Supplementary Appendix: Machine Learning Human Rights and Wrongs: How Supervised Learning Using Texts Can Inform the Debate About Changing Standards of Human Rights

Kevin T. Greene
Department of Political Science

Baekkwan Park
Department of Political Science

Michael Colaresi
Department of Political Science
mcolaresi@pitt.edu

December 10, 2017

Part I
Data

1 State Department Reports

The data for our analyses is taken from the State Department's Annual Country Reports on Human Rights Practices. The reports cover internationally recognized individual, civil, political, and worker's rights, as set forth in the Universal Declaration of Human Rights and other international agreements. The U.S. Department of State submits reports on all countries receiving assistance and all United Nations member states to the U.S. Congress in accordance with the Foreign Assistance Act of 1961 and the Trade Act of 1974. For the period 1978-1998, we rely on optical character recognition (OCR) scans of the primary documents[1]; for the period 1999-2010, we use web scrapped reports from the State Department's website. While our analysis drops words that appear in less than five percent of the documents, keeping most OCR errors from being included in our analysis, it is possible that meaningful words such as "killing" may be incorrectly rendered as "illing", possibly causing us to miss relevent information. To account for this and to further ensure our results are not driven by greater noise in the earlier reports, for the early period (1978-1998) we employ probabilistic spell correction, to correct OCR driven typos.

2 Political Terror Scale (PTS)

The Political Terror Scale (PTS) is a yearly, five point ordinal measure of a state's level of political violence and terror (Gibney and Cornett, 2015). The scale is coded based on the text of the annual human rights reports from both Amnesty International and the U.S. State Department. This study uses the PTS measure drawn from the U.S. State Department texts.

PTS Scores:
Level 1: Countries under a secure rule of law, people are not imprisoned for their view, and torture is rare or exceptional. Political murders are extremely rare.

Level 2: There is a limited amount of imprisonment for nonviolent political activity. However, few people are affected and torture and beatings are exceptional. Political murder is rare.

Level 3: There is extensive political imprisonment, or a recent history of such imprisonment. Execution or other political murders and brutality may be common. Unlimited detention, with or without a trial, for political views is accepted.

Level 4: The practices of level 3 are expanded to larger numbers. Murders, disappearances, and torture are a common part of life. In spite of its generality, terror on this level primarily affects those who interest themselves in politics or ideas.

Level 5: The terrors of level 4 have been expanded to the whole population. The leaders of these societies place no limits on the means or thoroughness with which they pursue personal or ideological goals.

[1]We thank Chris Fariss for sharing the documents.

Part II
Models

3 Supervised Learning Algorithms

Text classification is the task of assigning a given text document to one or more predefined categories depending on the contents of the document. Each document is categorized as one of the five PTS scores. Since instances (i.e., documents) are given with known labels (PTS ratings), we take a supervised machine learning approach.

We represent each document by a feature count vector. Meaning that a document (text) is modeled as a bag-of-words, a set of content words without any word order or syntactic relation information. Therefore, each unique word in documents becomes a separate feature. We use all the features available with TF (term frequency) and Tf-Idf (term frequency inverse document frequency) weighting in the hope that the informative or relevant features can be found. We also explore the role of higher order n-grams as features in discerning the subtleties reflecting human rights ratings. It is possible that higher order n-grams contain greater relevent information than simple unigrams. As suggested by Pang, Lee and Vaithyanathan (2002), employing higher order n-grams and combining them (unigram, bigram, and trigram together), could give us better performance than using them separately.

We train a number of linear and non-linear machine learning algorithms such as Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), and Random Forests (RF). Naive Bayes algorithm is a classification algorithm based on Bayes rule, that assumes all features of the training documents are independent of each other given the class (Lewis, 1998). This assumption simplifies the computation in handling data sets with many attributes. However, despite its performance, the conditional independence assumption may not hold in real-word situations. The Logistic Regression classifier is another supervised learning algorithm that we employ. It is a discriminative model which heavily relies on the logistic function (McCullagh and Nelder, 1989). Unlike NB, LR makes no assumptions about the relationship between features in the training documents. It take a linear combination of the input features, and assigns probabilities for each class, focusing on maximizing the probabilities. The further the data point lies from the separating hyperplane, the happier LR is. SVM is different from LR in that it tries to find the separating hyperplane that maximizes the distance of the closest data points to the margin (i.e., the support vectors) to reach classification of the input features (Cortes and Vapnik, 1995). If a data point is not a support vector, it does not really matter. Therefore, the main advantage of SVM algorithm is that it is relatively easier to overcome the high dimensionality problem that arises when there is a high number of input features relative to the number of available observations. For each learning algorithm, we employ regularization, because in supervised learning settings, with many input features, overfitting is a potential problem. The goal of regularization is to penalize large weights to avoid the problems of overfitting the data. Using too small a regularization parameter results in overfitting, and too large a value results in underfitting (Ng, 2004). We also use Random Forests (RF), an ensemble classifier. Random Forests operates by fitting a series of binary decision trees, where each split in the tree is determined by the variable that best reduces the misclassification rate (or best divides the data into similar subcomponents) (Breiman, 2001). Each of these binary decision tress are fit with a random sample of the data and a random sample of the independent variables (features). This randomization helps to prevent overfitting, and often leads to more accurate predictions. The predictions of each tree are then averaged to further reduce overfitting and produce more stable predictions.

