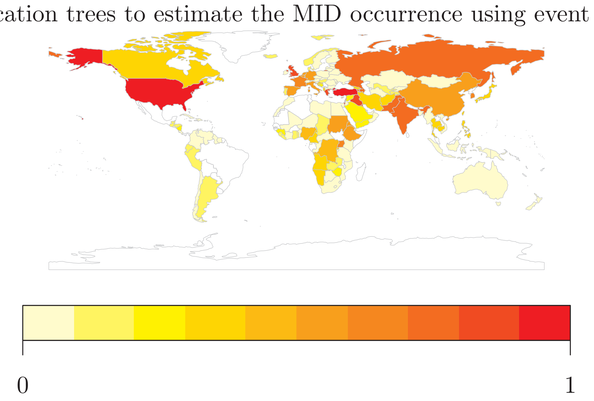
Automated Production of Political Indicators

Matthew C. Dickenson
mcd31@duke.edu

Duke UNIVERSITY

Motivation

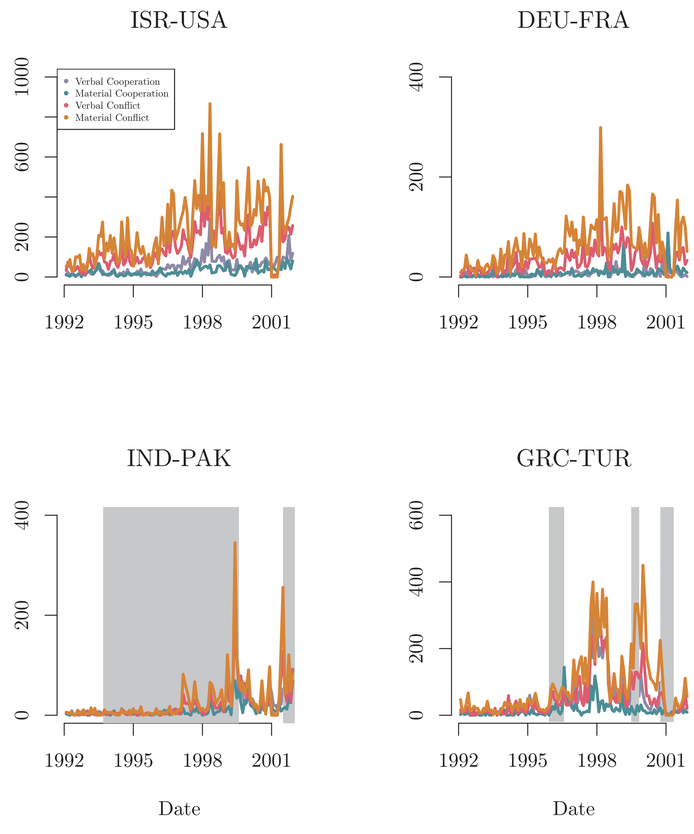
Can political indicators such as the Militarized Interstate Dispute (MID) dataset be approximated by automated classification procedures? Existing, human-intensive pipelines for the production of MID and related data (e.g. Polity and Freedom House) are costly and slow. This project uses classification trees to estimate the MID occurrence using event data.



Proportion of 1992-2001 Spent in MIDs

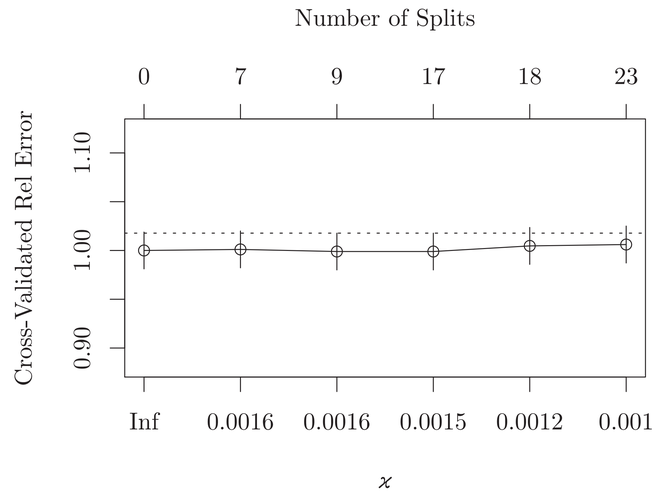
Data Source

Features were drawn from the Global Database of Events, Language, and Tone (GDELT), which classifies daily interactions as material or verbal, and cooperative or conflictual. Interactions between states were aggregated to the monthly level and first-differenced to account for the exponential growth in the number of records over time. Examples of dyadic time series are shown below, with shaded regions indicating the occurrence of a MID.

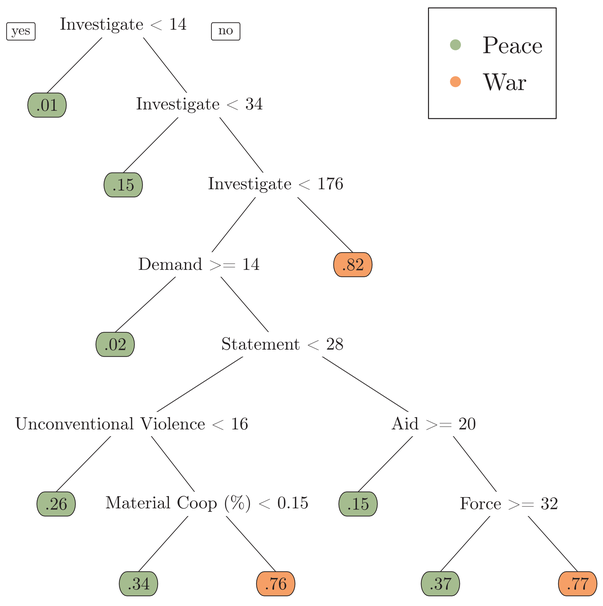


Classification Trees

A tree was fit to classify MIDs in dyad-months using changes in counts of each event types and the proportion of interactions in each quadrant of the material/verbal, cooperative/conflictual matrix. To fit the tree, the minimum complexity parameter was set to $\varkappa$=0.0001. By cross-validation, the optimal $\varkappa$ was found to be 0.00155, which also minimizes error on the test data.



