Journal of Data Science 24 (3), 477–481
July 2026
DOI: 10.6339/26-JDS243EDI
Editorial

Editorial: Government Advances in Statistical Programming (GASP) 2025

Lisa M. Frehill[1,*], Peter B. Meyer[2,†], and José Bayoán Santiago Calderón[3,†]

[1] *George Mason University, Schar School of Policy and Government, Arlington, VA, USA*
[2] *U.S. Bureau of Labor Statistics, Office of Productivity and Technology, Suitland, MD, USA*
[3] *U.S. Bureau of Economic Analysis, National Economic Accounts, Suitland, MD, USA*

Context: Increasing Relevance of Data Science in Government Settings

The Government Advances in Statistical Programming (GASP) conference started as a one-day workshop in 2018 to encourage statisticians and data scientists working on government projects to share technical approaches and strategies to enable use of the expanding set of data science tools in government settings, where data security, privacy, and confidentiality are of key importance. The GASP conference was established as the annual event by the Office of Management and Budget's (OMB) interagency Federal Committee on Statistical Methodology (FCSM) Data Science for Federal Statistics (DSFS) interest group. Originally called the Computational Statistics for the Production of Official Statistics (CSPOS) interest group, in December 2024 the group was rechartered to reflect the increasing use of data science techniques and the associated workforce changes to adapt to these new needs. (Federal Committee on Statistical Methodology, 2026).

The context has changed somewhat over time. A bipartisan U.S. Commission on Evidence-Based Policy began in 2016 to investigate ways to more effectively use the data collected across federal agencies to improve the efficacy, efficiency, operations, and delivery of government services. The Commission completed a report after an 18-month series of hearings and meetings with experts in areas like government data security, confidentiality and privacy protections, and the proliferating data science technical tools. The report addressed technical steps needed to match data across various government platforms while meeting confidentiality and privacy requirements (Commission on Evidence-Based Policy, 2017), which was incorporated into a new law, the Foundations for Evidence-Based Policymaking Act of 2018 (Public Law 115–435). This policy work provided a context in which skillsets of data scientists, statisticians, and social scientists would be increasingly important as agencies sought to deploy new tools to improve program operations and delivery. Notably, the White House's Office of Personnel Management (OPM) set up a new Data Science occupational series in 2021 in recognition of this skillset change. Sophisticated policy work along this line led to novel research, some of which appears in these pages.

The combined technical growth and evolution of data science tools along with the policy imperatives to better deploy those tools for increased efficacy and efficiency are part of the context for the expansion of GASP. By 2023, the GASP conference had grown to include sub-

*Corresponding author. Email: lisa.frehill@gmail.com.

†The views expressed in this paper are those of the authors and do not necessarily represent the U.S. Bureau of Economic Analysis, the U.S. Department of Commerce, the U.S. Bureau of Labor Statistics, or any other U.S. government agency.

stantially more participants and had become a multi-day online event using tools that became more common during the COVID-19 period.

With a host of federal workforce changes in early 2025, the June 2025 GASP conference (Data Science for Federal Statistics Interest Group, 2025) expanded again to include four days of content to accommodate the increased interest in presentations by social scientists, statisticians, and data scientists working on government projects. The American Statistical Association generously provided the platform for the conference and enthusiastically supported the expansion to four days of workshops, paper sessions, and lightning presentation sessions.

Special Issue: 2025 GASP Conference

This is the second special issue of the *Journal of Data Science* with articles based on GASP conference presentations. The earlier issue was published in July 2024 based on presentations from the 2023 GASP conference. All articles were in JDS topic areas of **Data Science in Action** (5 in 2024, 3 in 2026) or **Statistical Data Science** (2 in 2024, 5 in 2026). The newer work is quite technical, reflecting the ongoing incorporation of sophisticated techniques in government settings.

The lead article, by Belyaeva et al. (2026), tackles a key issue for data scientists who seek to use many federal data sets: how to accurately ingest survey metadata along with the actual data. While metadata are a rich trove of information, variations in their structure can pose challenges for large language models (LLMs) in ingesting this information to accurately analyze the associated data. Belyaeva et al. propose knowledge graphs as a framework for solving this problem. Because they use the U.S. American Community Survey (ACS, U.S. Census Bureau, 2026) – a widely used publicly available dataset – the tools they propose are likely to have wide application. Indeed, two other articles in this issue, one by Champney and Qin (2026) and the other by Saluja et al. (2026), show further examples of ACS' utility to data scientists.