4 Forecasting Models

4.1 Data Generating Process Simulations

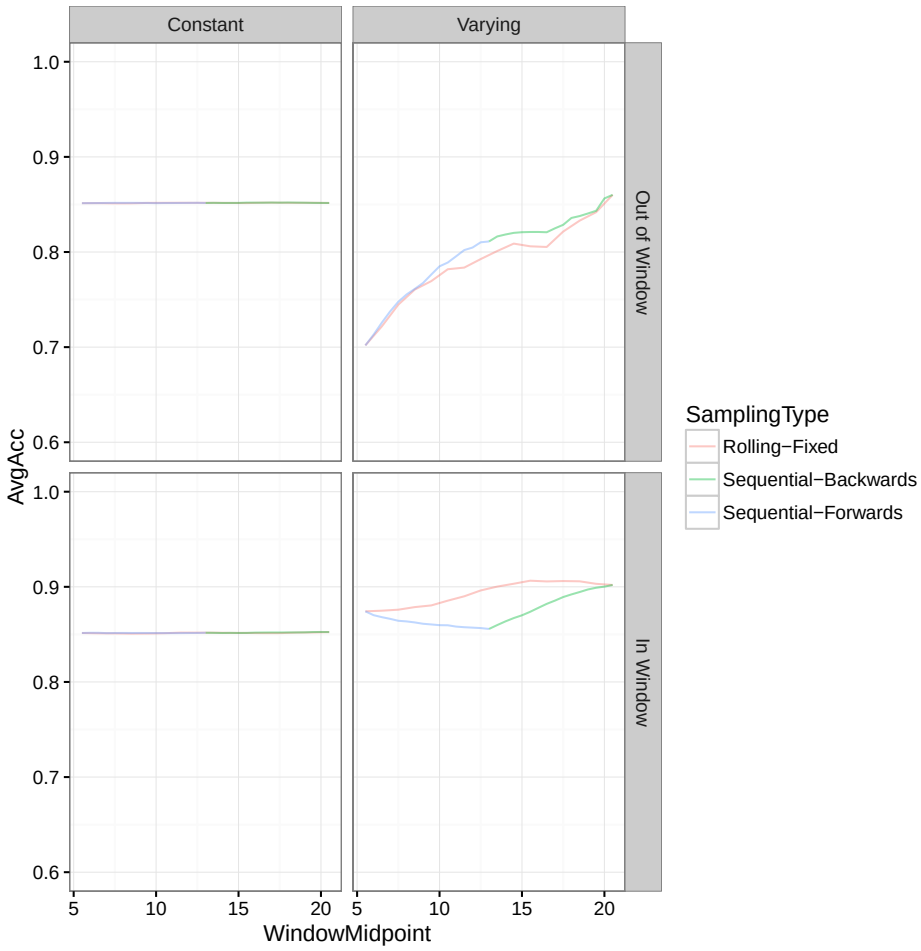

Figure 1: Out-of-window accuracy (first row) and in-window, 10-fold cross-validated accuracy (second row) from 250 simulated realizations of constant (first column) versus time-varying data (second column) generation processes. The time-varying process is generated from weights that follow a random walk.

Figure 5.2 illustrates an example of the distinct patterns of out-of-sample accuracy over different sets of training instances for constant versus varying data generation processes. We simulated 250 realizations of time-constant and varying data generation processes that produce observed pairs of x_{it} and y_{it} that represent data on units such as countries over time. We set aside the most recent quarter of the data as the out-of-training-window test set. The other three quarters of the data are kept as our available set of training instances. These instances are also used to measure the within-window performance for comparison. To vary the temporal distance between each training set and the fixed test set, we use three different sampling schemes to train models over different windows, (1) we roll a fixed width window through the training instances (Rolling-Fixed), (2) we sequentially extend a training window of minimum length n that extends from the last available training instance incrementally adding an observation each iteration (Sequential-Backwards), (3) and we use a window of minimum length n that begins with the

first available observation, and sequentially grows by one observation forward until the last instance is included (Sequential-Forwards).

Each window, within a sampling scheme, is uniquely defined by its middle time point and so that serves as the x-axis on each plot. We then use a generalized additive model, fitted to each realization, to learn the mapping from x_{it} to y_{it} given the training window. Average accuracy in the held-out test set across the 250 realizations is plotted on the y-axis in the first row to measure the performance of the models trained across the different windows on data from a distinct epoch. The in-window accuracy is presented on the second row for both data generation processes.

We see that out-of-window and in-window accuracy is stable over the different training sets and sampling schemes in the constant DGP case. In contrast, the out-of-window accuracy degrades for the time-varying process as the midpoints of the windows grow farther from the training set. The in-window results, fitted on temporally proximate data are much more stable, particularly over the Rolling-Fixed schemes. These are the observable fingerprints we are looking for in the actual human rights data, different DGPs should produce distinct patterns of out-of-window versus in-window accuracy over distinct samples of training instances. If information effects are part of the human rights data generation process, then we should see the out-of-sample performance in the test set degrade as the older data is used for training.

