Supplementary Materials for Understanding delegation through machine learning: A method and application to the European Union

1 Description of Performance Metrics

Each of these metrics provides a different perspective on each classifier's ability to detect delegation and constraint in the texts of EU legislation. In the machine learning and information retrieval literature, the most common measures used to measure the performance of an information retrieval system are precision, sensitivity and the combined F_1 metric (Ikonomakis, Kotsiantis, and Tampakas, 2005). Precision gives us information about how well a classifier is able to distinguish between a true and a false positive, sensitivity gives us information about how well a classifier is able to distinguish between a true positive and a false negative and the F_1 statistics combines these metrics to provide information about how well the classifier is able to distinguish between true positives, false positives and false negatives. We also report accuracy and specificity as part of our performance metrics although they are not particularly useful in this context due to the imbalanced nature of the data[1].

[1]This is a consequence of the nature of our training data for which there are far few "positive" classes, provisions delegating authority and provisions containing

Performance Metric	Symbol	Formula	Description
Accuracy	a	$\frac{TP+TN}{M}$	% correctly identified provisions.
Precision	π	$\frac{TP}{TP+FP}$	% of true v. false positives.
Sensitivity (Recall)	σ_1	$\frac{TP}{TP+FN}$	% of true positives
Specificity	σ_2	$\frac{TN}{TN+FP}$	% of true negatives.
F_1	F_1	$2 \times \frac{\pi \times \sigma_1}{\pi + \sigma_1}$	Combined performance metric.

Table 1: Accuracy a, precision π, sensitivity σ_1, specificity σ_2 and F_1 performance measures estimated for each GBT classifier determining which machine learning classifier is best suited to the task of measuring delegation and the imposition of constraints in EU provisions. Here TP = "true positive", FP = "false positive" , TN = "true negative" and FN = "false negative". Each of the metrics provide a different perspective on the classifier's ability to detect delegation and constraint in the texts of EU legislation.

2 Robustness Study: Classifier Performance From Other Machine Learning Algorithms

For the purpose of establishing a baseline level of performance with other machine learning algorithms, we trained several other commonly employed machine learning algorithms to identify delegation and constraint on national administrations and the European Commission. These include: naive Bayes, regularized logistic regression, support vector machines and vanilla random forests. Each of these algorithims have been employed in text classification problems with varying degrees of success (see eg Korde and Mahender (2012)).

The full results for each classifier are presented in the tables below with the F_1 score of each classifier in each category. All of these algorithms are trained using "out of the box" default settings with the exception of regularized logistic regression which was estimated using a penalty parameter λ that was determined through 10-fold cross validation. A glance at Table 2, which contains the delegation classifier F_1 scores from Tables 3 4, 5 and 6 show that the GBT classifiers discussed in the main text tends outperform other methods using this metric.

restraints, than there are "negative" classes, provisions neither delegating authority nor containing restraints.

Algorithim	EU Commission Delegation	Member States
Gradient Boosted Trees	0.754	0.730
Support Vector Machines	0.537	0.615
Naive Bayes	0.173	0.453
Regularized Logistic Regression	0.704	0.564
Random Forests	0.622	0.641

Table 2: F_1 score performance for delegation classifiers trained using 5 common machine learning algorithms.

Type	Accuracy	Precision	Recall	F_1
Authority Classifier				
Delegation to...				
EU Member States	0.905	0.727	0.533	0.615
EU Commission	0.954	0.917	0.379	0.537
Constraints Classifiers (EU Member States)				
Consultation Requirements	0.994	1	0.269	0.424
Appeals Procedures	0.995	1	0.111	0.2
Spending Limits	0.996	0.667	0.143	0.235
Rulemaking Requirements	0.907	0.439	0.196	0.271
Time Limits	0.991	0.7	0.212	0.326
Reporting Requirements	0.996	1	0.235	0.381
Executive Action Required	0.999	1	0.4	0.571
Executive Action Possible	0.999	0	0	NaN
Constraints Classifiers (EU Commission)				
Executive Action Possible	0.981	0.927	0.463	0.618
Public Hearings	0.997	1	0.111	0.200
Legislative Action Possible	0.995	1.000	0.188	0.316
Consultation Requirements	0.984	1.000	0.047	0.089
Rulemaking Requirement	0.969	1.000	0.061	0.115
Reporting Requirements	0.995	NaN	0.000	NA
Executive Action Required	0.997	NaN	0.000	NA
Time Limits	0.999	0.500	0.500	0.500

Table 3: Support Vector Machines: Performance metrics for authority and constraint classifiers.

Type	Accuracy	Precision	Recall	F_1
Authority Classifier				
Delegation to...				
EU Member States	0.793	0.364	0.6	0.453
EU Commission	0.401	0.096	0.897	0.173
Constraints Classifiers (EU Member States)				
Consultation Requirements	0.008	0.008	1	0.015
Appeals Procedures	0.005	0.005	1	0.011
Spending Limits	0.004	0.004	1	0.008
Rulemaking Requirements	0.824	0.23	0.429	0.3
Time Limits	0.013	0.01	1	0.019
Reporting Requirements	0.005	0.005	1	0.01
Executive Action Required	0.001	0.001	1	0.003
Executive Action Possible	0.001	0.001	1	0.002
Constraints Classifiers (EU Commission)				
Executive Action Possible	0.096	0.035	1	0.068
Public Hearings	0.004	0.004	1	0.007
Legislative Action Possible	0.006	0.006	1	0.013
Consultation Requirements	0.027	0.017	1	0.034
Rulemaking Requirement	0.198	0.038	0.963	0.073
Reporting Requirements	0.005	0.005	1	0.01
Executive Action Required	0.004	0.003	1	0.006
Time Limits	0.001	0.001	1.000	0.002

Table 4: Naive Bayes: Performance metrics for authority and constraint classifiers.

3 Description of Coding Process and NLP Parsing

Using a series of regular expressions, we broke down these 158 pieces of legislation into 8,417 provisions. Using the coding scheme provided by Franchino (2004), these were then coded as follows.

First, each provision was labeled using a binary classification system where 1 indicated that the provision delegated authority 0 that it does not. This resulted in the creation of two dummy variables: one member state delegation and one for

Type	Accuracy	Precision	Recall	F_1
Authority Classifier				
Delegation to...				
EU Member States	0.912	0.767	0.55	0.641
EU Commission	0.959	0.875	0.483	0.622
Constraints Classifiers (EU Member States)				
Consultation Requirements	0.995	1	0.308	0.471
Appeals Procedures	0.995	1	0.111	0.2
Spending Limits	0.996	1	0.143	0.25
Rulemaking Requirements	0.921	0.727	0.162	0.265
Time Limits	0.99	0.667	0.061	0.111
Reporting Requirements	0.996	0.75	0.176	0.286
Executive Action Required	0.999	1	0.4	0.571
Executive Action Possible	0.999	NaN	0	NA
Constraints Classifiers (EU Commission)				
Executive Action Possible	0.983	0.955	0.512	0.667
Public Hearings	0.997	1	0.111	0.2
Legislative Action Possible	0.996	1	0.438	0.609
Consultation Requirements	0.986	1	0.163	0.28
Rulemaking Requirement	0.969	1	0.061	0.115
Reporting Requirements	0.995	NaN	0	NA
Executive Action Required	0.997	NaN	0	NA
Time Limits	0.999	0.5	0.5	0.5

Table 5: Random Forests: Performance metrics for authority and constraint classifiers.

