JACOBS UNIVERSITY

Predicting military conflicts by data-driven techniques

Adalbert F.X. Wilhelm,

Herbsttagung AG DANK, WIAS Berlin
November 19th, 2016

Outline

1. Military conflicts: data structures and data projects
2. Logistic regression
3. Requirements for “accepted” data-driven classifiers
4. Some empirical evaluations
5. Class imbalance
6. Conclusion

Modeling armed conflict

- one of the major topics in international relations
- events are of great importance
- modeling onset, duration, or termination

- Definition:
  - Armed Conflict: An armed conflict is a contested incompatibility that concerns government and/or territory where the use of armed force between two parties, of which at least one is the government of a state, results in at least 25 battle-related deaths in one calendar year.
  - "Armed conflict" is also referred to as "state-based conflict", as opposed to "non-state conflict", where none of the warring parties are a government.
  - War = Armed conflict with at least 1000 battle related deaths

- Data typically in dyads (country-year) or triads (country-country-year)

- modeling
  - Onset
  - duration
  - termination of armed conflicts

Some data projects for conflict studies

- Correlates of War project
  - data from 1816 - 2010
  - military conflicts
    - between or among non-state entities (non-state war),
    - between states (inter-state war),
    - and within states (intra-state war).
  - Militarized Interstate Disputes
    - all instances of when one state threatened, displayed, or used force against another.

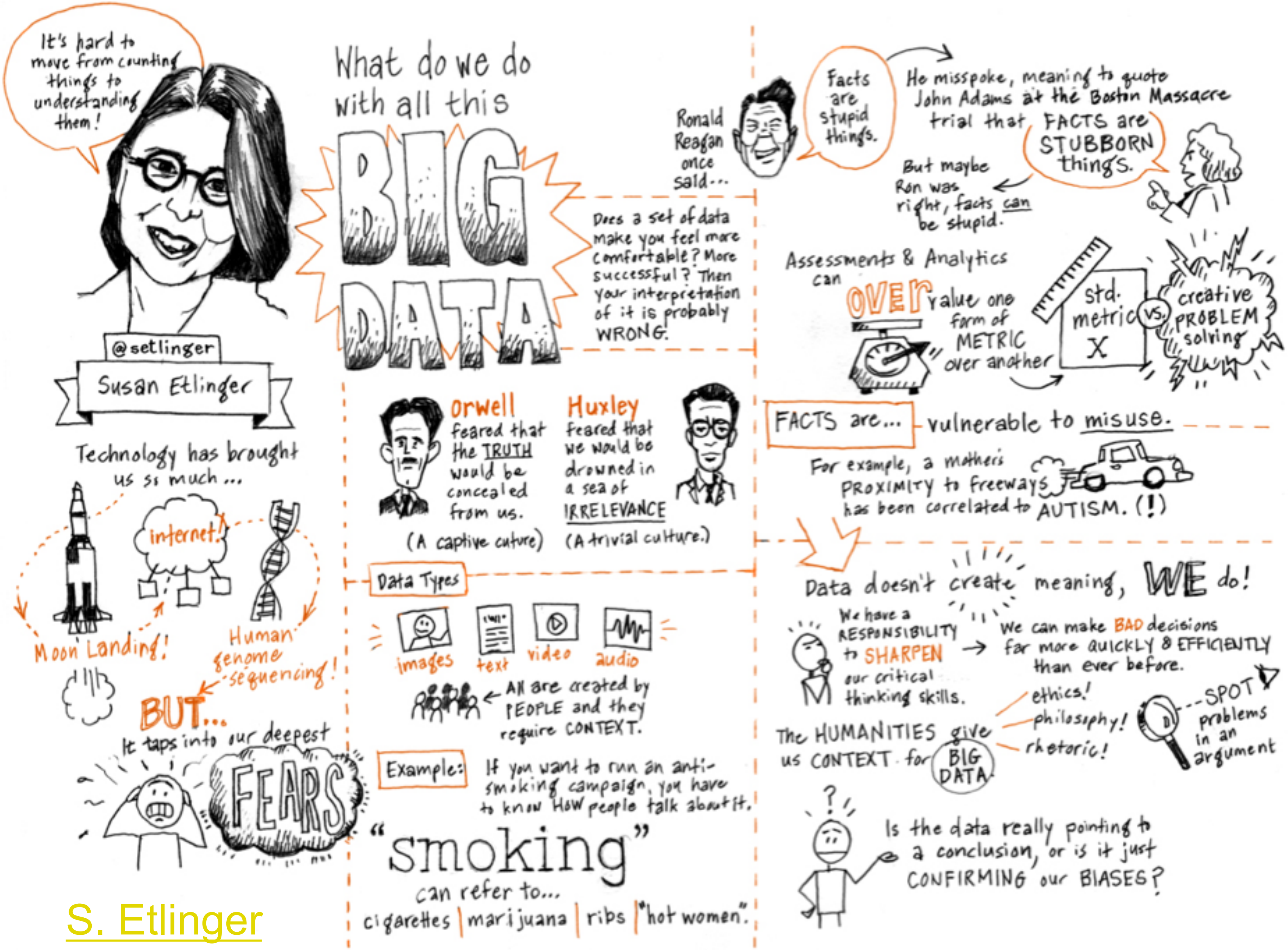
- Uppsala Conflict Data Program (UCDP/Prio)
  - a conflict-year dataset with information on armed conflict where at least one party is the government of a state in the time period 1946-2013.
  - comprises 2134 conflicts
  - involving 116 states
  - involving 547 opponents
  - covering 68 years

Some data projects for conflict studies

- KOSIMO
  - a conflict-year dataset with information on violent and non-violent conflicts where at least one actor is nation-state in the time period 1945-1999.
  - comprises 301 conflicts and 693 conflict episodes
  - involving 171 states
  - every conflict described by 28 variables

- ICB International Conflict Behavior
  - Four data sets covering the period from 1917 to 2001
  - Different units of analysis: nation-state, international system, nation-dyads, one-sided conflicts

What do we do with all this data?



S. Etlinger

JACOBS UNIVERSITY

Standard approaches for modeling occurrence of events in the Social Sciences

- Logit model
  - Dichotomous response
  - A set of predictors (continuous and categorical)
  - Model formulation on the linear predictor level using the link function

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

  - Model formulation on the response level using the inverse link function

$$P(Y=1) = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)}$$

JACOBS UNIVERSITY

Standard approaches for modeling occurrence of events in the Social Sciences

- ❑ Which predictors are significant?

- ❑ Focus on specific predictors: are they complementary or is one of them redundant?

- ❑ Sequential model comparison

- ❑ Quality of models?
- ❑ Prediction?

- ❑ Logistic regression misses out in predicting conflict cases!