Four articles propose various data science solutions to the ongoing problem of developing robust, reliable, and valid government statistics while survey response rates continue to decline. A **Statistical Data Science** article by Elkasabi et al. (2026) developed a statistical framework for combining data from probability and non-probability samples to develop robust estimates for official statistics. U.S. federal agencies are required to implement sampling methodologies for which the probability of selection of each respondent is a known quantity (e.g., random samples) to use in developing statistical estimates. As response rates have declined in many survey programs, the presence of other surveys that use non-probability sampling methods (e.g., web-based surveys that rely on self-response) offers one way to address the declining response rate issue. But with their unknown selection probabilities, accurately combining results from non-probability samples with those from probability samples has been a persistent challenge. These authors provide a framework for addressing this challenge and include their case study R code as supplementary material to facilitate use by others.

Three **Data Science in Action** articles offer strategies to enhance use of alternative data sources via data matching techniques at various levels. Champney and Qin (2026) provide a review of the literature on spatial matching to show how housing unit ACS data can be improved by taking advantage of data from commercial sources. They developed a multi-step process that "built upon the Census Bureau's standard production process by integrating cross-file reconciliation, geospatial joins using parcel boundaries, and fuzzy string matching." The improvement of match rates from 61.6% to 70.0% represents a critical improvement for official statistics operations.

Rhodes and his colleagues at the USDA's National Agricultural Statistics Service (NASS) explored how data logged directly from farm machinery could supplement or replace traditional establishment survey data for developing crop production statistics. In their pilot study, about 90% of the machine-logged fields could be matched to NASS records (Rhodes et al., 2026). Substantial adjustments were needed to impute survey values from derived values based on the farm machinery data. As the Internet of Things (IoT) continues to expand, such approaches can become increasingly useful. Challenges remain, requiring careful data science work, but such strategies could gather data quickly, reduce respondent burden, and enable more timely reporting.

Finally, Saluja et al. (2026) approach the ongoing problem of survey non-response by exploring how survey researchers could better dedicate data collection resources to increase responses from hard-to-survey populations. They used Medical Expenditure Panel Survey (MEPS, Agency for Healthcare Research and Quality, 2026) paradata and ACS "neighborhood" (Census tract-level) demographic data to propose a way to improve the process of increasing the MEPS response rate. A set of ACS neighborhood demographic characteristics, along with MEPS advance call records and contact paradata, was used to predict respondent cooperation patterns with multinomial logistic regression (MLR). The authors used a novel machine learning (ML) approach to group contact mode sequences, identifying five clusters of respondents, each with different completion rates, and then showed how these related to neighborhood demographics. Survey managers can use this information to more efficiently manage data collection early in their process.

The remaining three articles make various contributions in the area of **Statistical Data Science**. Yin et al. (2026) used an elegant approach to deploy several data science and statistical techniques to improve the estimation of university enrollment prediction. For colleges and universities, accurate prediction of fall enrollment helps avoid having either too few or too many resources to support the incoming class of first-year students. "For enrollment prediction, uncertainty is as consequential as accuracy for effective resource planning." Their article presents results of simulations using three model families (logistic, tree-based, and ensemble) and incorporating data from several sources (individual-level data about applicants, institutional data on historical enrollment patterns, and admissions-review assessments) to present a more nuanced approach to predicting incoming cohort size.

The article by Williams et al. (2026) also deals with uncertainty by presenting and evaluating various adjustments to variance-covariance estimators to obtain efficient confidence intervals with correct coverage for generalized linear mixed models under various survey designs, sampling mechanisms, and hierarchical structure conditions. The evaluation and practical recommendations are illustrated with an example from the National Survey on Drug Use and Health (RTI International, 2026).

Preiss et al. suggest a practical guide to differentially private deep learning. As discussed in the context section above, protecting respondent privacy is a critical ethical and legal requirement associated with the U.S. government's collection of data from individuals. Preserving privacy while implementing machine learning methods can be challenging. Respondents – or their guardians – routinely acknowledge their understanding of their right to privacy and the researchers' protections meant to protect these rights prior to answering survey questions or participating in an interview. Preiss et al. propose a pseudo posterior mechanism to balance "predictive performance with rigorous privacy protection... when training neural networks on sensitive individual-level data." Using Occupational Safety and Health Administration (OSHA) data on occupational injuries, they show that "diagnostic-guided tuning within SWAG-PPM

[Stochastic Weight Averaging-Gaussian Pseudo Posterior Mechanism] can achieve strong privacy protection with minimal loss in accuracy."