European Commission delegation. Similarly, another classification system was created wherein 1 indicated that a provision included each of the constraints on the authority of member-state administrations and/or the European Commission and 0 that it does not. This resulted in the creation of an additional 16 dummy variables, 8 for the constraints on member states and 8 for constraints on the Commission.

Of all coded provisions, 15.4% were identified as delegating executive authority to EU member states and 6.3% were identified as delegating authority to the European

Type	Accuracy	Precision	Recall	F_1
Authority Classifier				
Delegation to...				
EU Member States	0.879	0.623	0.516	0.564
EU Commission	0.962	0.731	0.679	0.704
Constraints Classifiers (EU Member States)				
Consultation Requirements	0.99	0.345	0.385	0.364
Appeals Procedures	0.994	0.333	0.111	0.167
Spending Limits	0.994	0.286	0.286	0.286
Rulemaking Requirements	0.849	0.251	0.361	0.296
Time Limits	0.988	0.387	0.364	0.375
Reporting Requirements	0.994	0.417	0.294	0.345
Executive Action Required	0.996	0.154	0.4	0.222
Executive Action Possible	0.999	0.333	0.25	0.286
Constraints Classifiers (EU Commission)				
Executive Action Possible	0.983	0.683	0.84	0.753
Public Hearings	0.997	0.5	0.6	0.545
Legislative Action Possible	0.994	0.5	0.65	0.565
Consultation Requirements	0.983	0.426	0.489	0.455
Rulemaking Requirement	0.948	0.233	0.304	0.264
Reporting Requirements	0.994	0.375	0.353	0.364
Executive Action Required	0.988	0.088	0.231	0.128
Time Limits	0.996	0.083	0.5	0.143

Table 6: Regularized Logistic Regression: Performance metrics for authority and constraint classifiers.

Commission [2].

[2]These statistics are lower than the original statistics in Franchino (2004, 2007) because of the increase in provisions generated by the parsing and coding process. While Franchino (2004, 2007) calculates these statistics using only full articles, we further partitioned each of the 158 pieces of legislation that we used for training by articles and subarticles as identified in Franchino (2004, 2007) codebook.

3.1 Units of analysis

HAS ADOPTED THIS DIRECTIVE:

Article 1

Subject matter and scope

This Directive lays down rules on a mechanism to resolve disputes between Member States when those disputes arise from the interpretation and application of agreements and that provide for the elimination of double taxation of income and, where applicable, capital. It also lays down the rights and obligations of the affected persons when such disput the purposes of this Directive, the matter giving rise to such disputes is referred to as a 'question in dispute'.

Article 2

Definitions

1. For the purposes of this Directive, the following definitions apply:

(a) 'competent authority' means the authority of a Member State which has been designated as such by the Member State concerned;

(b) 'competent court' means the court, tribunal or other body of a Member State which has been designated as such by the Member State concerned;

(c) 'double taxation' means the imposition by two or more Member States of taxes covered by an agreement or convention referred to in Article 1 in respect of the same tax or capital when it gives rise to either: (i) an additional tax charge; (ii) an increase in tax liabilities; or (iii) the cancellation or reduction of losses that could be used to o profits;

(d) 'affected person' means any person, including an individual, that is a resident of a Member State for tax purposes, and whose taxation is directly affected by a question in c

2. Any term not defined in this Directive shall, unless the context requires otherwise, have the meaning that it has at that time under the relevant agreement or convention re Article 1 that applies on the date of receipt of the first notification of the action that resulted in, or that will result in, a question in dispute. In the absence of a definition under suc or convention, an undefined term shall have the meaning that it had at that time under the law of the Member State concerned for the purposes of the taxes to which the said a convention applies, any meaning under the applicable tax laws of that Member State prevailing over a meaning given to the term under other laws of that Member State.

Figure 1: Part of EU Council Directive 2017/1852 (CELEX # 32017L1852)

We parse the EU legislative texts that we collect into *articles* as the units of analysis rather than provisions which are further parsed into sub–articles. We chose this unit of analysis because it offers a cleaner means of automatically processing texts via natural language processing and offers a more interpretable and meaningful unit of comparison among legislative texts across time. As an example of an article, in EU Council Directive 2017/1852 (CELEX # 32017L1852) on tax dispute resolution mechanisms in the European Union above all of *Article 1* is a single unit of analysis.

4 Categories of Restraint and Data Limitations

Franchino (2004) identified 12 categories of procedural constraints adapted to the European Union from Epstein and O'Halloran (1999). For our purposes, these constraints restrain members states' actions when authority is delegated to them. Below we include additional information for constraints that are not self–explanatory. A more detailed description of these definitions can be found in Franchino (2004).

While classification of provisions delegating authority to member states was straightforward and required training only one classifier, reconstructing constraint ratios using GBTs was more complicated as a separate classifier had to be trained for each of the restraint categories for which sufficient data were available. Of the 8,417 articles coded in the Franchino training data, there were only 11 (0.1%) coded articles available for the restraint category "Executive Action Possible" and 1 (0.01%) available for the "Public Hearings" and "Legislative Action Possible" categories. While these data limitations rendered training and testing machine learning classifiers to identify these categories of restraint impossible, the lack of data for these categories also suggests that the impact of removing these categories on the faithful reproduction of constraint ratios is minimal.

- **Time Limits**
- **Spending Limits**
- **Reporting Requirements** – requirements of member states to report to committees on actions taken.
- **Consultation Requirements** – a consultation procedure that member states must abide to when using their discretionary authority.
- **Public Hearings**
- **Rule–making Requirements**

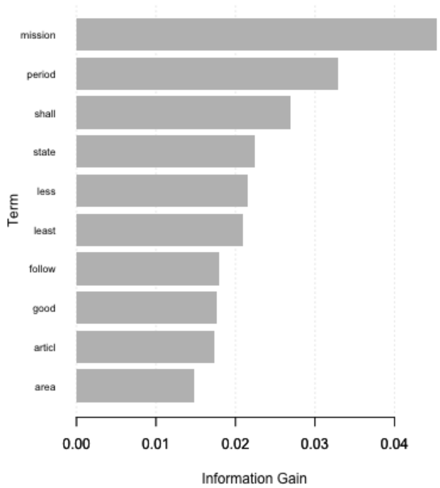
- **Appeals Procedures** – member states must justify the decisions that they make and have a right to appeal.
- **Exemptions** – limits to the scope of an act.
- **Legislative Action Required** – member state measure requires approval of the Commission before becoming effective.
- **Legislative Action Possible** - actions of a member state are referred to the Commission prior to becoming effective.
- **Executive Action Required** – executive agent must approve actions prior to becoming effective. This would include prior approval by the Commission for actions taken by the member states.
- **Executive Action Possible** – measures taken by member states can be overruled by actions taken by the commission.