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.355607	1.157507	-0.307	0.75868	
ARMSTRAN	0.302439	0.151659	1.994	0.04613	*
MILSPEND	0.766993	0.421489	1.820	0.06880	.
COLONIALFrance	0.182830	0.292283	0.626	0.53163	
COLONIALBritain	-0.445435	0.310336	-1.435	0.15119	
CUMWAR	1.591943	0.142832	11.146	< 2e-16	***
DEVELOP	-1.454921	0.529971	-2.745	0.00605	**
ETHNOPOL	0.928507	0.282423	3.288	0.00101	**
REPRESSI	0.004485	0.030116	0.149	0.88160	
SEMIDEMsemi democracy	1.408602	0.476337	2.957	0.00310	**
TRANSITItransition	0.526178	0.508101	1.036	0.30040	
TRANSITItransit semi-dem	2.654645	0.929470	2.856	0.00429	**

JACOBS UNIVERSITY

The Role of Prediction in the Social Sciences

- Prediction is a contentious issue in the Social Sciences
  - focus on estimation of causal parameters
    - Priority is given to identifying causal effects (Beck et al. 2000; Ward et al. 2010)
    - Refinement of established models to evaluate additional/alternative causal mechanisms
  - model fit often neglected
  - P-value overuse (->ASA Statement on statistical significance and p-values, 2016)
  - Growing literature on model evaluation and comparison (Goldstone et al. 2010; Ward et al. 2012; Hegre et al. 2013; Schrodt et al. 2013)
  - Growing literature on predicting occurrence of events
    - civil war (Hegre et al. 2013; Shellman et al. 2013; Brandt et al. 2014; Clayton and Gleditsch 2014)
    - interstate disputes (Gleditsch and Ward 2012),
    - political instability (Goldstone et al. 2010)

Requirements for “accepted” data-driven classifiers in the social sciences

- improved prediction accuracy
- explanatory capability
- adaptability to class-imbalanced data

- ideally, allowing discussion of “causal effects”

The Single Model Philosophy

Motivation: Occam's Razor

- "one should not increase, beyond what is necessary, the number of entities required to explain anything"
- Infinitely many models can explain any given dataset
- Might as well pick the smallest one…

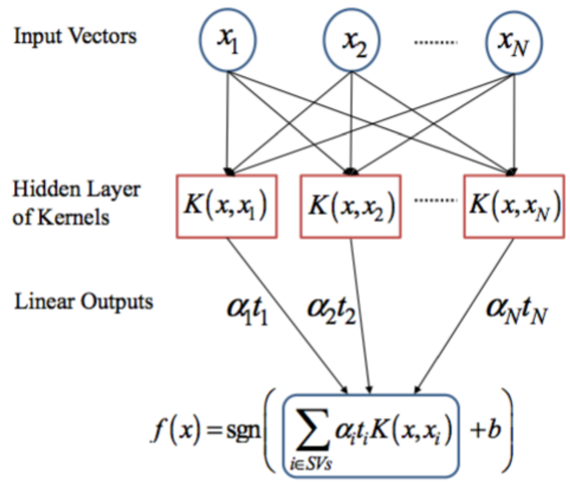
Ensemble Philosophy

Build many models and combine them
Only through averaging do we get at the truth!
It's too hard (*impossible*?) to build a single model that works best
Two types of approaches:

- Models that don't use randomness
- Models that incorporate randomness

Support Vector Machines

Choi, Pattipati & Asal (2008): A Data-driven Classification Framework for Conflict and Instability Analysis. *IEEE International Conference on Systems, Man and Cybernetics (SMC 2008)*



- Predicting stability level of a state (three levels)
- KOSIMO data base: consists of eleven macro-structural indicators (factors, attributes, features) for 171 countries over the period 1975-1999.

- Comparison against multinomial logit and unrestricted fuzzy analysis of statistical evidence (UnFASE)

	Multinomial Logit	UnFASE	Proposed Approach
Average Overall	79%	79%	**94**%
Average Recall	69%	75%	**91**%
Average Precision	62%	66%	**90**%

JACOBS UNIVERSITY

Single model philosophy

Data: Occurrence of military conflicts in sub-Saharan Africa (Craft & Smaldone, 2002)

Different splits into training and test data

Variables Description of the variables.

- **Warinvol** war involvement binary variable, from Gleditsch et.al.
- **Year** year, 1967 through 1997
- **Colonial** colonial indicator, from Blanton et. al.
- **Country** country name
- **Transiti** transition binary, from Polity IV
- **Ethnopol** ethno-political groups indicator, from Minorities at Risk
- **Repressi** repression indicator, from Polity IV
- **Semidem** semi-democracy indicator, from Polity IV
- **Armstran** arms imports, from WMEAT (log values)
- **Milspend** per capita military spending, from WMEAT (log values)
- **Develop** per capita GNP, from WMEAT/World Bank (log values)
- **Cumwar** 5-year moving average of war magnitude, from Gleditsh et.al.

	CART	Logistic Regression	Naïve Bayes Classifier	Linear Discriminant Analysis
1	0.8898	0.8976	0.8819	0.8898
2	0.8701	0.9094	0.9016	0.9213
3	0.9016	0.874	0.8661	0.8937
4	0.874	0.878	0.8543	0.8819
5	0.9016	0.8976	0.874	0.8976
6	0.9134	0.9055	0.8898	0.9134
7	0.8937	0.8898	0.8701	0.8819
8	0.8819	0.8661	0.8346	0.8661
9	0.9409	0.9173	0.9055	0.9173
10	0.9016	0.8937	0.874	0.9094
AVG	0.8963	0.8928	0.8753	0.8959

Koridze & W. (2015)

Ensemble Approaches

Bagging

- **B**ootstrap aggregating

Boosting

Random Forests

- Bagging reborn

- Well-established

Bagging

Main Assumption:

- Combining many unstable predictors to produce a ensemble (stable) predictor.
- Unstable Predictor: small changes in training data produce large changes in the model.
  - e.g. Neural Nets, trees
  - Stable: SVM (sometimes), Nearest Neighbor.