Our machine learning-inspired use of an explicit performance metric, accuracy, to compare models and learn about the data generation process has several unique advantages over a conventional null-hypothesis significance testing approach. Importently it allows us to identify observable implications from distinct and contradictory models without conditioning our inferences on strong identifying restrictions such as assuming that we know the one true underlying function/model. In our simulations above, a generalized additive model with time-constant parameters is used to learn the mapping from inputs to outputs, despite the DGP being time-varying.

Rolling Windows

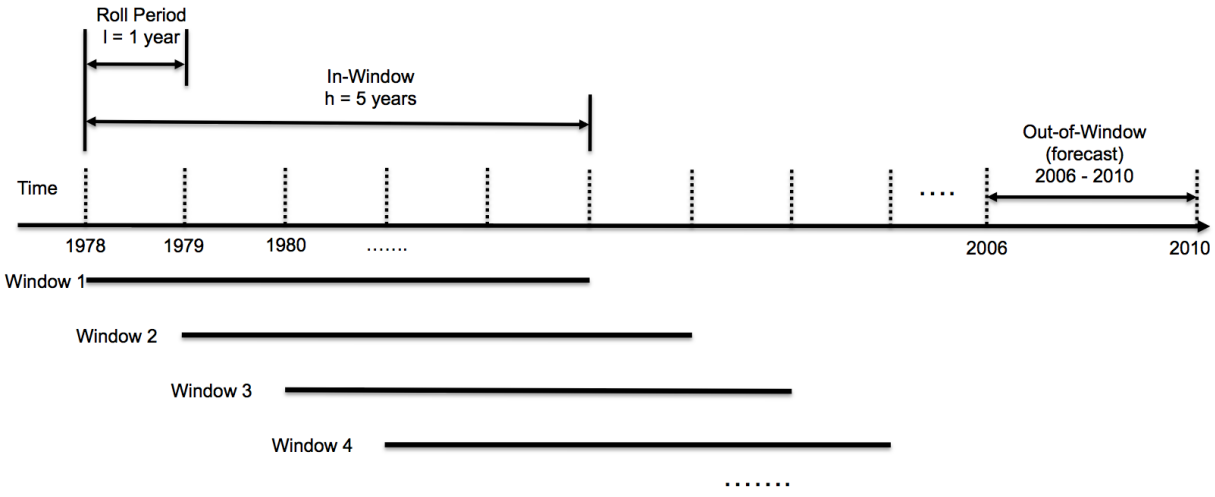

Figure 2: Example of Fixed Rolling Window: 5-Year Rolling Window Prediction

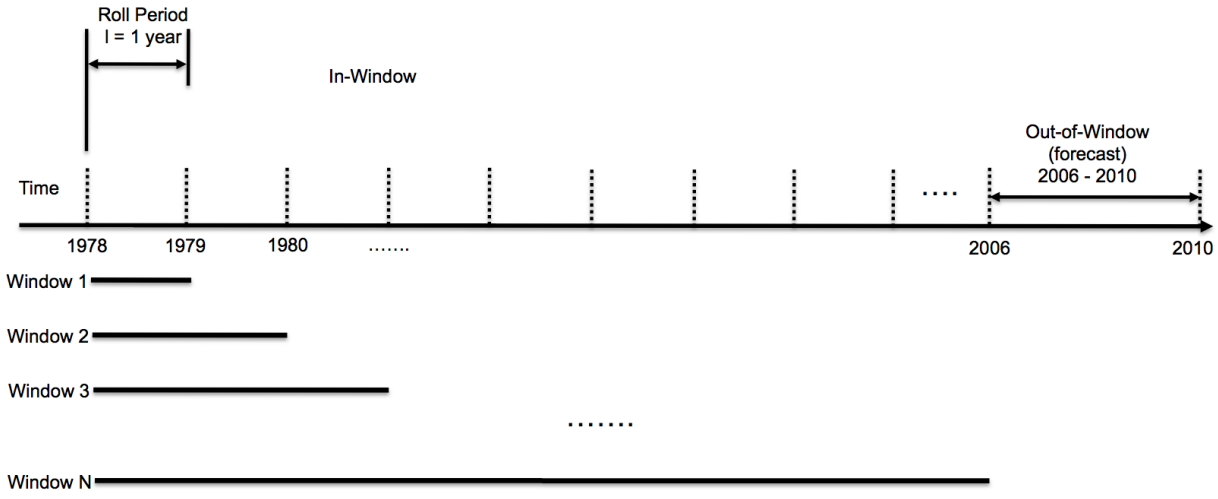

Figure 3: Example of Sequential Rolling Window: Sequential Forward Prediction

4.2 Sequential Forwards

The models here begin with a training window of one year (1978). Each movement of the window adds an additional year of training data until reaching the year 2005. The final window would contain the data for the entire in-window period 1978-2005. Consistent with our primary models the period 2006-2010 is used for out-of-window testing. Starting with a model fitted only with temporally distant data produces poor out-of-window prediction of a State's PTS score. However, as data gets closer to the out-of-window data set, the out-of-window accuracy increases. This provides further evidence of a changing DGP, and that our results are not an artifact of our rolling window setup.