5 Term Importance for Restraint Categories Measure by Information Gain

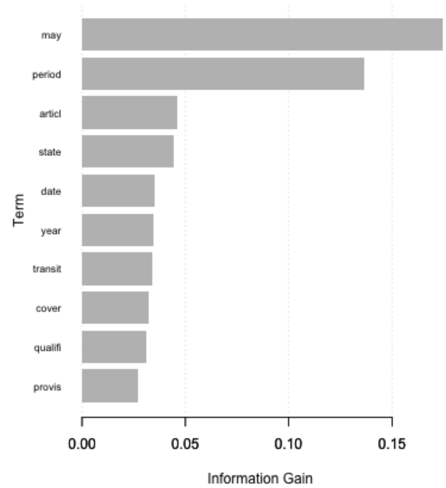
As mentioned in the main text, one of the great benefits of GBT classifiers is that they provide information for determining the most important terms for each delegation and restraint category. Below we include plots of term importance for each delegation and restraint category for EU member states and for the EU commission.

Figure 2: Gradient Boosted Tree Estimates of Term Importance for Delegation and Restraint Categories: Member States

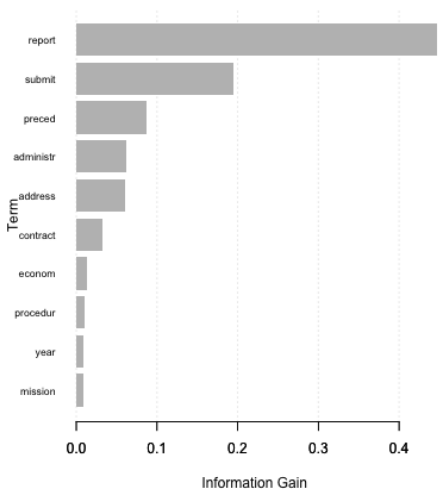
(a) Rule-making Requirements



(b) Time Limits



(c) Reporting Requirements



(d) Consultation Requirements

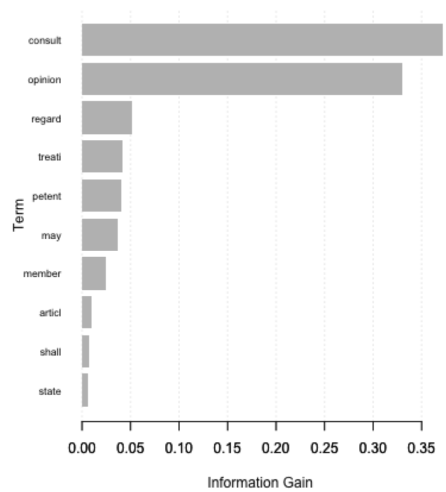
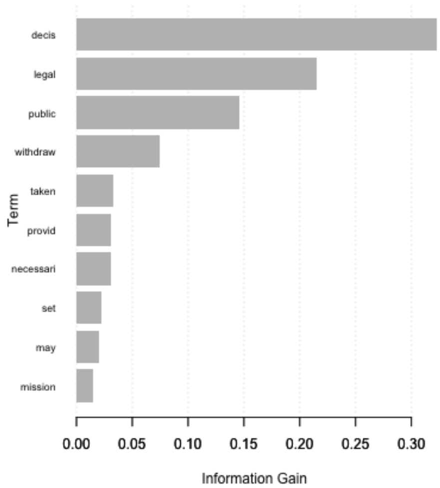
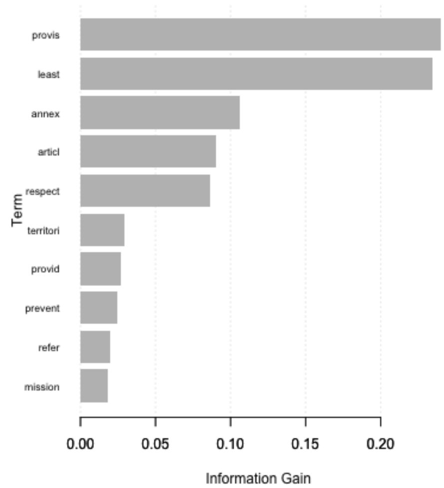


Figure 3: Gradient Boosted Tree Estimates of Term Importance for Delegation and Restraint Categories: Member States

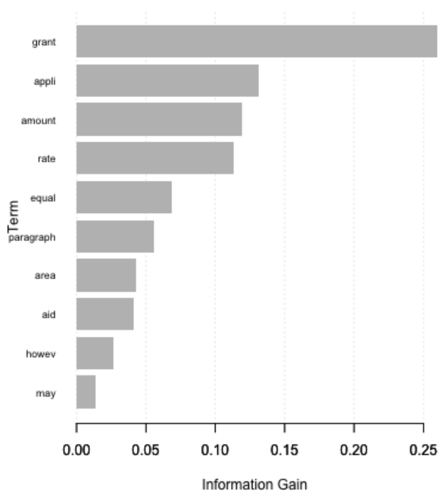
(a) Appeals Procedures



(b) Executive Action Required



(c) Spending Limits



(d) Executive Action Possible

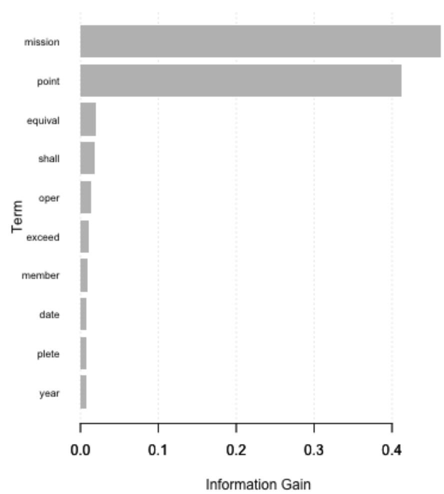
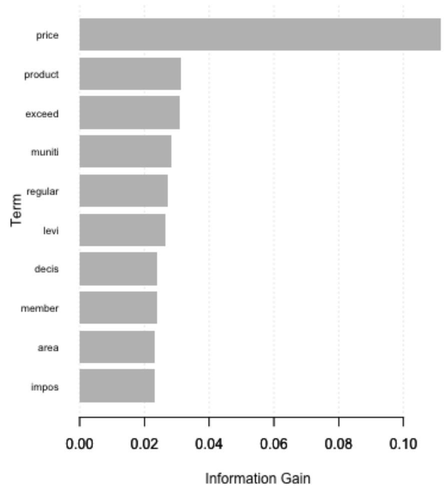
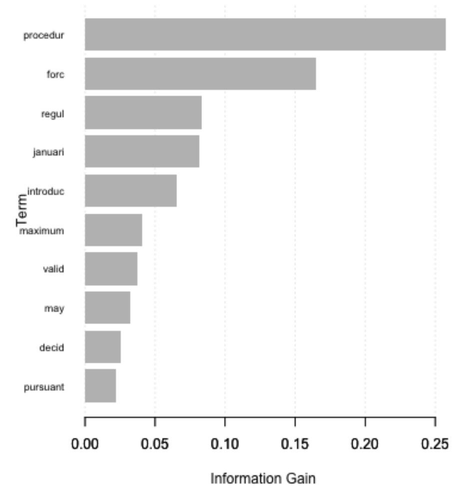


