Journal of Data Science 22 (3), 353–355
July 2024
DOI: 10.6339/24-JDS223EDI
Editorial

Introduction to the GASP Special Issue☆

Lisa M. Frehill[1,*] and Peter B. Meyer[2]

[1]*Social Science Research Specialist, U.S. Department of Energy, USA*
[2]*Research Economist, U.S. Bureau of Labor Statistics, USA*

Data science is increasingly used in government, as it is everywhere. In the context of official government statistics, there is particular attention to confidentiality, credibility, transparency, and open access. There is also growing interest in blending data from various sources including surveys, administrative data, or physical sensors. Guidelines meant to ensure data quality and reliability of official statistics prescribe minimal use of outliers or small samples. While many in the government statistical community desire to use open and transparent methods, consistent with government objectives, much government-collected data are confidential and/or administrative data that necessitates a high degree of data security.

The Government Advances in Statistical Programming (GASP) workshops were initiated by the Computational Statistics for the Production of Official Statistics[1] interest group of the Federal Committee on Statistical Methodology (FCSM) in 2018. GASP's purpose was to provide members of the government statistical community a venue to share techniques and skills, give them support, and share information on ways to manage government security protocols within an increasingly open data/analysis space. There was high interest, with many participants seeking to engage in open-source sharing, at least across the federal government. Each of the five GASP events to date—it is now a conference—included about 40 speakers in regular discussion or paper panels, lightning sessions, keynotes, and workshops. Initially one-day in-person events, with the Covid-19 pandemic, a two-afternoon virtual conference emerged as the norm since 2020, expanding opportunities for more geographically dispersed participants. This broadened participation from statistical agencies outside the Washington, DC area, and, indeed, outside the United States, as well as in academia.

This special issue includes selected papers from the 2023 GASP conference, often illustrating issues of particular interest to government programmers. Many of the articles are in the JDS *Data Science in Action* category, since statistical programming is a key tool for agencies to provide services to the American public.

In the first contribution, Shrivastava and Korkmaz (2024) apply standard accounting methods to estimate the contributions of U.S. federal government agencies to open source code on github. Leveraging data from Code.gov, these authors document the increasing investment in open source software, estimating the value at $407 million in 2021. They further use network analytics to describe the contributor network to show the web of sharing across U.S. federal agencies.

Next, a series of three articles provide insights about text analysis. Government text data come from many sources such as verbatim responses on government surveys and the continually expanding volume of text-based administrative data. As such, automated text analytics represent

☆This paper represents the views of the authors and does not necessarily reflect those of the United States Government or any agency thereof.

*Corresponding author. Email: lisa.frehill@hq.doe.gov.

[1]As of 10 April 2024, the interest group name was changed to Data Science for Federal Statistics (DSFS).

an important area for federal analysts who seek efficiency, while adhering to the principles of data quality and transparency necessary to maintain credibility. Efficiently grouping text units can often be a first step in manual text analysis. Preiss et al. (2024) compare the efficacy of various machine learning methods to label a cluster of texts and show that analysts' choice of methods can be complicated by the content and context of the texts to be analyzed. The authors provide a supplemental guide about issues to consider for those who might seek to use machine learning methods within text analytics projects.

Hadley et al. (2024) examine healthcare provider community benefit trends during COVID using text data included by non-profit organizations on the Internal Revenue Service Form 990H. Deploying some of the same tools as Preiss et al. (2024), Hadley et al. (2024) show the benefits and potential drawbacks associated with deploying Generative Artificial Intelligence for these analyses. Rounding out this trio of text analytics articles, Knappenberger (2024) discusses an interactive tool that can help survey respondents match their products to the complex codes used by government analysts in the North American Product Classification System (NAPCS). Government agencies seek efficiencies and must monitor the burden placed on the public in applying for services or completing surveys such as the Economic Census, which is Knappenberger's focus. The SINCT tool provides users with the ability to quickly select an appropriate code, avoiding the need to write-in a response, thereby lowering the burden on survey respondents. Additionally, the American public benefits by having more accurate and higher quality data with less federal worker time expended.