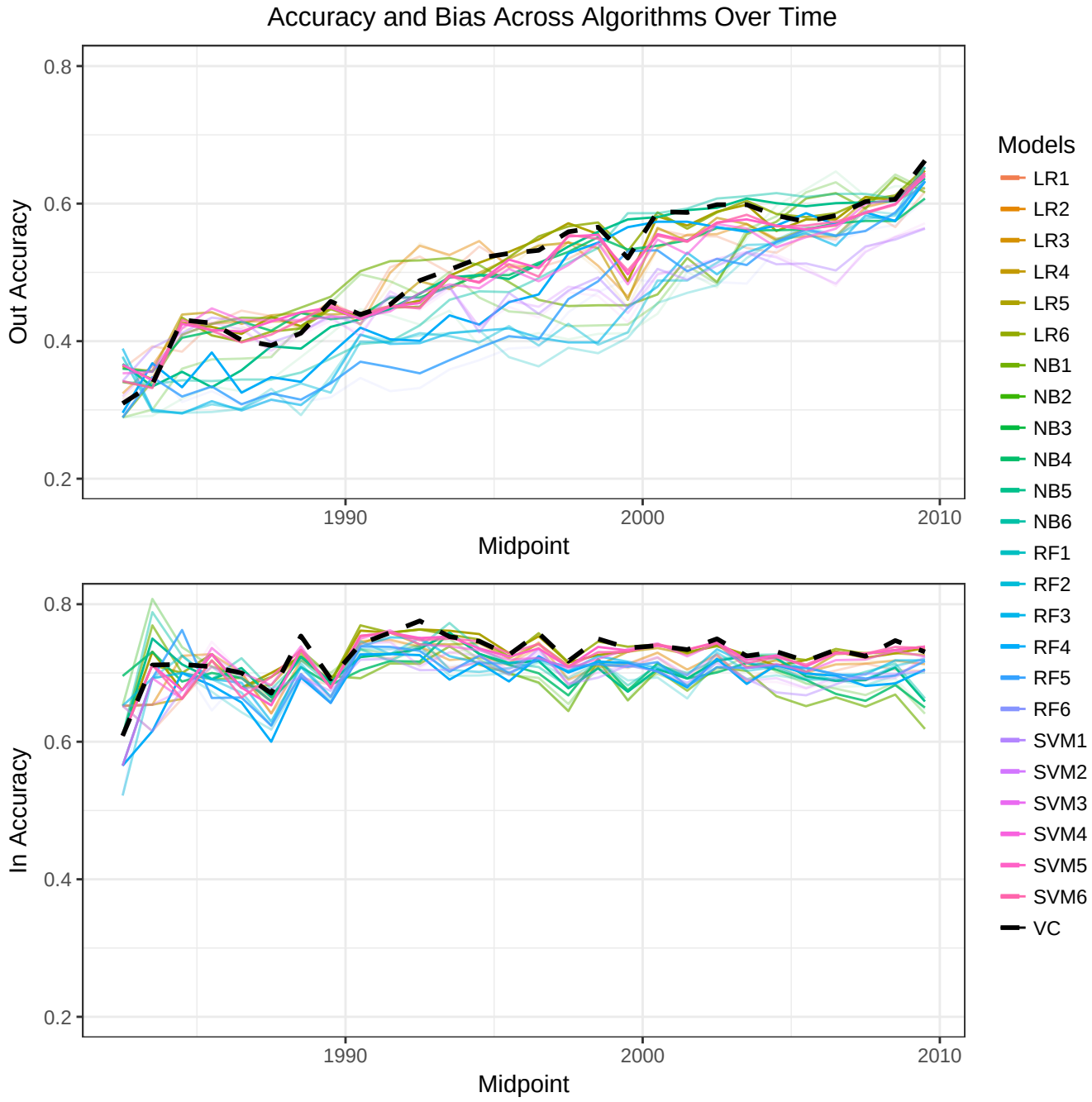

Figure 4: Sequential Forward from 1978 to 2005

4.3 Rolling 5 Year

The graphs below are identical to the main results in the article, except that the rolling window is changed from ten to five years. The first window is now 1978-1982, the out of sample window (2006-2010) is the same. This smaller window should allow for more dynamic changes in accuracy. More importantly

it demonstrates that our findings are not an artifact of the size of the rolling windows. The results are consistent with our primary models using 10 year rolling windows.