Figure 4: Gradient Boosted Tree Estimates of Term Importance for Delegation and Restraint Categories: European Commission

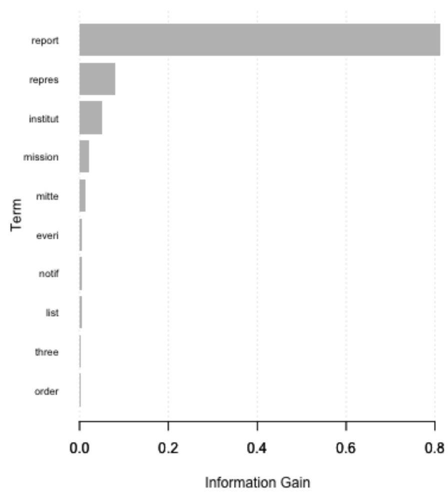
(a) Rule-making Requirements



(b) Time Limits



(c) Reporting Requirements



(d) Consultation Requirements

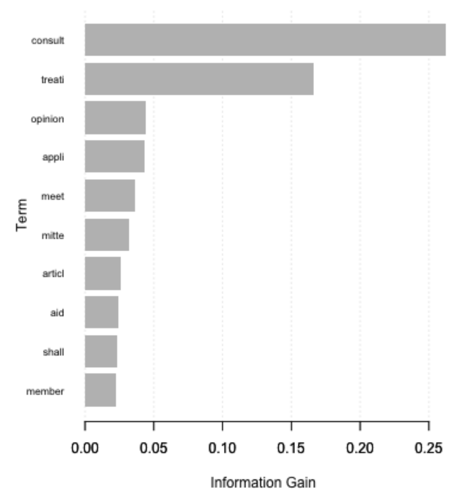
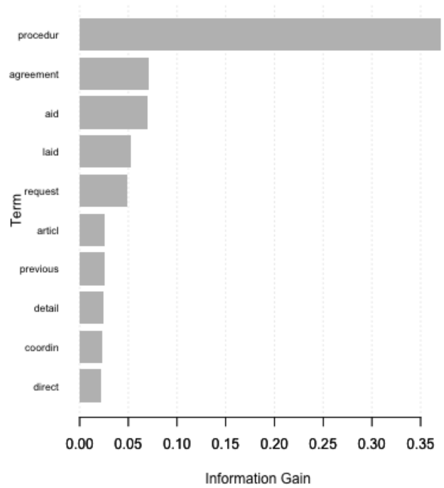
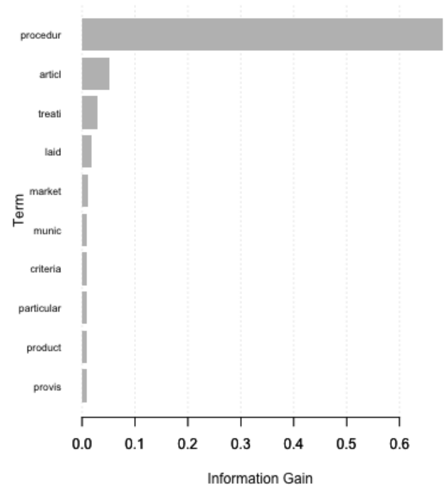


Figure 5: Gradient Boosted Tree Estimates of Term Importance for Delegation and Restraint Categories: European Commission

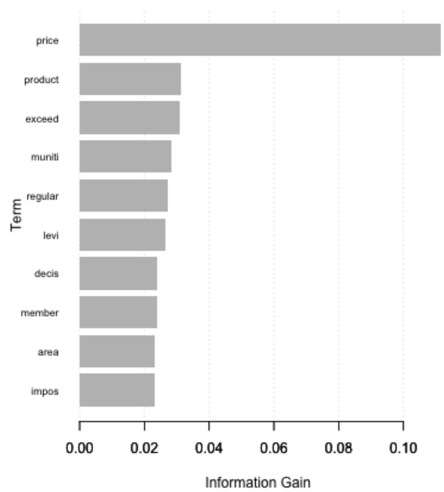
(a) Executive Action Required



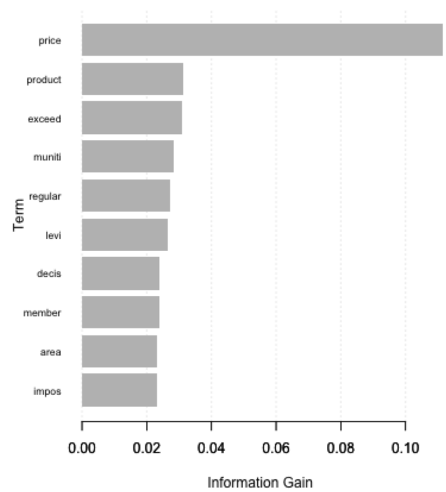
(b) Executive Action Possible



(c) Public Hearings



(d) Legislative Action Possible



6 Description of Gradient Boosted Trees and Text Pre–Processing Steps

Each classifier trained was trained using identical text pre–processing methods with a series of gradient–boosted tree classifiers with regularization and hyper–parameter tuning. Gradient-boosted trees are a variant of a decision tree algorithm which, like their random forests predecessor, grow multiple trees from random subsets of the training data and use a majority vote rule of the trees to generate the final class label. This method has become popular in the social sciences and frequently used for text classification problems because it has been found to be among the most transparent and accurate methods for a variety of applications (Athey et al., 2016; Chalfin et al., 2016; Chen and Guestrin, 2016; Hinton and Salakhutdinov, 2009; Kleinberg et al., 2017).

Gradient-boosted trees tend to exhibit significantly improved classification performance over ordinary random forests because they have several hyperparameters that can be fine-tuned using cross-validation methods. Training the algorithm to identify delegation and constraint in the 8,417 EU provisions in the training data involved the following steps: (1) text pre-processing; (2) conversion of text into a document-term matrix; (3) algorithm training and fine-tuning via cross-validation; and (4) performance assessment on the test data. The text pre-processing stage involved standardizing the text of each provision such that only the words (or parts of words) with the highest amount of useful information are retained (Denny and Spirling, 2018; Gentzkow, Kelly, and Taddy, 2017; Grimmer and Stewart, 2013)). Prior to analysis, the text of each provision for training was pre-processed using the *quanteda* package in **R**. Provisions were tokenized into unigrams and bigrams using the “tokens” function and special characters, punctuation and stopwords were removed using the “tokens_select” function. The cleaned, tokenized texts were then transformed into a document-feature matrix (DFM) using the “dfm” function with options for removing stop words, stemming and removing punctuation added resulting in a very sparse 8,417 x 79,088 DFM.

Because tree-based methods are sensitive to poorly predictive terms, automatic sparsity reduction using the “dfm_trim” function with a 99% threshold was imposed on the DFM, yielding a 8,417 x 1,278 DFM used for algorithm training and performance assessment. Final models were selected using a process of hyperparameter tuning with five-fold cross-validation. Details regarding cross validation and training with gradient boosted trees can be found in the manual of the *XGBoost* **R** package: https://cran.r-project.org/web/packages/xgboost/xgboost.pdf.

