Statistical Forecasting using Political Event Data

Philip Schrodt
Political Science and
International Center for the Study of Terrorism
Pennsylvania State University
schrodt@psu.edu
http://eventdata.psu.edu

Development of Political Event Data

- 1965-1980:
  Initial development under DARPA funding; human coding of newspapers; WEIS ontology
- 1990-2005:
  NSF and proprietary development of automated coding systems using newswire reports; CAMEO and IDEA ontologies
- 2007-2010:
  DARPA Integrated Conflict Early Warning System (ICEWS); 5 indicators of political instability for 29 countries in Asia; 80%+ forecasting accuracy at 6 month horizon; fully-automated models

ICEWS Developments

- Successfully scaled by two orders of magnitude
- Parallel processing allows unlimited speed
  - 8 million sentences in ten minutes
- Integration with open-source natural language processing software
  - JABARI-NLP (Lockheed-Martin)
- Refinement of sub-state actor ontologies
- Development of extensive dictionaries and systematic named-entity recognition software
- Some work with geolocation, machine translation, real-time updating and new social media

Event Data in 2011

- With the combination of the web and automated coding methods, we now have instruments that can automatically monitor the state of the political world in near real time.
- We are only beginning to develop the appropriate ways of systematically using this information

New social media

- The good
  - Widely available to elites
  - More or less uncensored
  - Should provide early information on changing sentiment prior to observing actual collective action
- The bad
  - No filters and mostly politically irrelevant: "`Wanna getta pizza? ;)`"
  - Easily manipulated by anyone—business, government, NGOs—who wants to go to the trouble of doing so
- The ugly
  - No standardization of content

Statistical challenges

- Rare events
  - Incorporate much longer historical time lines?—Schelling used Caesar's *Gallic Wars* to analyze nuclear deterrence
  - Calibration can be very tricky
- Analysis of event sequences, which are not a standard data type
- Causality
  - *Oxford Handbook of Causation* is 800 pages long
- Integration of qualitative and SME information
  - Bayesian approaches are promising but to date they have not really been used

Statistical challenges

- Systematically dealing with measurement error and missing values rather than assuming “missing at random”
- Correctly leveraging ensemble methods which utilize multiple statistical and computational pattern recognition methods
  - PITF forecasting tournament; Bayesian model averaging
  - There are known and irreducible random elements in political behavior
- Upshot: you can't simply specify a desired rate of accuracy and assume by throwing sufficient funding at the problem you will get there.
  - PITF and many other models all converge to about 80%

Sources of error

- Specification error: no model of a complex, open system can contain all of the relevant variables;
- Measurement error: with very few exceptions, variables will contain some measurement error
  - presupposing there is even agreement on what the “correct” measurement is in an ideal setting;
  - Predictive accuracy is limited by the square root of measurement error: if your reliability is 80%, your accuracy can't be more than 90%
- Free will
  - Rule-of-thumb from our rat-running colleagues:“A genetically standardized experimental animal, subjected to carefully controlled stimuli in a laboratory setting, will do whatever it wants.”
- Quasi-random structural error: Complex and chaotic deterministic systems behave as if they were random under at least some parameter combinations

Challenges in integrating models into decision-making

- Forecasting is hard (Tetlock)
- Probabilistic reasoning is hard (Kahneman)
- Statistics is new compared to deterministic modeling and is still changing, even at very fundamental levels
  - Frequentist vs Bayesian approaches
  - New approaches made possible by computational advances

Earliest date of a textbook that could be used in a contemporary class

- Geometry: 300 BCE
- Algebra: 1750
- Calculus and differential equations: 1820
- Statistics: ???
  - Current social science statistics books are mostly frequentist (hypothesis testing using significance levels) whereas most statistics departments are now Bayesian (estimating probability distributions of coefficients)
  - The topics in the introductory curriculum vary little from pre-computer times

Challenges in integrating models into decision-making

- Forecasting is hard (Tetlock)
- Probabilistic reasoning is hard (Kahneman)
- Statistics is new compared to deterministic modeling and is still changing, even at very fundamental levels
  - Frequentist vs Bayesian approaches
  - New approaches made possible by computational advances
- Government seems to be locked-in to a set of 20 to 40 years old approaches
  - The answers aren't simple, even if some colonel wants them to be simple
  - Our 20th century peer competitors were trained as ideologues; our 21st century peer competitors are trained as engineers.