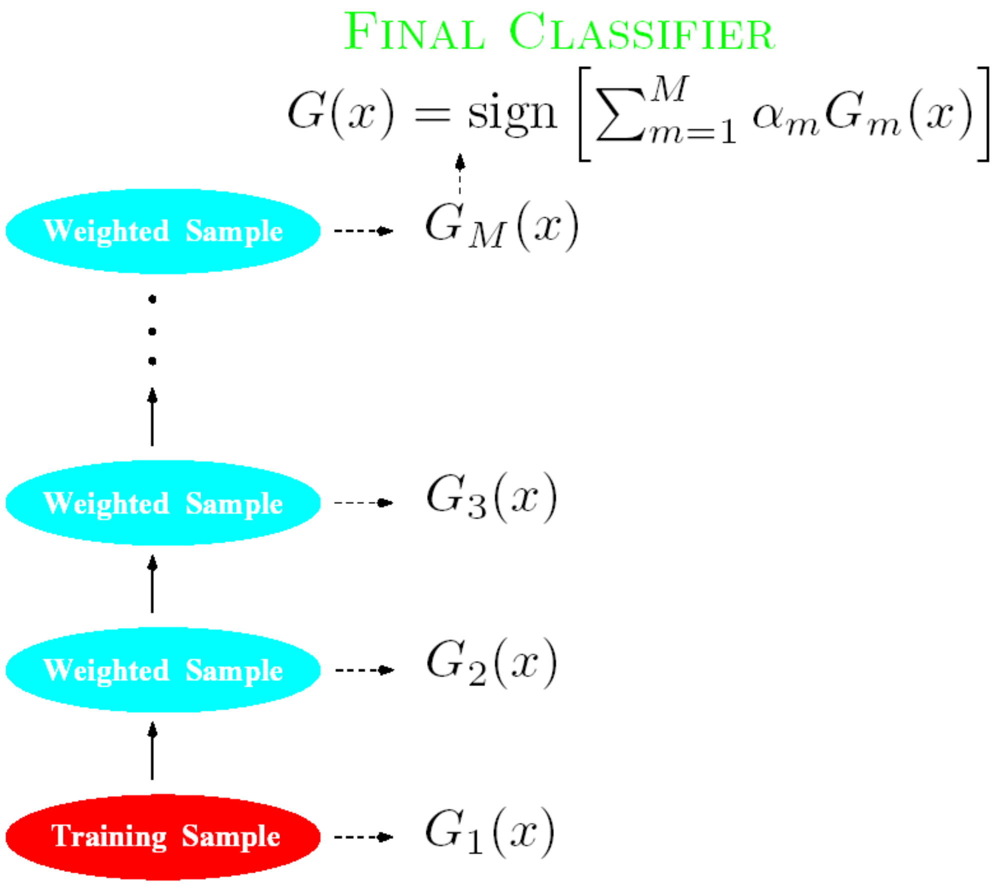
Hypothesis Space

- Variable size (nonparametric):
  - Can model any function if you use an appropriate predictor (e.g. trees)

Boosting

- Originally developed by computational learning theorists to guarantee performance improvements on fitting training data for a ***weak learner*** that only needs to generate a hypothesis with a training accuracy greater than 0.5 (Schapire, 1990).
- Revised to be a practical algorithm, AdaBoost, for building ensembles that empirically improves generalization performance (Freund & Shapire, 1996).
- Key Insights:
  - Instead of sampling (as in bagging) re-weigh examples!
  - Examples are given weights. At each iteration, a new hypothesis is learned (weak learner) and the examples are reweighted to focus the system on examples that the most recently learned classifier got wrong.
  - Final classification based on weighted vote of weak classifiers

Boosting



Each classifier $G_m(\mathbf{x})$ is trained from a weighted sample of the training data

Each predictor is created by using a biased sample of the training data

- Instances (training examples) with high error are weighted higher than those with lower error

Difficult instances get more attention

- This is the motivation behind boosting

Random Forest

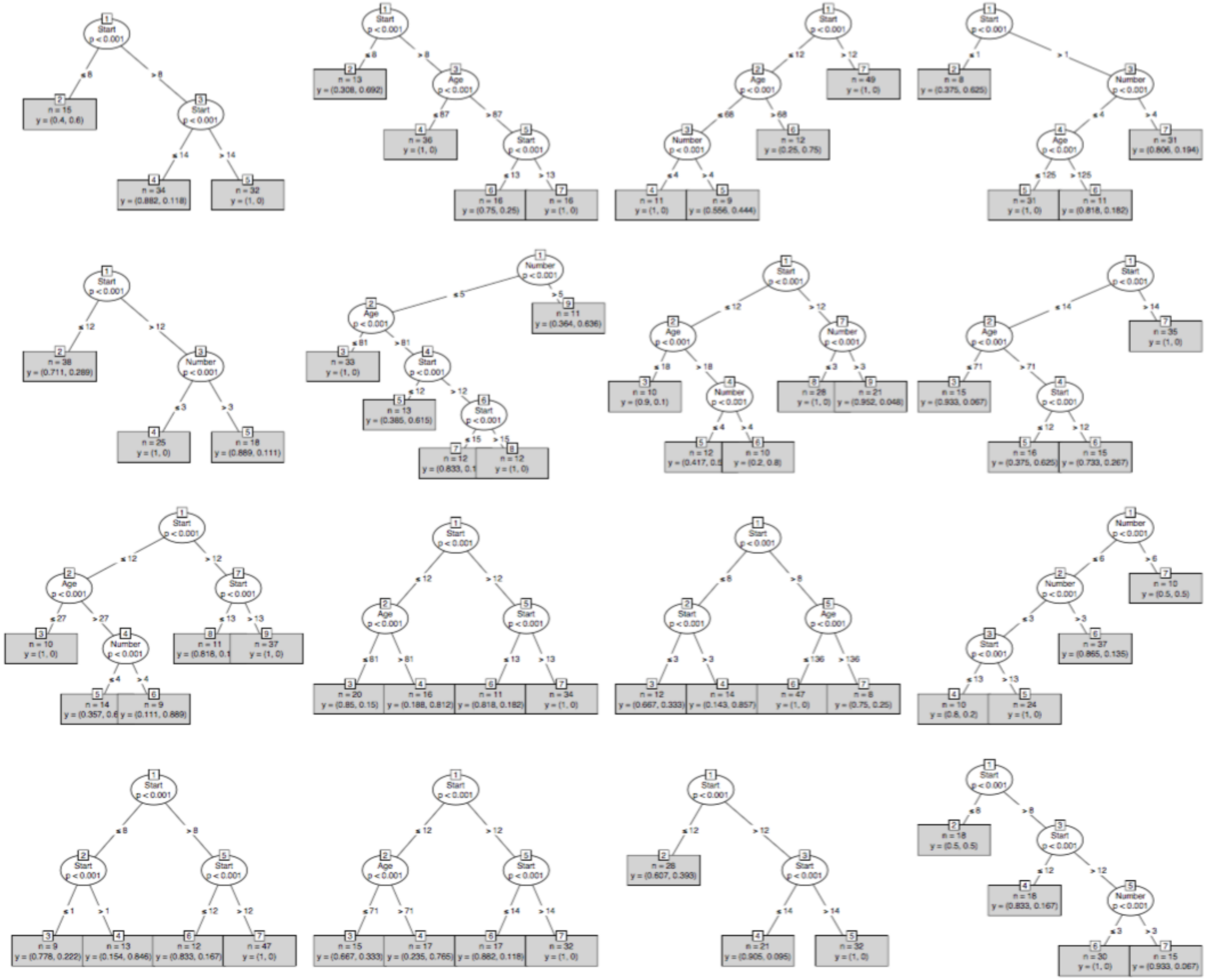
- *Leo Breiman, Random Forests, Machine Learning, 45, 5-32, 2001*
- Motivation: reduce error correlation between classifiers
- Main idea: build a larger number of un-pruned decision trees
- Key: using a random selection of features to split on at each node

JACOBS
UNIVERSITY

How Random Forest Work

- Each tree is grown on a bootstrap sample of the training set of **N** cases.
- A number **m** is specified much smaller than the total number of variables **M** (e.g. m = sqrt(M)).
- At each node, m variables are selected at random out of the M.
- The split used is the best split on these **m** variables.
- Final classification is done by majority vote across trees.