As Table 2 demonstrates there were significant class imbalances in the training data. These were handled via a process of re-weighting positive labels using a positive predictive- weight tuning parameter, namely, the fraction of the negative over positive examples in the training data. The training process then involved randomly selecting a training and test set. We opted for a 90/10 train/test split when training the delegation (authority) classifier and a 70/30 split for each of the constraint classifiers to maximize the information available to the classifiers for training and to ensure that there are enough positive observations available to assess classifier performance in a reliable manner. We split the data differently for delegation and constraint as a consequence of the relative lack of positive examples for each of the constraint categories as can be seen from Tables 7 and 8.

Category	Provisions	Percent
Delegation	1296	15.4%
Rule-making Requirements	705	8.4%
Time Limits	89	1.1%
Consultation Requirements	63	0.8%
Appeals Procedures	35	0.5%
Reporting Requirements	39	0.5%
Spending Limits	33	0.4%
Executive Action Required	19	0.2%
Legislative Action Required	12	0.2%
Executive Action Possible	11	0.2%
Exemptions	23	0.03%
Public Hearings	1	0.00%
Legislative Action Possible	1	0.00%

Table 7: Member state training data statistics: total number and proportion of provisions in each delegation and constraint category

Category	Provisions	Percent
Delegation	528	6.3%
Rule-making Requirements	232	2.7%
Time Limits	25	0.3%
Consultation Requirements	108	1.3%
Appeals Procedures	0	0%
Reporting Requirements	40	0.5%
Spending Limits	10	0.1%
Executive Action Required	45	0.5%
Legislative Action Required	10	0.1%
Executive Action Possible	244	2.9%
Exemptions	37	0.4%
Public Hearings	22	0.3%
Legislative Action Possible	50	0.6%

Table 8: European Commission training data statistics: total number and proportion of provisions in each delegation and constraint category

Model training involves prediction of each of each of delegation and constraint categories for which there was sufficient training data available to evaluate algorithm performance using only the words contained in the document term matrix. This is accomplished through growing multiple trees via an iterative loss minimization process using an objective function, $\mathcal{O}(\theta)$, which is comprised of a logistic regression loss function $\mathcal{L}(\theta)$ of the tree parameters θ and a regularization term, $\gamma(f_k)$, which is a function of the number of k trees grown where each tree is represented by a function $f_k \in \mathcal{F}$ in the function space $\mathcal{F}$ of all possible trees:

$$\mathcal{O}(\theta) = \mathcal{L}(\theta) + \sum_{k=1}^{K} \gamma(f_k) = \sum_{i=1}^{T} l(c_i, \hat{c}_i^p) + \sum_{k=1}^{K} \gamma(f_k)$$

Figure 6 presents the top portion of one tree grown as part of the training process. The relevant parts of the tree are the terms which can be found right below the tree number. This sample tree is one of several that were grown as part of the training process and the term "member_state" is at the top of the tree suggesting it is an important term for classifying delegation to member states. Trees are grown according to terms that provide the most information gain for classifying the training label that the tree is presented with. Thus the importance of terms in a tree flows from most important at the top to least important at the bottom.

The goal of training is to minimize $\mathcal{O}(\theta)$ by simultaneously accounting for the difference between the true and predicted classification of each provision in the training data (c_i and $\hat{c}_i^p$) and the regularization term $\sum_{k=1}^{K} \gamma(f_k)$ which prevents overfitting of the model.

An important part of the training process involved hyper-parameter tuning using

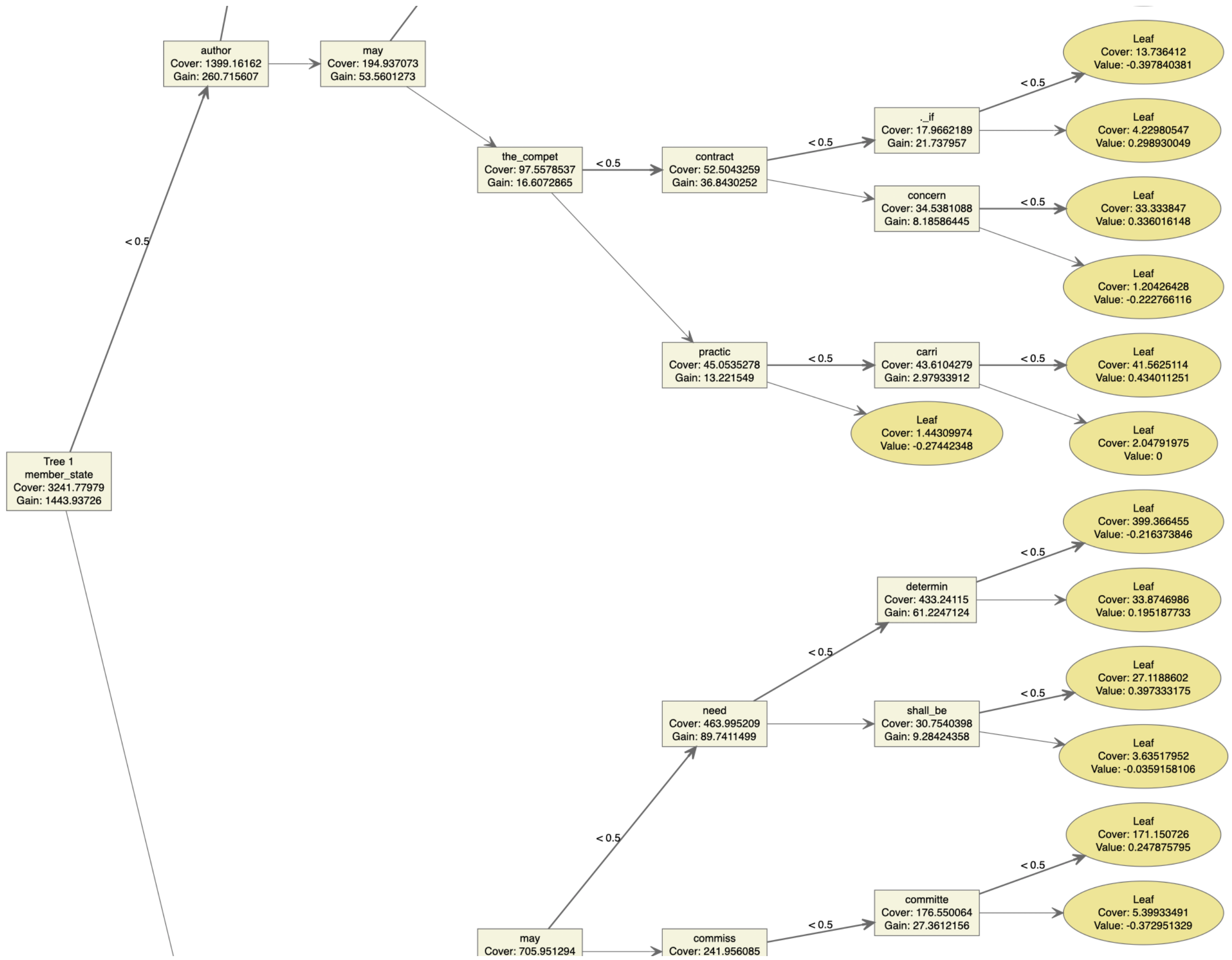


Figure 6: A few branches from one of the multiple decision trees grown by the member state delegation GBT classifiers. Trees in this context are grown using terms in each of the provisions starting with the most important provision at the primary (top) node of the tree. Each of the nodes in this tree contains information about the cover and gain of a particular term. Cover is an estimate of the relative number of observations affected by a split on the feature while gain is the information gain achieved by splitting on the feature.

