Journal of Data Science 20 (3), 277–278
July 2022

DOI: 10.6339/22-JDS203EDI
Editorial

Editorial: Data Science Meets Social Sciences

Elena A. Erosheva[1], Shahryar Minhas[2], Gongjun Xu[3], and Ran Xu[4,*]

[1]*Department of Statistics, School of Social Work, and the Center for Statistics and the Social Sciences, University of Washington, Seattle, WA 98195, USA*
[2]*Department of Political Science, Michigan State University, East Lansing, MI 48823, USA*
[3]*Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA*
[4]*Department of Allied Health Sciences, University of Connecticut, Storrs, CT 06269, USA*

This special issue features eight articles on "Data Science Meets Social Sciences." Data science is playing an increasingly important role in the social sciences through the use of a wide variety of data – structured and unstructured, quantitative and qualitative – to facilitate our understanding of human and society. The social sciences, historically rich in theories and contextual knowledge, are well positioned to guide development and translation of data products into meaningful decisions and practices. This special issue is dedicated to highlighting scientific approaches at the conjunction of the social sciences and data science. It covers a wide range of topics including applications of current state-of-the-art data science methods as well as development of new data science approaches in education, psychology, political science, economy, public health and other social sciences.

All articles in this special issue were peer-reviewed. Anglin et al. (2022) compared two different methods of text annotation: a "complex" method where all codes were annotated for all pieces of text in context and a "simple" method in which codes were annotated one at a time for individual pieces of text, and found that the complex annotation scheme was more accurate and efficient. Social scientists and computational linguists that are tasked with making annotation decisions will find this paper's contribution to the data science of annotation of text data especially appealing. Bluhm and Cutura (2022) demonstrated how to handle big data with Apache Spark, an open-source distributed computing toolset, using reproducible econometrics examples. The paper targets explicitly a non-expert audience, making it a valuable reference for researchers with a limited background in data handling and distributed computing. Lehoucq (2022) argued that optimism among American liberals and conservatives about technology, specifically, the fairness of predictive automation, is becoming increasingly polarized. The author relied on nationally-representative survey data from the Pew Research Center's American Trends Panel and utilized a variety of machine learning techniques to understand difference in American perceptions about predictive automation. The author showed that Americans who think predictive automation is fair tend to lean conservative, accept more controversial social media practices, and have a positive view of technology corporations. Ma and Xu (2021) discussed and compared the methods to test latent hierarchical structures in Cognitive Diagnosis Models (CDMs), a model that has been widely used in social and biological sciences. They demonstrated the effectiveness and superiority of a parametric bootstrap method in testing latent hierarchical structures in CDMs using comprehensive simulations and an educational assessment dataset. Mo et al. (2022) explored the application of classification and regression trees (CRT) in educational research, particularly in studying students' performance levels and achievement gains. Through a case study of Early Childhood Longitudinal Study-Kindergarten 2011 (ECLS-K: 2011) data, the advantages and

[*] Corresponding author. Email: ran.2.xu@uconn.edu.

limitations of using CRT on achievement data are demonstrated, which provides a practical illustration of scenarios when CRT is appropriate and beneficial. Sanders et al. (2022) empirically investigated the relationship between grit and age. Using a within and between generational cohort age difference-in-difference approach, they found a negative-parabolic relationship between grit and age, which was driven by generational variation and not by age variation. Yang and Bradley (2021) developed a more efficient implementation of the Conway-Maxwell (COM) Poisson model, which is unique in its ability to handle under- and over-dispersion. The authors made a number of novel improvements over the model to deal computational issues in the estimation process and impose a conditional independence assumption to avoid inflating variance of the data with spatial random effects. They highlighted the utility of their approach using simulated examples and a real world application to voting data from the 2016 US presidential election. Yang and Shang (2022) investigated different disaggregation structures in grouped functional time series and their implications for forecasting. Using Japanese sub-national age-specific mortality rates from 1975 to 2016 as an example, they found that the dynamic multivariate functional time series method, combined with reconciliation methods, obtained improved point and interval forecasts when applied to the functional time series formed by disaggregated series.

We are grateful to the authors for their timely contributions and to the anonymous reviewers for their thoughtful reviews of the manuscripts. It is our hope that this special issue will be of interest to social scientists and data scientists alike.

References

Anglin K, Boguslav A, Hall T (2022). Improving the science of annotation for natural language processing. *Journal of Data Science*, 20(3): 339–357. https://doi.org/10.6339/22-JDS1054.

Bluhm B, Cutura J (2022). Econometrics at scale: Spark up big data in economics. *Journal of Data Science*. 20(3): 413–436. https://doi.org/10.6339/22-JDS1035.

Lehoucq E (2022). Do Americans think the digital economy is fair? Using supervised learning to explore evaluations of predictive automation. *Journal of Data Science*, 20(3): 381–399. https://doi.org/10.6339/22-JDS1053.

Ma C, Xu G (2021). Hypothesis testing for hierarchical structures in cognitive diagnosis models. *Journal of Data Science*, 20(3): 279–302. https://doi.org/10.6339/21-JDS1024.

Mo Y, Habing B, Sedransk N (2022). Tree-based methods in educational research: A tool for modeling nonlinear complex relationships and generating new insights from data. *Journal of Data Science*, 20(3): 359–379. https://doi.org/10.6339/22-JDS1056.

Sanders S, Gedara NIM, Walia B, Boudreaux C, Silverstein M (2022). Does aging make us grittier? Disentangling the age and generation effect on passion and perseverance. *Journal of Data Science*, 20(3): 401–411. https://doi.org/10.6339/22-JDS1041.

Yang HC, Bradley JR (2021). Bayesian inference for spatial count data that may be over-dispersed or under-dispersed with application to the 2016 US presidential election. *Journal of Data Science*, 20(3): 325–337. https://doi.org/10.6339/21-JDS1032.

Yang Y, Shang HL (2022). Is the group structure important in grouped functional time series? *Journal of Data Science*. 20(3): 303–324. https://doi.org/10.6339/21-JDS1031.