Random Forest (part of it)



JACOBS UNIVERSITY

Advantages of random forest

- Error rates compare favorably to Adaboost
- More robust with respect to noise.
- More efficient on large data
- Provides an estimation of the importance of features in determining classification

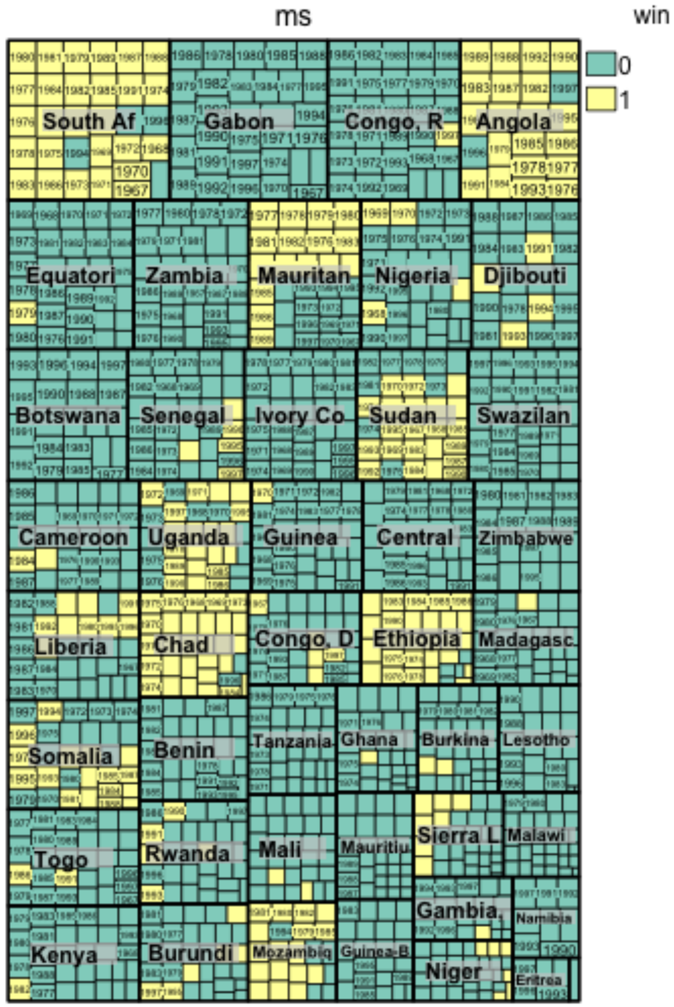
- http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm

Data sets

Table 1 Summary information about the data sets used in the evaluative comparison.

Data set	# cases	#countries	# time periods	incidence rate
Sub-Saharan Africa I	1017	46	26	0.22
Sub-Saharan Africa II	743	41	19	0.27
Petrostates	7768	188	59	0.17

Sub-Saharan Africa I



Ensemble model philosophy

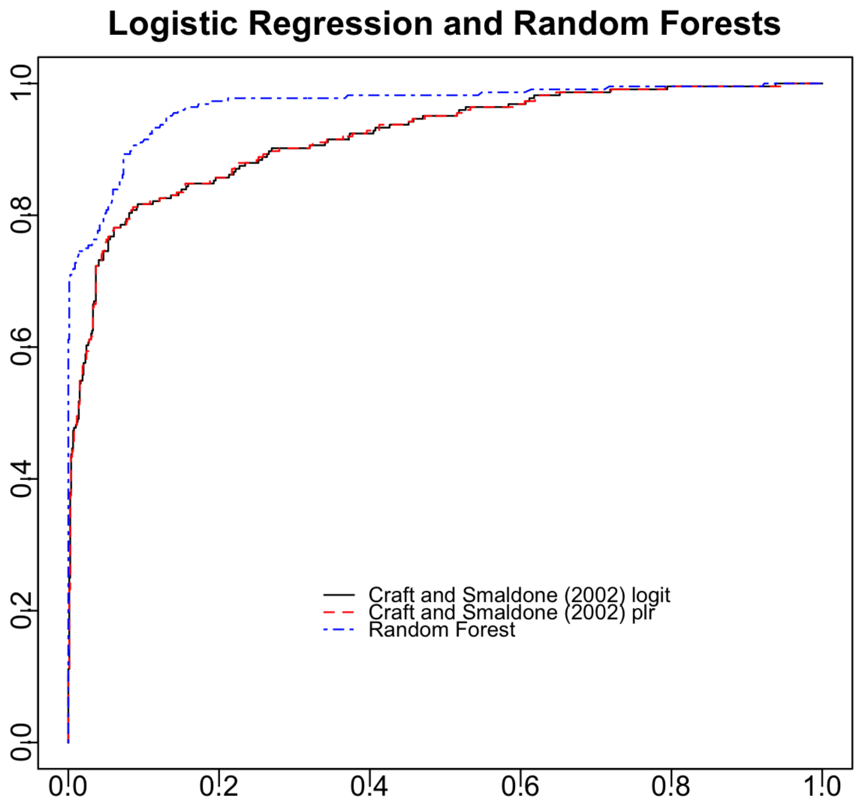
Data: Occurrence of military conflicts in Sub-Saharan Africa (Craft & Smaldone, 2002)

Random forests

10-fold Cross-validation

Variables Description of the variables.

- **Warinvol** war involvement binary variable, from Gleditsch et.al.
- **Year** year, 1967 through 1997
- **Colonial** colonial indicator, from Blanton et. al.
- **Country** country name
- **Transiti** transition binary, from Polity IV
- **Ethnopol** ethno-political groups indicator, from Minorities at Risk
- **Repressi** repression indicator, from Polity IV
- **Semidem** semi-democracy indicator, from Polity IV
- **Armstran** arms imports, from WMEAT (log values)
- **Milspend** per capita military spending, from WMEAT (log values)
- **Develop** per capita GNP, from WMEAT/World Bank (log values)
- **Cumwar** 5-year moving average of war magnitude, from Gleditsh et.al.



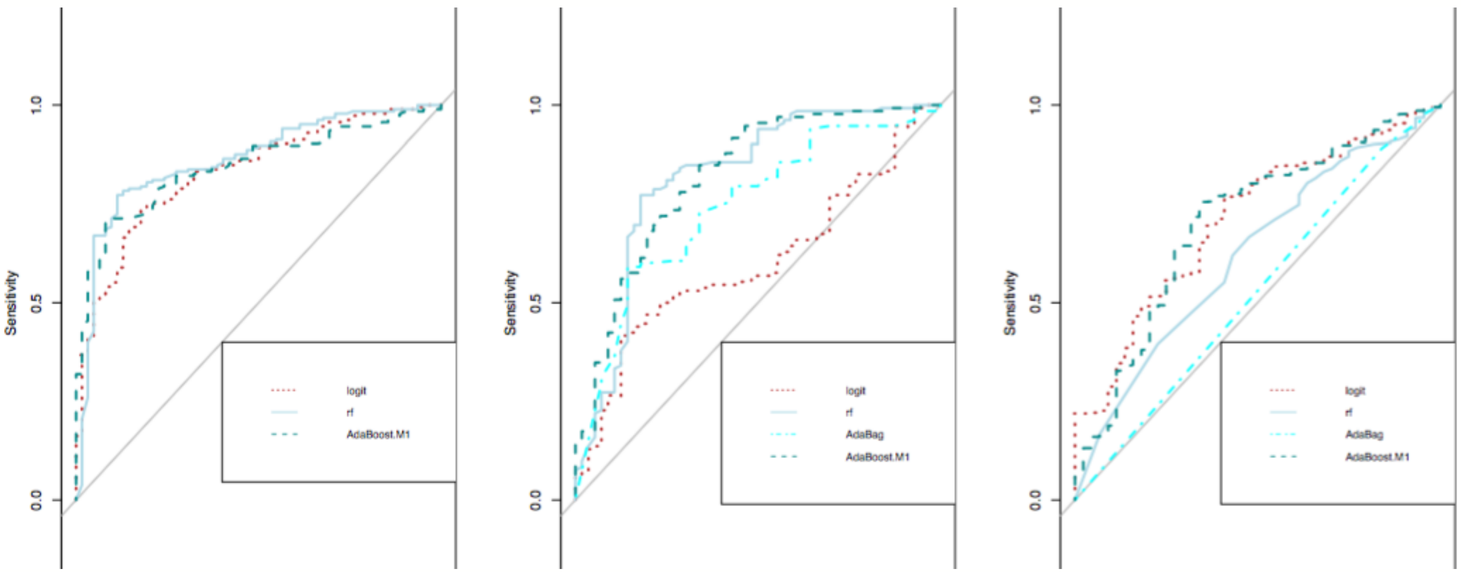
W. (2016)