5-fold cross-validation on the training data to select the model with the minimum average cross-validated test error as defined by the objective function in equation.

The training and test error of the models trained via the cross validation procedure are below. The final model selected for making predictions in the larger database of EU legislation contains the minimum test error denoted by the dotted line.

7 Estimation of delegation ratios, constraint ratios and discretion indices for all EU legislation

Labeling for each provision was conducted using probabilities estimated by each classifier in the following manner. For the j^{th} provision, it was labeled as delegating authority D_j if the probability of delegation given the words X_j was greater than 50% :

$$D_j = \begin{cases} 1 & \text{if } p(D_j|X_j) > 0.50 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The estimated delegation ratio $\hat{\Delta}_i$ for the i^{th} law is then:

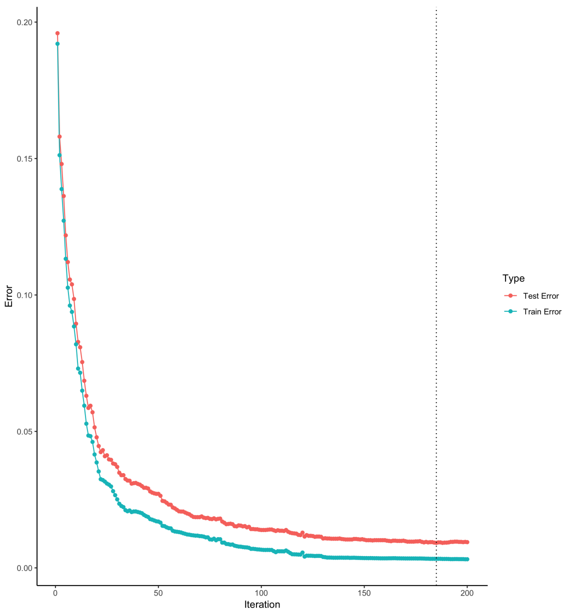
$$\hat{\Delta}_i = \sum_{j \in J_i} \frac{D_{ji}}{J_i}$$

where $\hat{\Delta}_i$ is simply the % of provisions in each law delegating authority to national administrations.

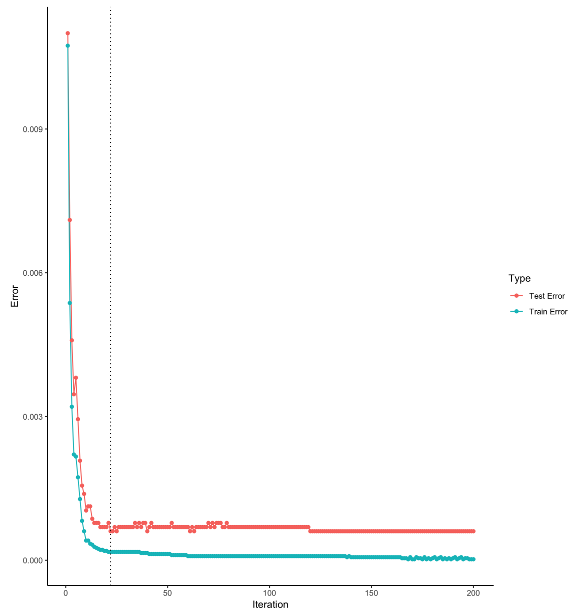
Estimation of the constraint ratio for each piece of legislation was slightly more complicated. For each j^{th} provision, we applied each of the $k = 8$ constraint classifiers to determine whether either of the constraint categories were present using the probabilities estimated by the classifier.

Figure 7: Average 5-fold, cross-validated training and test error for each GBT gradient descent iteration, by classifier. The optimal classifier chosen had the lowest RMSE test error.

