appreciation for the four associate editors from the SDSS 2024 Program Committee – Xiaoyue Cheng, Stephanie Shipp, Maria Tacket, and Robert Tumasian – for their time and care in managing the peer-review procedure for all submissions. We thank the anonymous referees with expertise in the topics represented who ensured a thorough, well-rounded review of each paper.

References

Abeykoon CS, Beknazaryan A, Sang H (2025). The double descent behavior in two layer neural network for binary classification. *Journal of Data Science*, 23(2): 370–388. https://doi.org/10.6339/25-JDS1175

Chen T, Habans R, Douthat T, Losh J, Chalangar Jalili Dehkharghani L, Lin LH (2025). Exact inference for transformed large-scale varying coefficient models with applications. *Journal of Data Science*, 23(2): 353–369. https://doi.org/10.6339/25-JDS1181

DePratti R, Singh G (2025). The journey to improve LCPM: An R package for ordinal regression. *Journal of Data Science*, 23(2): 399–415. https://doi.org/10.6339/25-JDS1183

Dumbacher B, Whitehead D, Jeong J, Pfeiff S (2025). BEACON: A tool for industry self-classification in the economic census. *Journal of Data Science*, 23(2): 429–448. https://doi.org/10.6339/25-JDS1180

Flood J, Mostafa SA (2025). Matched mass imputation for survey data integration. *Journal of Data Science*, 23(2): 332–352. https://doi.org/10.6339/25-JDS1179

Johnson III T, Mostafa SA (2025). Impact of data perturbation for statistical disclosure control on the predictive performance of machine learning techniques. *Journal of Data Science*, 23(2): 312–331. https://doi.org/10.6339/25-JDS1186

Maboudou-Tchao EM, Agbemade E, Chung J (2025). Rescale hinge loss support vector data description. *Journal of Data Science*, 23(2): 287–311. https://doi.org/10.6339/25-JDS1185

McGee M, Sadler BP (2025). Generative AI takes a statistics exam: A comparison of performance between ChatGPT3.5, ChatGPT4, and ChatGPT4o-mini. *Journal of Data Science*, 23(2): 269–286. https://doi.org/10.6339/25-JDS1174

Tezbasaran A, Ricci S (2025). Embracing the AI revolution: ChatGPT's role in advancing data science consultation services. *Journal of Data Science*, 23(2): 416–428. https://doi.org/10.6339/25-JDS1184

Urbanek S (2025). RCloud – Collaborative visualization and analysis platform. *Journal of Data Science*, 23(2): 389–398. https://doi.org/10.6339/24-JDS1153