JACOBS UNIVERSITY

Ensemble model philosophy

Table 2 Predictive accuracy for some classification techniques as measured by AUC (area under the ROC curve) on presented data sets.

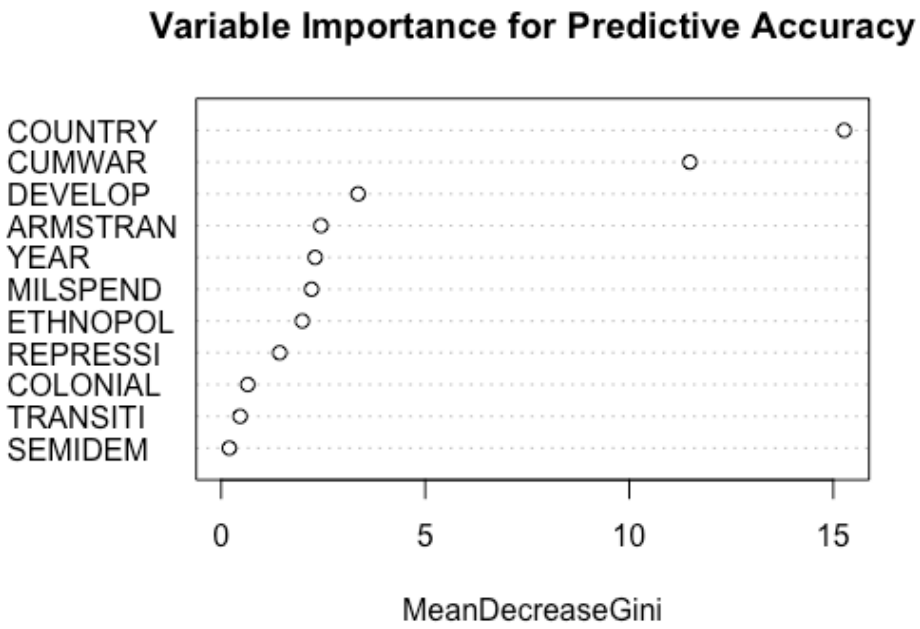
Data set	Logistic	AdaBag	AdaBoost	Random Forests
Sub-Saharan Africa I	0.8307	2	0.8481	0.8605
Sub-Saharan Africa II	0.6017	0.743	0.7551	0.8078
Petrostates	0.7086	0.5	0.7124	0.6048

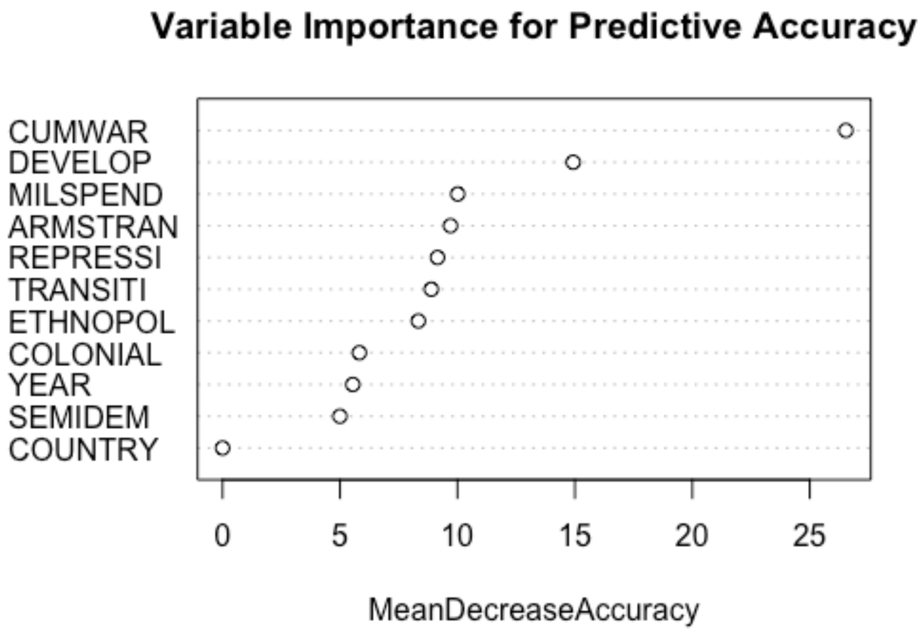


W. (2016)

JACOBS UNIVERSITY

Variable Importance Plots (Sub-Saharan Africa Data)





Gini importance
mean Gini gain produced by X_j over all trees

for variables of different types: biased in favor of continuous variables and variables with many categories (Strobl et al., 2007)

Permutation importance
mean decrease in classification accuracy after permuting X_j over all trees

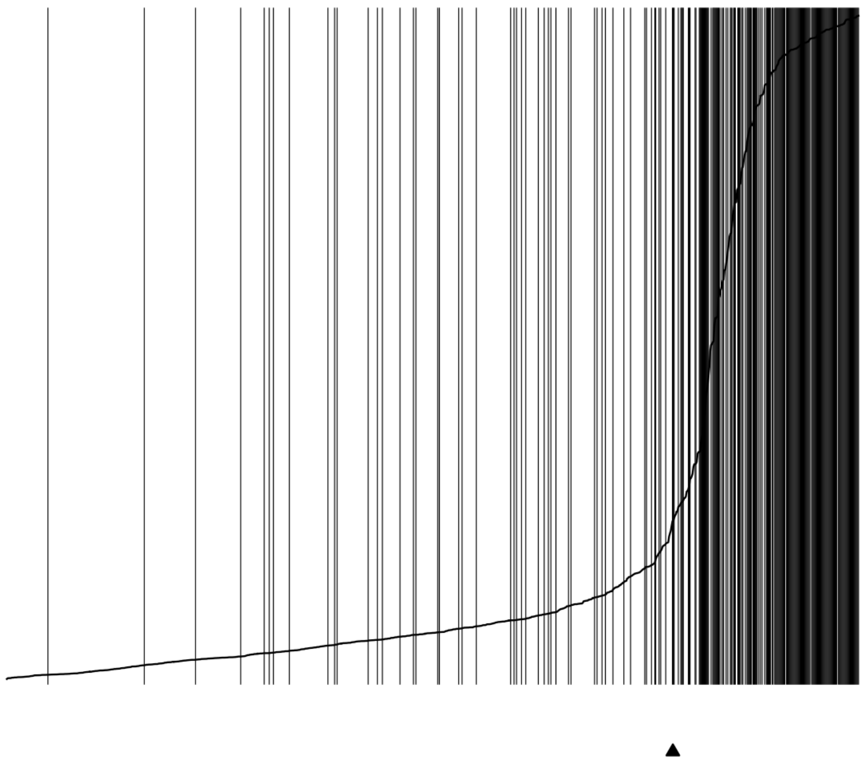
for variables of different types: unbiased only when subsampling is used (Strobl et al., 2007)