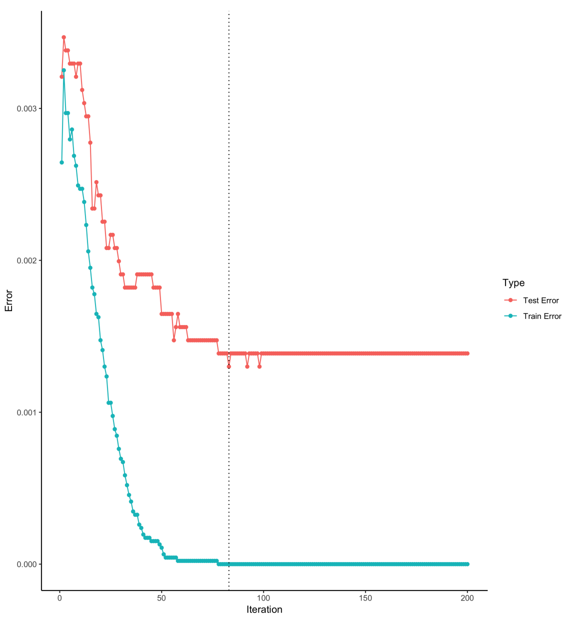
(a) Rule-making Requirements



(b) Time limits



(c) Reporting Requirements



(d) Consultation requirements

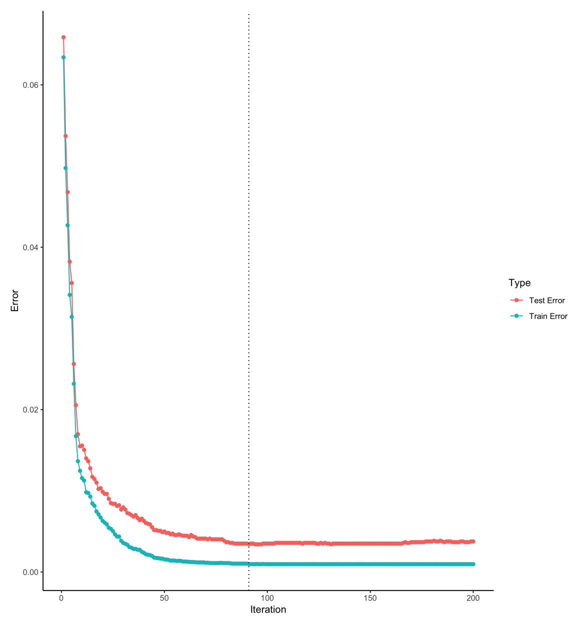
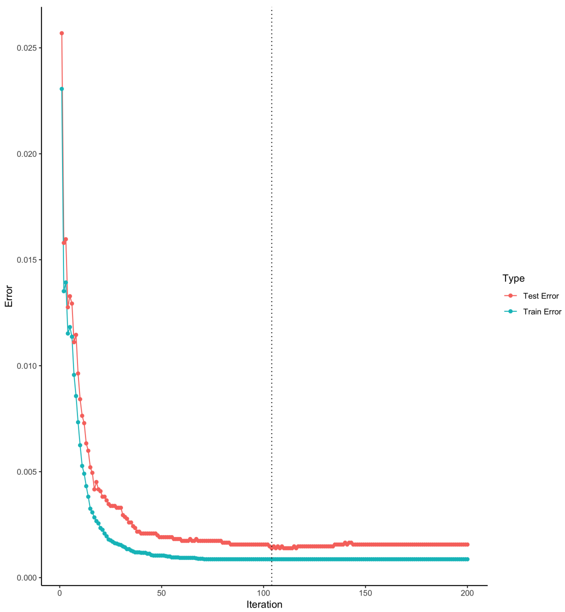
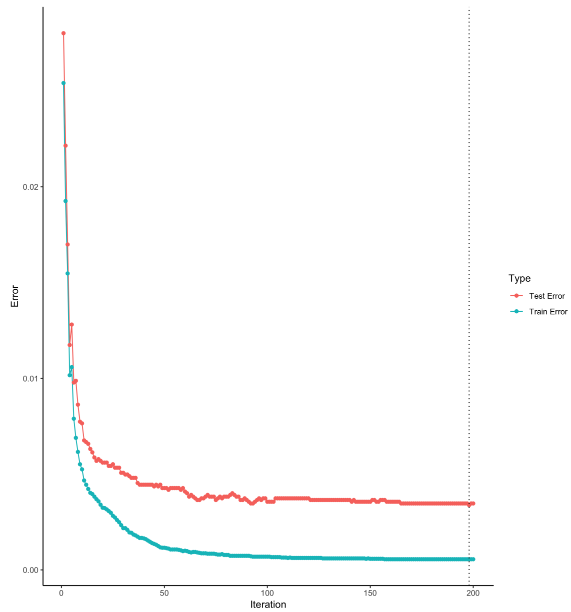


Figure 8: Average 5-fold, cross-validated training and test error for each GBT gradient descent iteration, by classifier. The optimal classifier chosen had the lowest RMSE test error.

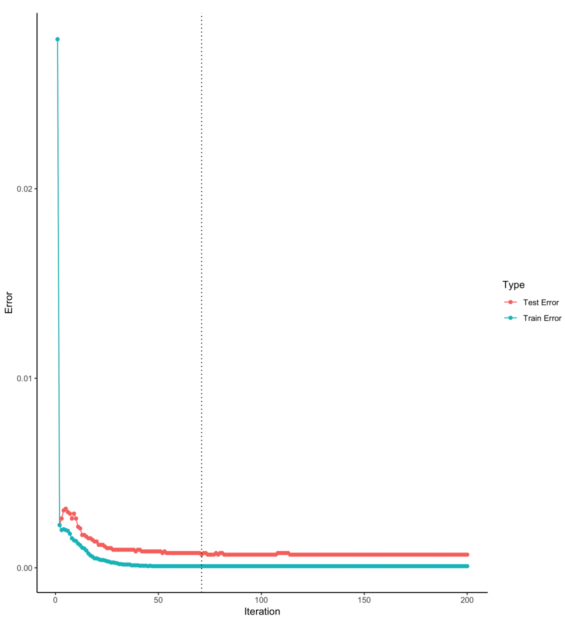
(a) Appeals Procedures



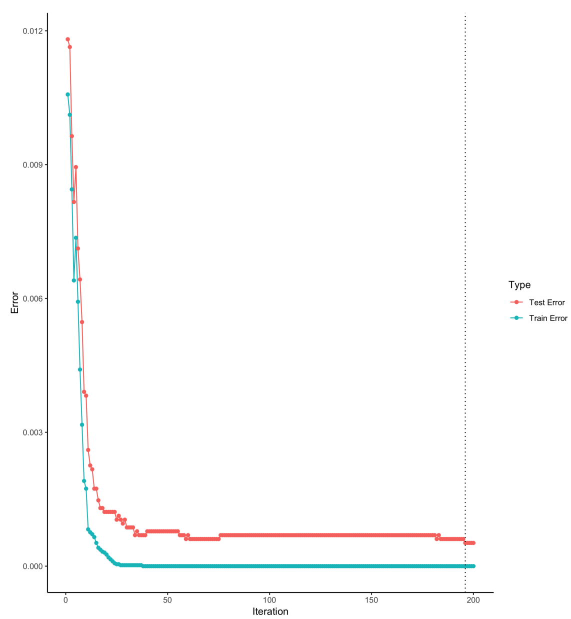
(b) Executive Action Required



(c) Spending Limits



(d) Executive Action Possible



$$C_{kj} = \begin{cases} 1 & \text{if } p(C_{kj}|X_j) > 0.50 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The estimated constraint ratio for each law was then computed according to whether any of the provisions in the i^{th} law contained at least one constraint category:

$$\hat{\mathcal{C}}_i = \frac{C_i}{12}$$

Where C_i is

$$C_i = \sum_{k=1}^{8} \mathbb{I}\left[\sum_{j \in J_i} C_{kj} \geq 1\right]$$

Finally, the estimated discretion index $\hat{\delta}_i$ is calculated by combining the estimated delegation and constraint ratios:

$$\hat{\delta}_i = \hat{\Delta}_i - [\hat{\mathcal{C}}_i \times \hat{\Delta}_i]$$

8 Further details regarding the discontinuity test

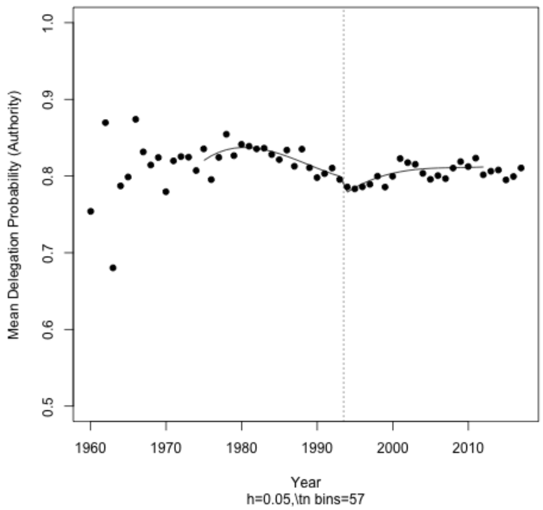
The discontinuity test is implemented using the predicted probabilities of delegation and constraint for each of the delegation and constraint categories at the article level produced by the trained gradient boosted tree model $P(D = 1|X)$. Using the Imbens–Kalyanaraman optimal bandwidth we estimate 9 local linear regressions of the following form:

$$P(D = 1|X)_i = \alpha + \beta_1 1[Y_i > 1993]_i + f(Y_i) + \eta_i$$

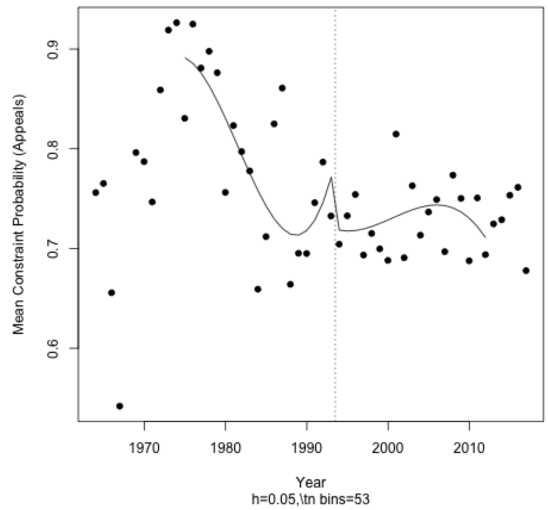
In the equation above $1[Y_i > 1993]_i$ is a dummy variable which is 1 for the out of sample years (1994–2017), $f(Y_i)$ is a function of the year estimated using the triangular kernel and $P(D = 1|X)_i$ is the estimated probability of the positive class label for each of the 9 delegation and constraint categories. A classifer is said to have "failed" the discontinuity test if $|\beta_1| > 1$. Plots of local linear regressions from each of the delegation and constraint categories show that each of the classifiers pass the discontinuity test.