The last Data Science in Action article by Emmet and colleagues (2024) at the U.S. Department of Agriculture's (USDA) National Agricultural Statistical Service (NASS) addresses how to identify agricultural fields with satellite sensor information in lieu of other USDA program participation records. This predicament occurs more often with farmers from groups who do not participate in certain USDA programs, notably the Amish farmers. By blending data from satellite sensors along with county assessor parcel data and other sources the authors developed methods to more accurately enumerate the nation's farming activities and fill knowledge gaps about farmers from underrepresented groups.

In the next section are two GASP 2023 contributions in the JDS Statistical Data Science category. The first article, by another NASS team led by Sartore (2024), used a distribution free method and fuzzy logic approach to identify outliers, which can signal potential data errors. Within the federal agricultural data context, rapid identification of anomalous entries is critical to developing accurate and timely official statistics. The final article by Chen and Xu (2024) examines another prevalent survey data quality issue: missing data. While there are a host of approaches to imputation of missing values, Chen and Xu (2024) demonstrate that statisticians need to be mindful of the ways that popular machine learning approaches embody underlying assumptions that may not be readily apparent. Their novel approach to Predictive Mean Matching balances robustness and generating reliable imputed values.

This volume would not have been possible without the work of many specialized reviewers, including Nathan Cruze, Kelsey Gray, Travis Hoppe, Ben Klemens, Wendy Martinez, Nipa Phojanamongkolkij, Paul Reimer, Benjamin Rogers, José Bayoán Santiago-Calderón, Yumiko Siegfried, Lupe Villatoro, and Matt Williams, and other reviewers who requested anonymity. We are grateful to the rich insights reviewers provided to the authors. We are also grateful to Jun Yan for giving us an opportunity to share GASP 2023 articles with a wider audience. The prerogatives under which statisticians operate at federal agencies to deliver high quality, timely, and useful data products to the American public provide a ripe proving ground for data science approaches. The GASP conference in its recent virtual format has provided an important forum

in which data scientists from academia, the private sector, and government can share techniques and knowledge to contribute to the growth of open data and better meet the needs of the American public.

July 2024

References

Chen S, Xu C (2024). Predictive mean matching imputation procedure based on machine learning models for complex survey data. *Journal of Data Science*, 22(3): 456–468. https://doi.org/10.6339/24-JDS1135

Emmet RL, Hunt K, Jennings R, Daniel K, Abreu DA (2024). Evaluating a method for georeferencing agricultural fields. *Journal of Data Science*, 22(3): 423–435. https://doi.org/10.6339/24-JDS1146

Hadley E, Marcial L, Quattrone W, Bobashev G (2024). Traditional and GenAI text analysis of COVID-19 pandemic trends in hospital community benefits IRS documentation. *Journal of Data Science*, 22(3): 393–408. https://doi.org/10.6339/24-JDS1144

Knappenberger C (2024). Bringing search to the economic census – the NAPCS classification tool. *Journal of Data Science*, 22(3): 409–422. https://doi.org/10.6339/24-JDS1147

Preiss AJ, Arbeit CA, Berghammer A, Bollenbacher J, McCarthy JV, Brom MG, et al. (2024). Evaluation of text cluster naming with generative large language models. *Journal of Data Science*, 22(3): 376–392. https://doi.org/10.6339/24-JDS1149

Sartore L, Chen L, van Wart J, Dau A, Bejleri V (2024). Identifying anomalous data entries in repeated surveys. *Journal of Data Science*, 22(3): 436–455. https://doi.org/10.6339/24-JDS1136

Shrivastava R, Korkmaz G (2024). Measuring public open-source software in the federal government: An analysis of Code.gov. *Journal of Data Science*, 22(3): 356–375. https://doi.org/10.6339/24-JDS1148