W. (2015)

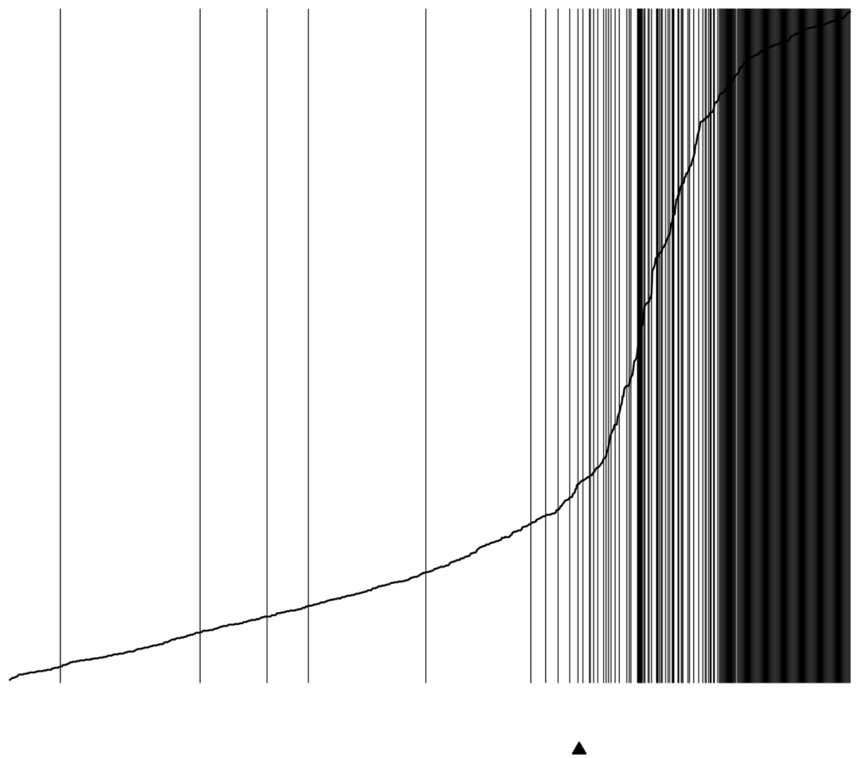
JACOBS UNIVERSITY

Separation Plots (Sub-Saharan Africa Data)

CRAFT AND SMALDONE (2002) -- logit



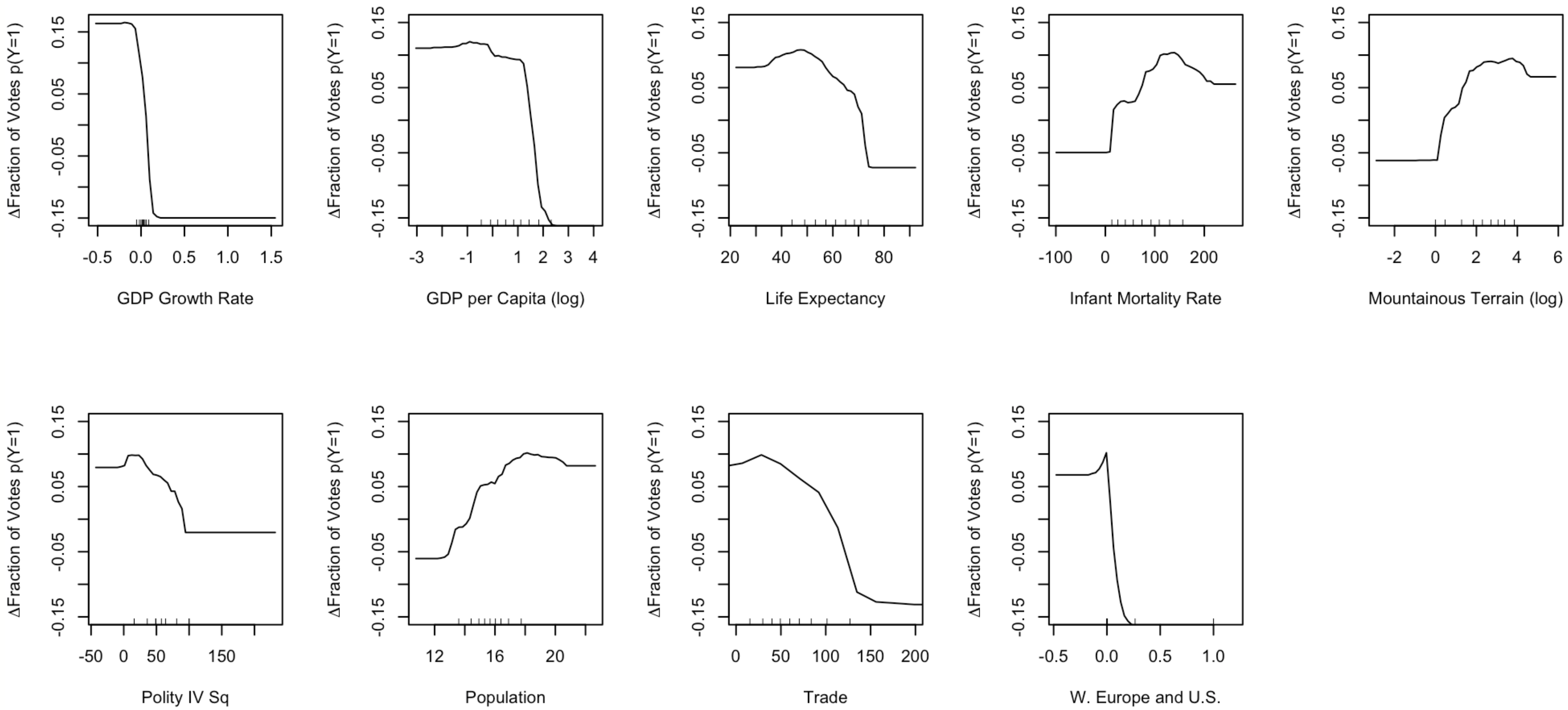
Random Forests



R Package: separationplot
Greenhill et al. (2015)

W. (2016)

Partial Dependence Plots (Civil War Data)



Muchlinski et al. (2016)

Class-imbalance

- Broad range of incidence rates
  - Restriction to politically relevant cases
  - Restriction to specific regions
  - Restriction to specific time frames
  - All these selections implicitly correct for class-imbalance!

