Jian Cao

Department of Political Science
Trinity College Dublin
2 College Green
Dublin, Ireland

https://www.jiancao.net
+353 (89) 944 1053
caoj@tcd.ie
https://github.com/jian-frank-cao

Academic Background

Trinity College Dublin
Research Fellow — *January 2022 –Present*

California Institute of Technology
Visitor — *January 2022 –Present*
Postdoctoral Scholar in Data Science and Election Integrity — *July 2019 –December 2021*

Florida State University
Senior Researcher — *August 2018 –June 2019*
Ph.D. Economics — *August 2018*
Dissertation: Multiple Imputation Methods for Large Multi-Scale Data Set with Missing or Suppressed Values
M.S. Economics — *August 2016*

China University of Mining & Technology
Master of Financial Engineering — *June 2014*

Henan University of Economics & Law
B.A. Economics — *June 2010*

Research Interests

Econometrics, political methodology, computational social sciences
— Comparative studies, multiple imputation methods, social media analyses

Peer-Reviewed Articles

Li, Zhuofang, **Jian Cao**, Nicholas Adams-Cohen, and R. Michael Alvarez. 2023. “The Effect of Misinformation Intervention: Evidence from Trump’s Tweets and the 2020 Election.” *Multidisciplinary International Symposium on Disinformation in Open Online Media 2023 Proceedings*. doi: 10.1007/978-3-031-47896-3_7.

Cao, Jian, Seo-young Silvia Kim, and R. Michael Alvarez. 2022. “Bayesian Analysis of State Voter Registration Database Integrity.” *Statistics, Politics and Policy*. doi: 10.1515/spp-2021-0016.

Cao, Jian, Christina M. Ramirez, and R. Michael Alvarez. 2021. “The politics of vaccine hesitancy in the United States.” *Social Science Quarterly*. doi: 10.1111/ssqu.13106.

Alvarez, R. Michael, **Jian Cao**, and Yimeng Li. 2021. “Voting Experiences, Perceptions of Fraud, and Voter Confidence.” *Social Science Quarterly*. doi: 10.1111/ssqu.12940.

Srikanth, Maya, Anqi Liu, Nicholas Adams-Cohen, **Jian Cao**, R. Michael Alvarez, and Anima Anandkumar. 2021. "Dynamic Social Media Monitoring for Fast-Evolving Online Discussions." *Knowledge Discovery and Data Mining 2021 Proceedings*. doi: 10.1145/3447548.3467171.

Cao, Jian, Nicholas Adams-Cohen, and R. Michael Alvarez. 2021. "Reliable and Efficient Long-Term Social Media Monitoring." *Journal of Computer and Communications*. doi: 10.4236/jcc.2021.910006.

Working Papers

"Multiple Imputation for Large Multi-Scale Data With Linear Constraints."
(with Paul Beaumont) (Job Market Paper)

"Dynamic Synthetic Controls: Accounting for Varying Speeds in Comparative Case Studies."
(with Thomas Chadefaux) (Under Review)

"Ballot Rejections and Ballot Curing in Washington State."
(with Canyon Foot, Jay Lee, R. Michael Alvarez, Paul Manson, and Paul Gronke)

Work in Progress

"Clustering Historical Matrices of Dependent and Independent Variables as Unobserved Effects in Dynamic Panel Data Modeling."
(with Thomas Chadefaux)

"Enhancing Regression Analysis through Self-Aligned DTW-Derived Speed Profiles in Time Series Data."
(with Thomas Chadefaux)

"The Parallel Quasi-Monte Carlo Bayesian Multi-Scale Multiple Imputation Method."
(with Paul Beaumont)

"Mailing It In: Voter Confidence in Vote-By-Mail In the 2020 Presidential Election."
(with R. Michael Alvarez and Seo-young Silvia Kim)

Research Experience

Trinity College Dublin

Research Fellow *January 2022 –Present*
Project: Patterns of Conflict Emergence

Identify patterns in the pre-conflict actions using data on conflict events and in their perceptions using data from financial markets, news articles, and diplomatic documents.

Evaluate the utility of these patterns to improve forecasts of conflict with both historical and live out-of-sample predictions.

Summarize the core features of dangerous patterns into motifs that can help build new theories of conflict emergence and escalation.

CALIFORNIA INSTITUTE OF TECHNOLOGY

Visitor *January 2022 –Present*
Postdoctoral Scholar in Data Science and Election Integrity *July 2019 –December 2021*
Project: Election Auditing

Developed probabilistic matching and Bayesian multivariate models using GCP for large election database auditing in California and Florida.

Implemented entity resolution and anomaly detection on daily snapshots of voter registration databases that contain more than 20 million records and detected 10× more true anomalies than the existing methods did.

Project: Twitter Monitoring

Developed serverless architectures using GCP, AWS, and Oracle for long-term Twitter monitoring. They ingest, process, and store more than 4.5 billion tweets (30 TB in size) related to COVID-19, primary/general elections, and protests.

Work closely with the Computer Science team and implemented topic, spatial, network, and sentiment analyses on the collected tweets and identified COVID-19 misinformation and voting issues in the 2020 Election cycle.

FLORIDA STATE UNIVERSITY

Senior Researcher *August 2018 –June 2019*
Project: Large Missing Data Multiple Imputation

Developed the fastest and most accurate Bayesian inference method for missing data multiple imputation.

Developed a parallel-sequential imputation method that can impute large multi-scale data sets with 1.5 billion observations (500 GB in size).

Project: Economic Impact Modeling

Analyzed the economic impact of Florida's housing and small business policies.

Developed a NETS-based impact analysis tool that has 1000 times finer resolution than the existing methods.

Teaching

Caltech	Post-doc lecturer, *SS 224*: Social Science Data (2019, 2021).
Florida State University	Teaching assistant, *ECO 3431*: Analysis of Economic Data (2017).
Henan University of Economics & Law	Teaching assistant, Statistics (2008).

Conference and Seminar Presentations

2023 APSA 2023 (American Political Science Association Annual Conference); MISDOOM 2023 (Multi-disciplinary International Symposium on Disinformation in Open Online Media Annual Conference).

2021 RRoCCET 2021 (Research Running on Cloud Compute & Emerging Technologies); KDD 2021 (Knowledge Discovery and Data Mining).

2020 ESRA 2020 (Election Sciences, Reform, & Administration Conference); VoteCal (Vote California); Caltech COVID Dynamic Talk.

2019 Caltech SISL Talks (Social and Information Sciences Laboratory); FSU GSMC Talks (Graduate Modern Statistics Club).

Software

`dsc` A R package for the Dynamic Synthetic Control method, an enhanced synthetic control method that accounts for misalignment in time series caused by inherent speed differences.

`spike` A Python package for deploying long-term Twitter monitors and managing files on Google Drive, Google Cloud Storage, and SFTP file systems.

References

Paul Beaumont (primary advisor)
Associate Professor of Economics
Florida State University
beaumont@fsu.edu

Thomas Chadefaux
Professor in Political Science
Trinity College Dublin
thomas.chadefaux@tcd.ie

R. Michael Alvarez
Professor of Political and
Computational Social Science
California Institute of Technology
rma@hss.caltech.edu