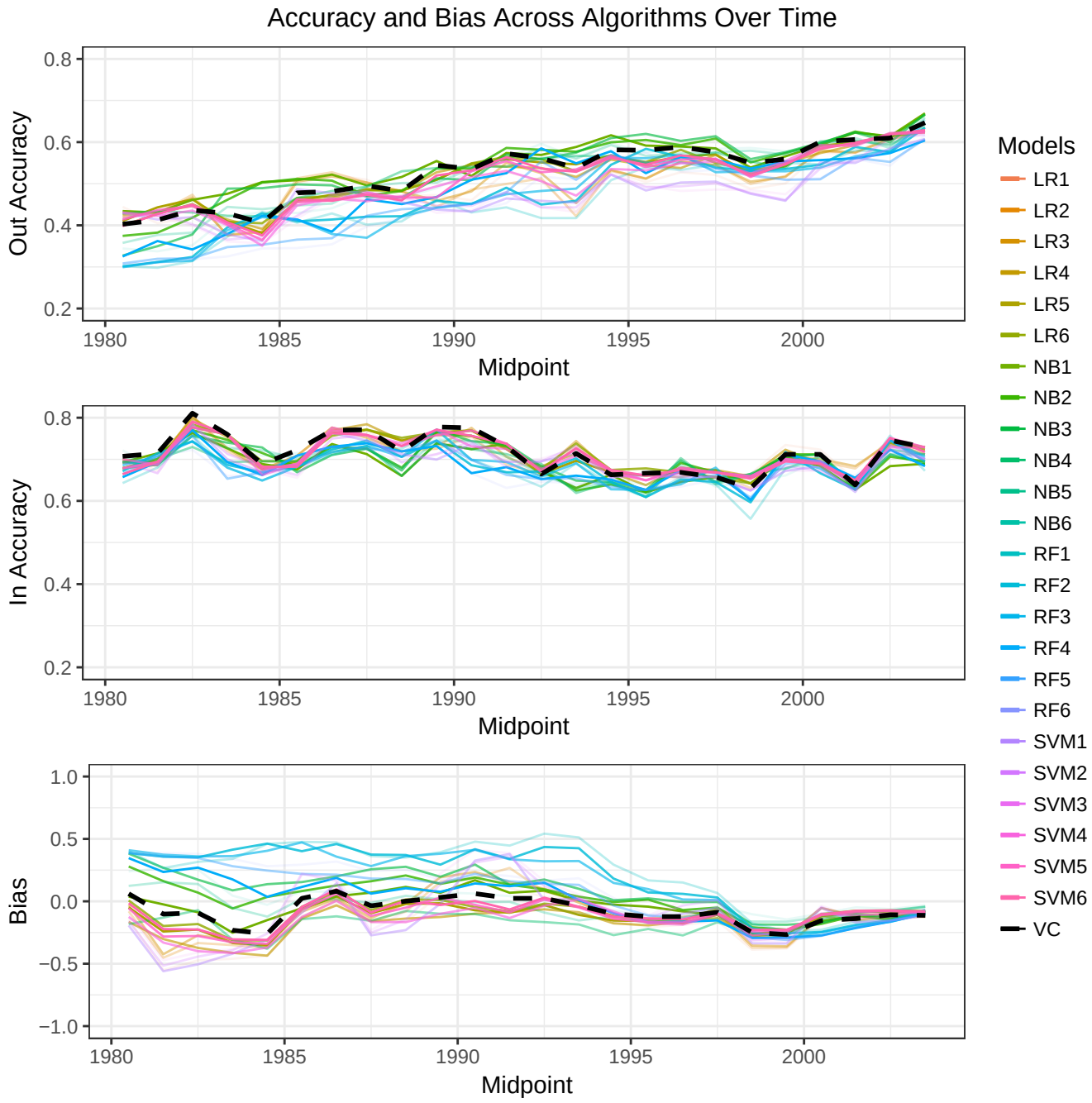

Figure 5: 5 Year Rolling Window

5 Evaluation Metrics

5.1 Precision, Recall, F-1

	Precision PTS					Recall PTS					F1-Score PTS				
In-Window	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1978-1987	0.87	0.71	0.68	0.48	0.50	0.90	0.74	0.64	0.52	0.31	0.88	0.72	0.66	0.50	0.38
1979-1988	0.89	0.67	0.65	0.67	0.44	0.88	0.80	0.60	0.42	0.57	0.88	0.73	0.63	0.52	0.50
1980-1989	0.95	0.64	0.59	0.68	0.83	0.90	0.70	0.62	0.58	0.71	0.92	0.67	0.61	0.62	0.77
1981-1990	0.91	0.67	0.64	0.61	0.67	0.92	0.67	0.69	0.59	0.44	0.92	0.67	0.67	0.60	0.53
1982-1991	0.95	0.61	0.64	0.71	0.69	0.91	0.80	0.58	0.46	0.75	0.93	0.69	0.61	0.56	0.72
1983-1992	0.85	0.61	0.56	0.61	0.55	0.87	0.62	0.59	0.57	0.40	0.86	0.62	0.58	0.59	0.46
1984-1993	0.93	0.71	0.61	0.67	0.50	0.90	0.76	0.71	0.43	0.53	0.91	0.73	0.66	0.52	0.51
1985-1994	0.88	0.65	0.59	0.55	0.57	0.91	0.67	0.61	0.46	0.57	0.89	0.66	0.60	0.50	0.57
1986-1995	0.84	0.71	0.71	0.60	0.67	0.94	0.64	0.65	0.60	0.70	0.89	0.68	0.68	0.60	0.68
1987-1996	0.84	0.71	0.60	0.6 0	0.65	0.92	0.71	0.63	0.55	0.46	0.88	0.71	0.61	0.57	0.54
1988-1997	0.86	0.71	0.56	0.51	0.64	0.88	0.73	0.58	0.54	0.32	0.87	0.72	0.57	0.53	0.42
1989-1998	0.86	0.58	0.58	0.51	0.75	0.86	0.68	0.51	0.49	0.65	0.86	0.63	0.54	0.50	0.70
1990-1999	0.84	0.65	0.64	0.60	0.54	0.83	0.71	0.65	0.52	0.50	0.83	0.68	0.65	0.56	0.52
1991-2000	0.84	0.68	0.55	0.59	0.70	0.91	0.66	0.62	0.47	0.56	0.88	0.67	0.58	0.52	0.62
1992-2001	0.78	0.65	0.78	0.47	0.54	0.88	0.71	0.56	0.46	0.62	0.83	0.68	0.65	0.47	0.58
1993-2002	0.74	0.64	0.57	0.68	0.76	0.83	0.70	0.54	0.47	0.70	0.78	0.67	0.55	0.55	0.73
1994-2003	0.76	0.62	0.63	0.74	0.73	0.83	0.68	0.60	0.60	0.58	0.79	0.65	0.62	0.66	0.65
1995-2004	0.76	0.57	0.64	0.71	0.48	0.76	0.61	0.66	0.44	0.72	0.76	0.59	0.65	0.55	0.58
1996-2005	0.70	0.59	0.71	0.63	0.81	0.75	0.63	0.66	0.56	0.81	0.72	0.61	0.69	0.59	0.81