Data set	# cases	#countries	# time periods	incidence rate
Sub-Saharan Africa I	1017	46	26	0.22
Sub-Saharan Africa II	743	41	19	0.27
Petrostates	7768	188	59	0.17
Civil War Data	7141	172	45	0.0165
ICOW	36156	34	109	0.0077
UCDP	4 314 736	116	68	0.0005

JACOBS UNIVERSITY

Rare events correction – Class imbalance problem

- For logistic regression, options to correct predicted probabilities for imbalanced data or to use penalized logistic regression (Firth' method)
- yields unbiased estimates for class-imbalanced data

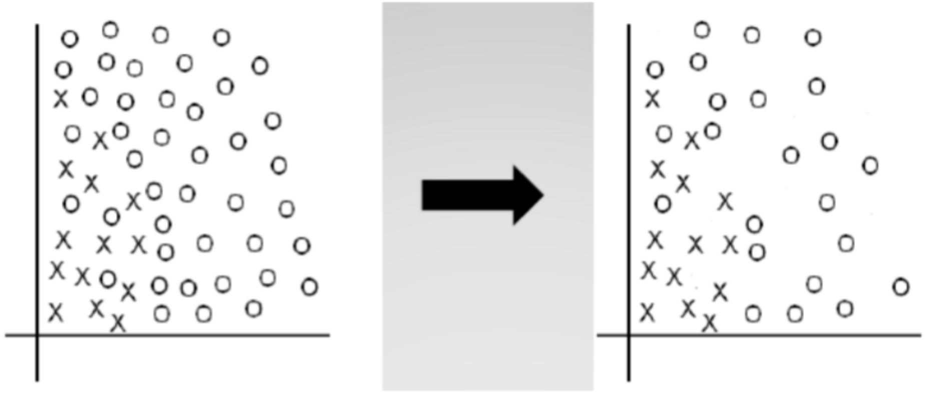
Rare events correction – Class imbalance problem

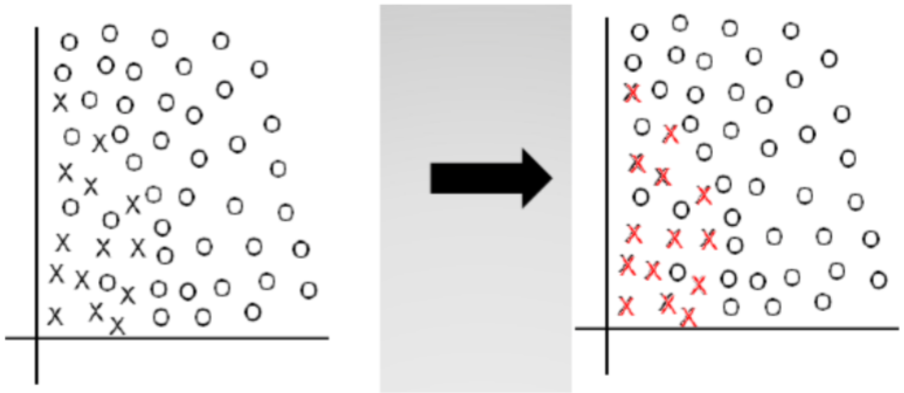
- For data-driven classifiers use sampling
  - Down-sampling
    - Looses information on majority class
  - Up-sampling
    - Repeats information of minority class

- Specify sampling counts per strata
  - Balanced design
  - Over-sampling minority class

Class imbalance problem - Solution approaches

- Sampling
  - Up-Sampling (Oversampling)
    - Reoeats information of minority class
  - Down Sampling (Undersampling)
    - Looses information of majority class
  - SMOTE (Synthetic Minority Over Sampling Technique)
  - Cluster-based or strata based sampling





JACOBS UNIVERSITY

Class imbalance problem - Solution approaches

SMOTE-Algorithm (k-NN approach)

Algorithm *SMOTE*(T, N, k)

Input: Number of minority class samples *T*;
Amount of SMOTE *N*%;
Number of nearest neighbors *k*

Output: (*N*/100)* *T* synthetic minority class samples

1. (∗ *If N is less than 100%, randomize the minority classsamples as only a random percent of them will be SMOTEd.* ∗)
2. **if** *N* <100
3. **then** Randomize the *T* minority class samples
4. *T* = (*N*/100) ∗ *T*
5. *N* = 100
6. **endif**
7. *N* = (*int*)(*N*/100)
(∗ *Amount of SMOTE is in integral multiples of100.* ∗)
8. *k* = Number of nearest neighbors
9. *numattrs* = Number of attributes
10. *Sample*[ ][ ]: array for original minority class samples
11. *newindex*: keeps a count of number of synthetic samples generated, initialized to 0
12. *Synthetic*[ ][ ]: array for synthetic samples

(∗ *Compute k nearest neighbors for each minority class sample.* ∗)

13. **for** *i* ← 1 **to** *T*
14. Compute *k* nearest neighbors for *i*, and save the indices inthe *nnarray*
15. Populate(*N*, *i*, *nnarray*)
16. **endfor**

Populate(*N, i, nnarray*)
(∗*Function to generate the synthetic samples.* ∗)

17. **while** *N* != 0
18. Choose a random number between 1 and *k*, call it *nn*.
(*This step chooses one ofthe *k* nearest neighbors of *i*.*)
19. **for** *attr* ← 1 **to** *numattrs*
20. Compute: *dif* = *Sample*[*nnarray*[*nn*]][*attr*] – *Sample*[*i*][*attr*]
21. Compute: *gap* = random number between 0 and 1
22. *Synthetic*[*newindex*][*attr*] = *Sample*[*i*][*attr*] + *gap* ∗*dif*
23. **endfor**
24. *newindex*++
25. *N* = *N* –1
26. **endwhile**
27. **return**(∗ *End of Populate.* ∗)

JACOBS UNIVERSITY

Class imbalance problem - Solution approaches

- Cost-sensitive learning
  - Weighted learning

Table 1: Cost matrix

		Prediction	
		Class i	**Class j**
True	**Class i**	**0**	λ_{ij}
	Class j	λ_{ji}	**0**

- Recognition based learning
- Ensemble methods
- Combinations of the above

Class imbalance problem - Solution approaches

Method	Advantages	Limitations
Under-sampling	• Independent on underlying classifier. • Can be easily implemented	• May remove significant patterns and cause loss of useful information
Over-sampling		• Time consuming: Introduce additional computational cost • May lead to over-fitting
Cost sensitive	• Minimize the cost of misclassification (by biasing the classifier toward the minority class)	• The misclassification costs (the actual cost of errors) often are unknown
Recognition based	• Have better performance especially on high dimensional data	• Many classifiers such as decision trees and Naive Bayes cannot be built by one class learning.
Ensemble	• Better classification performance than individual classifiers • More resilience to noise	• Time consuming • Over fitting

Elraham & Abraham 2013, Journal of Network and Innovative Computing, Volume 1 (2013) pp. 332-340

Case study 2

Data: Sub-Saharan Africa I
Random Forest

```
##    mtry       ROC      Sens      Spec      ROCSD     SensSD     SpecSD
## 1     2 0.9334589 0.9852186 0.7716912 0.04887503 0.01210057 0.09989737
## 2     6 0.9298289 0.9818579 0.7900735 0.05555949 0.01448013 0.07759298
## 3    11 0.9263345 0.9736885 0.8025735 0.05418566 0.02213473 0.08566255
```