Acknowledgements

We greatly appreciated Dr. Jun Yan's invitation to produce a JDS special issue with GASP presentations. This has been an important way for government statistical scientists to gain greater exposure for their work beyond government settings. Scientists completing projects within U.S. government settings – whether they are federal employees or in contracting organizations – have additional administrative challenges compared to researchers in academic settings. Data sources have considerable privacy and confidentiality requirements and manuscripts undergo rigorous internal scrutiny before they can be submitted. These privacy and confidentiality issues created additional burdens for the JDS reproducibility check process, which was ably managed by Dr. Jing Zhou. We are grateful to her for her rapid work to address the various special circumstances associated with validating the reproducibility of government data science research projects.

This volume would not have been possible without the work of many specialized reviewers and associate editors, including: John Eltinge, Gwynn Gebeyehu, Emily Gentles, Tomo Lazovich, Taylor Lewis, Wendy Martinez, Adrian Perez Bote, Benjamin Reist, Rashi Saluja, Luca Sartore, Aaron Schroeder, Yunhan Wu, Alex Young, and other reviewers who requested anonymity. We are grateful to the rich insights reviewers provided to the authors.

References

Agency for Healthcare Research and Quality (2026). Medical Expenditure Panel Survey online at https://meps.ahrq.gov/mepsweb/. (Accessed 11 July 2026).

Belyaeva I, Carino C, Wang LC (2026). Leveraging survey metadata for LLM reasoning via knowledge graphs. *Journal of Data Science*, 24(3): 482–503. https://doi.org/10.6339/26-JDS1230

Champney TF, Qin H (2026). Maximizing linkage in address data: Spatial, exact, and fuzzy matching. *Journal of Data Science*, 24(3): 600–615. https://doi.org/10.6339/26-JDS1236

Commission on Evidence-Based Policy (2017). "The Promise of Evidence-Based Policymaking" online at https://www2.census.gov/adrm/fesac/2017-12-15/Abraham-CEP-final-report.pdf (Note: the final report and background papers prepared for the Commission are available online at https://acf.gov/opre/project/commission-evidence-based-policymaking-cep) (Accessed 11 July 2026).

Data Science for Federal Statistics Interest Group (2025). Program: Government Advances in Statistical Programming. Online at: https://statspolicy.gov/assets/fcsm/files/docs/gasp/GASP2025_Program.pdf (Accessed 11 July 2026).

Elkasabi M, Lewis T, Williams MR (2026). An estimation framework for combining probability and non-probability samples. *Journal of Data Science*, 24(3): 544–563. https://doi.org/10.6339/26-JDS1234

Federal Committee on Statistical Methodology (2026). Online at https://statspolicy.gov/FCSM/about/ (Accessed 11 July 2026).

Foundations for Evidence-Based Policymaking Act of 2018. Public Law 115 – 435. Online at https://www.govinfo.gov/app/details/PLAW-115publ435 (Accessed 11 July 2026).

Preiss AJ, Konet A, Chew R, Williams MR, Segarra EA, ..., Savitsky TD (2026). A practical guide to differentially private deep learning using the pseudo posterior mechanism. *Journal of Data Science*, 24(3): 504–522. https://doi.org/10.6339/26-JDS1237

Rhodes S, Johnson D, Sartore L, Garber S, Miller D, Abreu D (2026). Use of farm equipment machine-logged data to inform crop production statistics. *Journal of Data Science*, 24(3): 584–599. https://doi.org/10.6339/26-JDS1235

RTI International (2026). National Survey on Drug Use and Health online at https://nsduhweb.rti.org/respweb/homepage.cfm (Accessed 11 July 2026).

Saluja R, Sun H, Korkmaz G, Carle J, Hubbard R, ..., Dulaney R (2026). Predicting "Yes": Machine learning and diverse data to boost respondent cooperation. *Journal of Data Science*, 24(3): 616–629. https://doi.org/10.6339/26-JDS1242

U.S. Census Bureau (2026). American Community Survey online at https://www.census.gov/programs-surveys/acs/ (Accessed 11 July 2026).

Yin X, Xie Y, Yan J, Wang S, Liu PY, ..., Chen MH (2026). Prediction intervals and group variable importance for classification models in university enrollment yield. *Journal of Data Science*, 24(3): 523–543. https://doi.org/10.6339/26-JDS1241

Williams MR, McGuire FH, Savitsky TD (2026). Uncertainty quantification for multi-level models using the survey-weighted pseudo-posterior. *Journal of Data Science*, 24(3): 564–583. https://doi.org/10.6339/26-JDS1238