Table 1: Evaluation Metrics for SVM (unigram, TF): 10-Year Rolling Window

$$\begin{bmatrix} 71 & 8 & 0 & 0 & 0 \\ 10 & 73 & 15 & 1 & 0 \\ 1 & 21 & 50 & 6 & 0 \\ 0 & 1 & 5 & 11 & 4 \\ 0 & 0 & 4 & 5 & 4 \end{bmatrix}$$

In-Window (1978-1987)

$$\begin{bmatrix} 54 & 18 & 0 & 0 & 0 \\ 20 & 66 & 15 & 4 & 0 \\ 2 & 23 & 67 & 9 & 0 \\ 1 & 3 & 12 & 27 & 5 \\ 0 & 2 & 0 & 3 & 21 \end{bmatrix}$$

In-Window (1996-2005)

Figure 6: Confusion Matrices for SVM (unigram, TF): 10-Year Rolling Window

	Precision PTS					Recall PTS					F1-Score PTS				
In-Window	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1978-1987	0.80	0.75	0.66	0.62	1.00	0.89	0.74	0.82	0.24	0.08	0.84	0.74	0.73	0.34	0.14
1979-1988	0.88	0.65	0.62	0.78	0.60	0.84	0.84	0.65	0.21	0.43	0.86	0.74	0.63	0.33	0.50
1980-1989	0.90	0.71	0.53	0.90	0.00	0.89	0.75	0.78	0.27	0.00	0.89	0.73	0.63	0.42	0.00
1981-1990	0.92	0.65	0.55	0.57	1.00	0.86	0.79	0.73	0.24	0.22	0.89	0.71	0.63	0.33	0.36
1982-1991	0.91	0.61	0.56	0.75	0.67	0.87	0.76	0.68	0.32	0.17	0.89	0.68	0.62	0.45	0.27
1983-1992	0.84	0.67	0.53	0.79	1.00	0.85	0.71	0.74	0.45	0.13	0.84	0.69	0.62	0.57	0.24
1984-1993	0.93	0.72	0.50	0.62	0.50	0.91	0.70	0.76	0.31	0.18	0.92	0.71	0.61	0.41	0.26
1985-1994	0.88	0.60	0.60	0.59	0.75	0.89	0.76	0.61	0.50	0.26	0.88	0.67	0.60	0.54	0.39
1986-1995	0.82	0.60	0.62	0.54	0.77	0.90	0.75	0.54	0.44	0.43	0.86	0.67	0.58	0.49	0.56
1987-1996	0.87	0.77	0.48	0.73	0.56	0.92	0.77	0.75	0.39	0.21	0.89	0.77	0.59	0.51	0.30
1988-1997	0.81	0.73	0.51	0.58	1.00	0.89	0.74	0.62	0.46	0.23	0.85	0.74	0.56	0.51	0.37
1989-1998	0.86	0.59	0.62	0.58	0.93	0.84	0.79	0.53	0.51	0.61	0.85	0.67	0.57	0.54	0.74
1990-1999	0.88	0.59	0.51	0.54	0.65	0.83	0.70	0.69	0.25	0.39	0.85	0.64	0.59	0.34	0.49
1991-2000	0.86	0.68	0.45	0.58	0.88	0.90	0.69	0.69	0.15	0.56	0.88	0.68	0.54	0.24	0.68
1992-2001	0.79	0.61	0.57	0.80	0.75	0.89	0.72	0.57	0.29	0.62	0.84	0.66	0.57	0.43	0.68
1993-2002	0.75	0.62	0.53	0.71	0.83	0.83	0.74	0.59	0.31	0.56	0.79	0.68	0.56	0.43	0.67
1994-2003	0.77	0.61	0.59	0.73	0.80	0.81	0.66	0.70	0.42	0.42	0.79	0.63	0.64	0.54	0.55
1995-2004	0.81	0.60	0.62	0.79	0.80	0.69	0.69	0.76	0.42	0.67	0.75	0.64	0.69	0.55	0.73
1996-2005	0.83	0.68	0.67	0.67	0.86	0.76	0.76	0.77	0.54	0.46	0.80	0.72	0.72	0.60	0.60