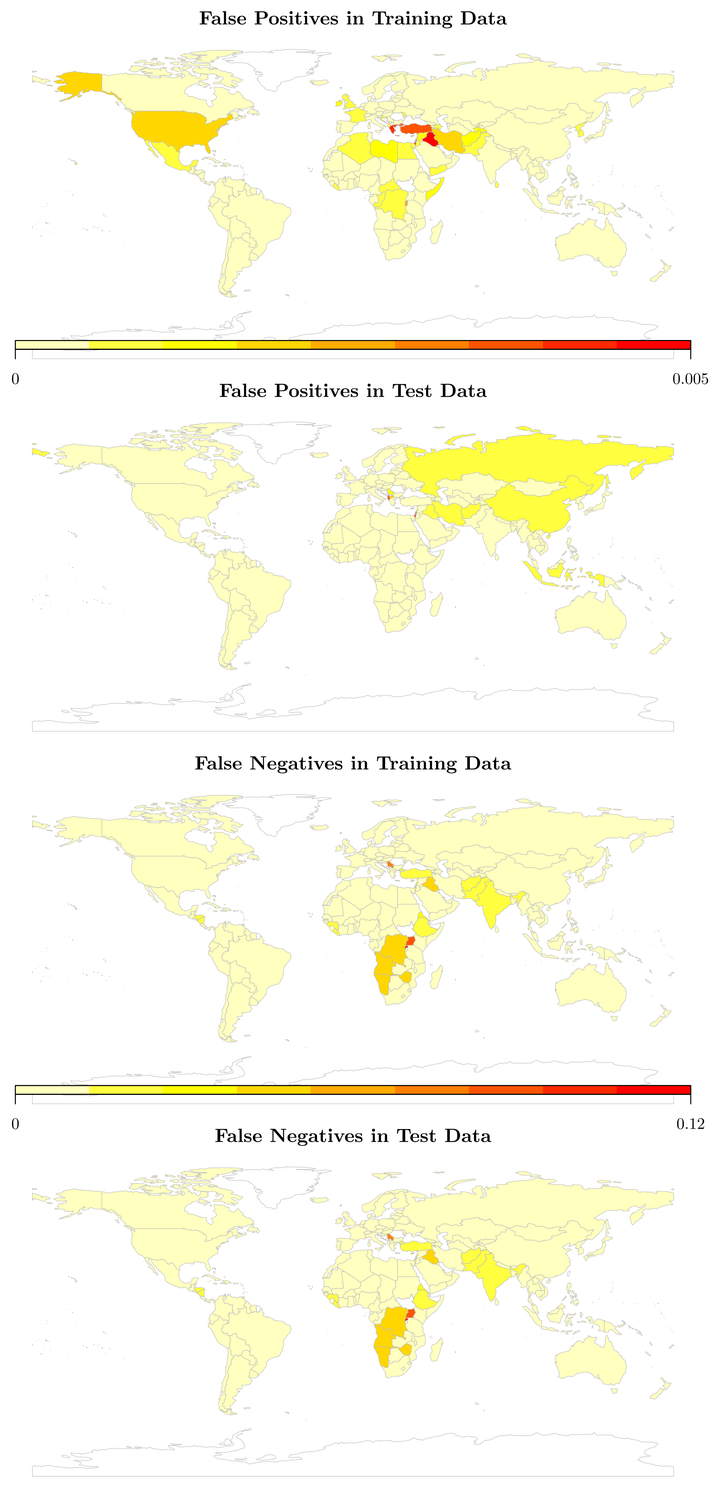
This value was used to prune the tree shown below. The leaves of the tree are shaded according to whether war (MID hostilities $\geq$ 4) or peace is more likely, and the value at each leaf indicates the relative frequency of war in the training set.



The table below presents several measures of accuracy for the classification tree compared to a null model (all predicted values are zero) and a logistic regression model using the same features. The test set covers 1992-1998 (n=372,271) and the training set covers 1999-2001 (n=213,218). The classification tree outperforms the other two models in all respects for the training data, and in most respects for the test data.

	Training Data			Test Data		
Model	MSE	Precision	Recall	MSE	Precision	Recall
Null	0.0075	0.000	0.000	0.0066	0.000	0.000
GLM	0.0082	0.158	0.022	0.0079	0.142	0.038
CART	0.0067	0.702	0.192	0.0066	0.422	0.027

Results



Outlook

- CART handles interactions between features better than GLM.
- Next steps include SMOTE for rare events, using spatial and network lags to account for interdependence between countries, and a comparison to random forest.
- More advanced methods and a final pass by human coders will likely be required to obtain fully accurate classifications.
- Automated methods can offer a quick and inexpensive first approximation of political indicators.