Downsampling

```
##    mtry       ROC      Sens      Spec      ROCSD     SensSD     SpecSD
## 1     2 0.9325959 0.9621858 0.8209559 0.05274956 0.01733350 0.07974260
## 2     6 0.9177396 0.9621311 0.8213235 0.06242765 0.01754704 0.09307596
## 3    11 0.9157309 0.9621585 0.8147059 0.06426289 0.01355468 0.09577972
```


JACOBS
UNIVERSITY

Case study 2

Data: Sub-Saharan Africa II
Random Forest

```
##    mtry       ROC      Sens      Spec      ROCSD     SensSD    SpecSD
## 1     2 0.9489306 0.9733449 0.7766667 0.04110217 0.02414699 0.1275474
## 2     6 0.9502497 0.9709059 0.7971429 0.04190666 0.02231801 0.1087551
## 3    10 0.9488522 0.9635889 0.7904762 0.04606556 0.03093548 0.1105883
```

Downsampling

```
##    mtry       ROC      Sens      Spec      ROCSD     SensSD    SpecSD
## 1     2 0.9419993 0.9297329 0.8380952 0.03681699 0.04016076 0.1176104
## 2     6 0.9458408 0.9346109 0.8523810 0.03399345 0.03592463 0.1192062
## 3    10 0.9447466 0.9297329 0.8523810 0.03302378 0.04016076 0.1192062
```


JACOBS UNIVERSITY

Case study 2

Data: Petrostates
Random Forest

```
##    mtry       ROC      Sens Spec      ROCSD      SensSD SpecSD
## 1     2 0.6419646 1.0000000    0 0.08496380 0.000000000      0
## 2     9 0.6577368 0.9968446    0 0.05122113 0.003071956      0
## 3    17 0.6619219 0.9957919    0 0.05567687 0.003224936      0
```

Downsampling

```
##    mtry       ROC      Sens      Spec      ROCSD     SensSD     SpecSD
## 1     2 0.6789208 0.6529996 0.6516667 0.06018751 0.02529584 0.08022442
## 2     9 0.6767277 0.6729960 0.6112500 0.06561405 0.02320574 0.12981543
## 3    17 0.6756234 0.6600098 0.6320833 0.06742040 0.02359248 0.13999628
```

Class imbalance for conflict data

- dependency on magnitude of class imbalance
- correction needed for strong imbalances
- for pre-adjusted data sets correction may actually harm
- balanced design produces stable results

What can we conclude?

- Machine learning classifiers (in particular, random forests) improve prediction accuracy for onset of conflicts
- Variable importance results are fairly stable and a reasonable alternative to predictor significance in regression models
- Partial dependence plots enhance interpretability of ”causal effects”
- Existing non-linearities in relationships can be easily handled
- Theoretically existing rare event situations are avoided by sample pre-selection
- Rare event situations can be tackled by down-/up-sampling
- Data-driven classifiers are a valuable addition to the tool-kit of the quantitative-oriented social scientist
- First step towards a paradigmatic shift between explanation, prediction and modeling
- Wider acceptance of data-driven classifiers in the social sciences needs additional linkage to theory-driven approaches and their results

Future work?

- Causal Random Forests (Duncan, 2014)
- Mixed-effects random forests for clustered data (Haijem et al.. 2014)
    - to address
        - Serial correlation
        - Spatial correlation
        - Clustering
        - Hierarchical data
        - Panel structure
- Further evaluation of class imbalance effects

References:

- Beck, N., G. King, and L. Zeng. 2000. Improving quantitative studies of international conflict: A conjecture. American Political Science Review 94(1):21–35.
- Brandt, P., J. R. Freeman, and P. Schrodt. 2014. Evaluating forecasts of political conflict dynamics. International Journal of Forecasting 30:944–62.
- Breiman, L. 1996. Out-of-bag estimation. Technical report, Citeseer.
  ———. 2001a. Random forests. Machine Learning 45(1):5–32.
  ———. 2001b. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science 16(3):199–231.
- Clayton, G., and K. S. Gleditsch. 2014. Will we see helping hands? Predicting civil war mediation and likely success. Conflict Management and Peace Science 31:265–84.
- Duncan, G. M. 2014. Causal random forests. http://econ.washington.edu/sites/econ/files/old-site-uploads/2014/08/Causal-Random-Forests_Duncan.pdf
- Firth, D. 1993. Bias reduction of maximum likelihood estimates. Biometrika 80(1):27–38.
- Gleditsch, K. S., and M. Ward. 2012. Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes. Journal of Peace Research 50(1):17–31.
- Goldstone, J. A., R. H. Bates, D. L. Epstein, T. R. Gurr, M. B. Lustik, M. G. Marshall, J. Ulfelder, and M. Woodward (2010). A global model for forecasting political instability. American Journal of Political Science 54(1):190–208.
- Greenhill, B., M. D. Ward, and A. Sacks (2011). The separation plot: A new visual method for evaluating the fit of binary models. American Journal of Political Science 55(4):991–1002.
- Hajjem, A., F. Bellavance, and D. Larocque. 2014. Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation 84(6):1313–28.

References:

- Hegre, H., J. Karlsen, H. M. Nygard, H. Strand, and H. Urdal. 2013. Predicting armed conflict, 2010–2050. International Studies Quarterly 57(2):250–70.
- King, G., and L. Zeng. 2001. Logistic regression in rare events data. Political Analysis 9(2):137–63.
- Koridze, G., and A.F.X. Wilhelm, (2015). Exploratory evaluation of prediction accuracy of some machine learning algorithms applied to military conflicts. (Working Paper Jacobs University).
- Liaw, A. 2015. Package “randomforest”. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf
- Montgomery, J. M., F. M. Hollenbach, and M. D. Ward. 2012. Improving predictions using ensemble Bayesian model averaging. Political Analysis 20(3):271–91.
- Muchlinski, D., D. Siroky, J. He, and M. Kocher (2016). Comparing Random Forests with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. Political Analysis 24:87–103
- Schrodt, P., J. Yonamine, and B. E. Bagozzi. 2013. Data-based computational approaches to forecasting political violence. In Handbook of computational approaches to counterterrorism, ed. V. Subrahmanian, 129–62.
- Shellman, S. M., B. P. Levy, and J. K. Young. 2013. Shifting sands: Explaining and predicting phase shifts by dissident organizations. Journal of Peace Research 50:319–36.
- Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8:25.
- Ward, M. D., N. W. Metternich, C. Dorff, M. Gallop, F. M. Hollenbach, A. Schultz, and S. Weschle. 2012. Learning from the past and stepping into the future: The next generation of crisis predition. International Studies Review 15(4): 473–90.
- Wilhelm, A.F.X. (2015). Quality of classification approaches for the quantitative analysis of international conflict. CLADAG 2015, Cagliari
- Wilhelm, A.F.X. (2016). Data-driven classifiers in the analysis of conflict onset data. In preparation.

R packages used:

- library(randomForest) #for random forests
- library(caret) # for CV folds and data splitting
- library(ROCR) # for diagnostics and ROC plots/stats
- library(pROC) # same as ROCR
- library(stepPlr) # Firth’s logit implemented thru caret library
- library(doMC) # for using multipe processor cores
- library(separationplot)

Thank you very much for your attention!

Questions?

Comments?

JACOBS UNIVERSITY