Table 2: Evaluation Metrics for Random Forests (unigram, TF): 10-Year Rolling Window

$$\begin{bmatrix} 70 & 9 & 0 & 0 & 0 \\ 14 & 73 & 12 & 0 & 0 \\ 1 & 12 & 64 & 1 & 0 \\ 1 & 1 & 14 & 5 & 0 \\ 1 & 2 & 7 & 2 & 1 \end{bmatrix}$$

In-Window (1978-1987)

$$\begin{bmatrix} 55 & 16 & 1 & 0 & 0 \\ 11 & 80 & 14 & 0 & 0 \\ 0 & 17 & 78 & 6 & 0 \\ 0 & 2 & 18 & 26 & 2 \\ 0 & 2 & 5 & 7 & 12 \end{bmatrix}$$

In-Window (1996-2005)

Figure 7: Confusion Matrices for Random Forests (unigram, TF): 10-Year Rolling Window

5.2 Full Variable Importance Output

The following sections contains the full output for the Random Forests variable importance. The graphs display the top 25 features PTS scores 4 and 5 for both the early and late years of the ten year rolling window. The early years are the first three windows from the 10 year rolling (1978-1987,1979-1988,1980-1989), while the late years are the last three windows from the 10 year rolling models. The rank of each word's variable importance was then averaged within the early and late windows to determine their average early and late rankings. The top 25 words from both the early and late windows, measured by average rank, are then plotted across our entire temporal range.

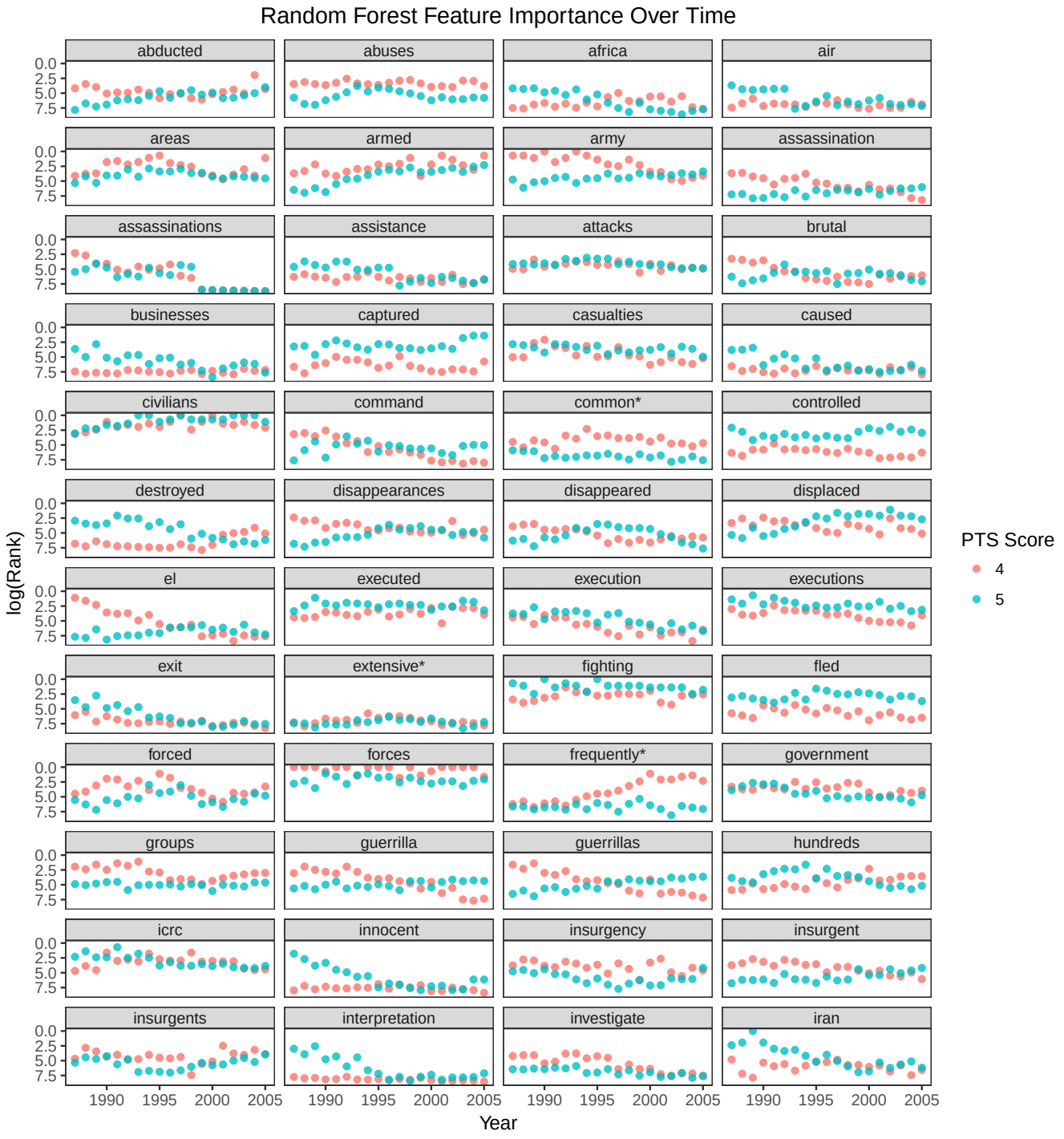

Figure 8: Full Variable Importance 4 and 5 Early Window

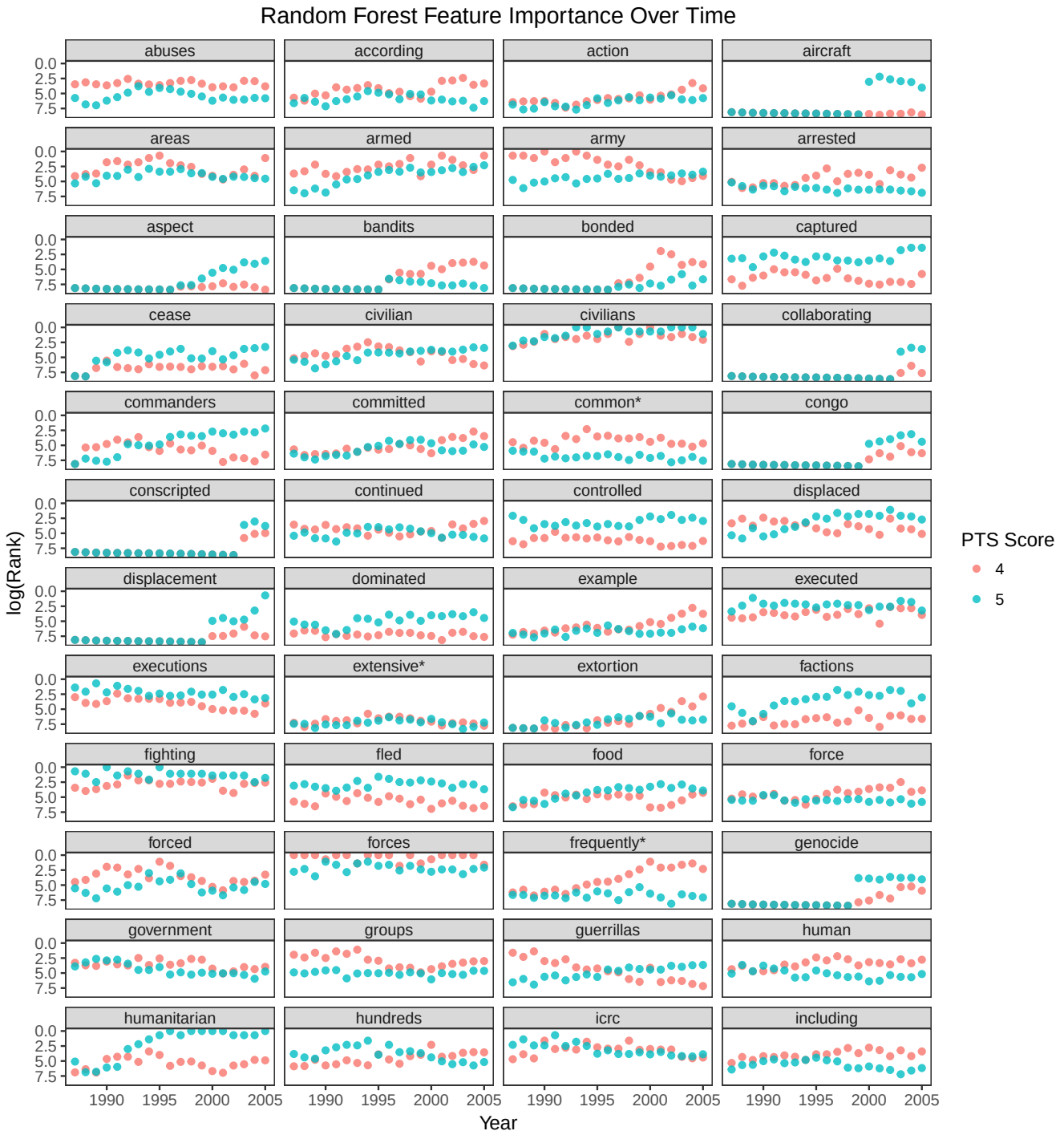

Figure 9: Full Variable Importance 4 and 5 Late Window

References

Breiman, Leo. 2001. "Random forests." *Machine learning* 45(1):5–32.

Cortes, Corinna and Vladimir Vapnik. 1995. "Support-vector networks." *Machine learning* 20(3):273–297.

Gibney, Mark and Reed Wood Peter Haschke Daniel Arnon Cornett, lInda. 2015. "The Political Terror Scale 1976-2015.".

Lewis, David D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Machine learning: ECML-98*. Springer pp. 4–15.

McCullagh, Peter and John A Nelder. 1989. *Generalized linear models*. Vol. 37 CRC press.

Ng, Andrew Y. 2004. Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*. ACM p. 78.

Pang, Bo, Lillian Lee and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics pp. 79–86.