Figure 9: Member state discontinuity tests for each of the delegation and constraint categories

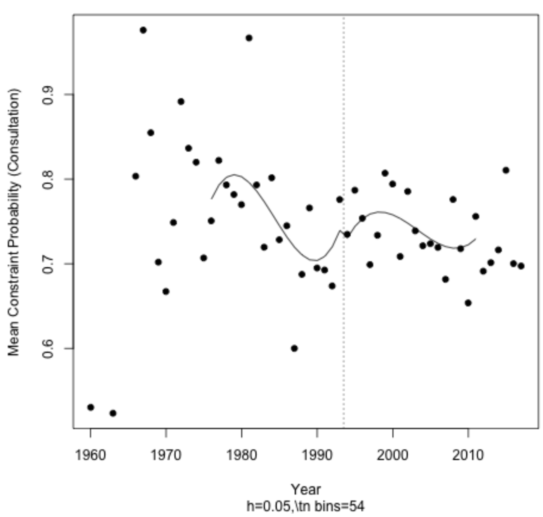
(a) Delegation



(b) Appeals Procedures



(c) Consultation Requirements



(d) Executive Action Required

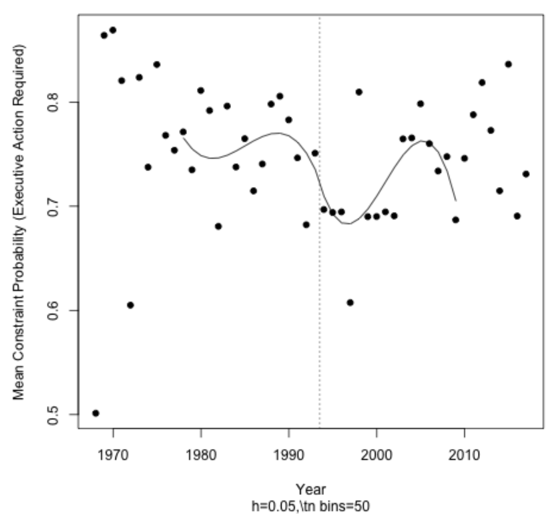
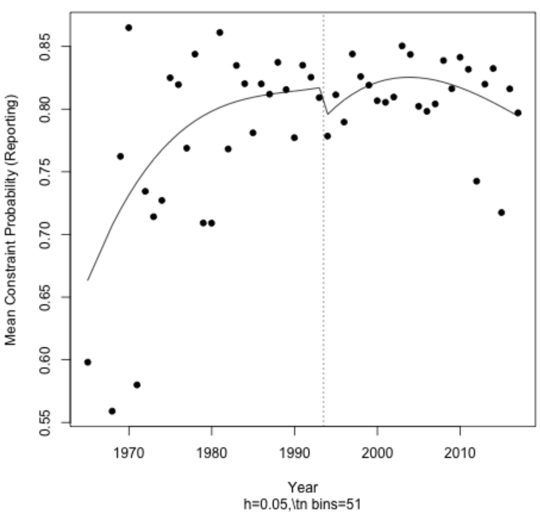
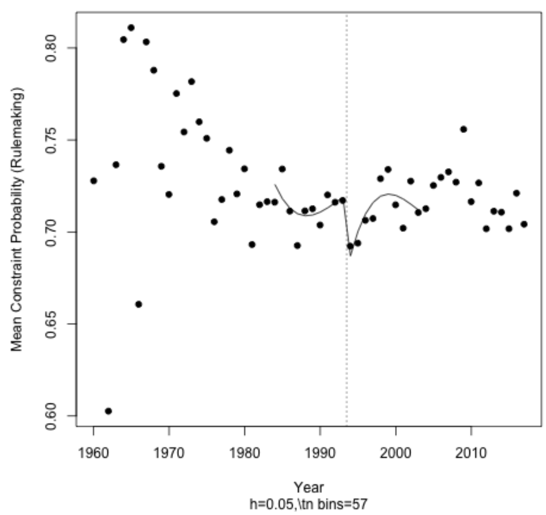


Figure 10: Member state discontinuity tests for each of the delegation and constraint categories

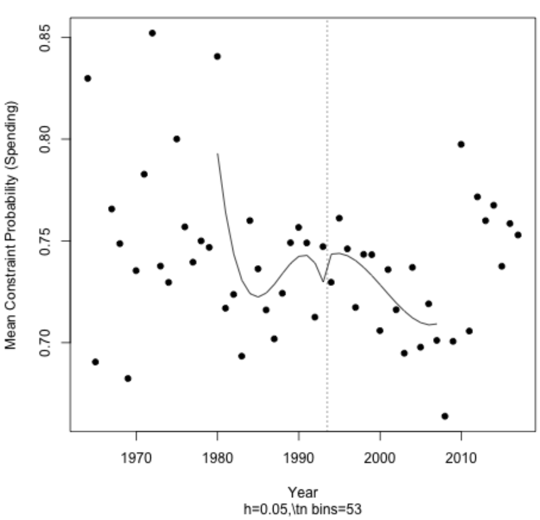
(a) Reporting Requirements



(b) Rulemaking Requirements



(c) Spending Limits



(d) Time Limits

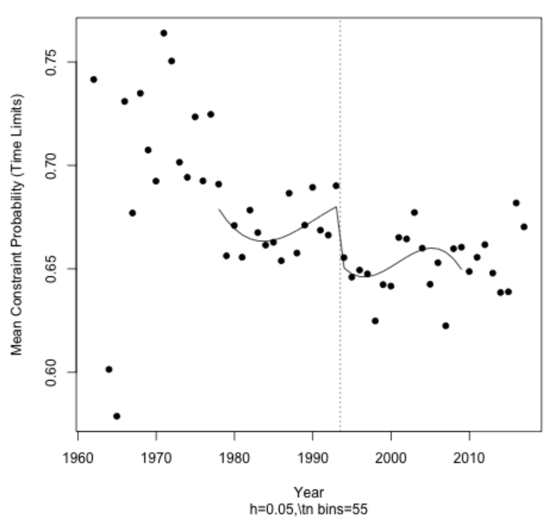
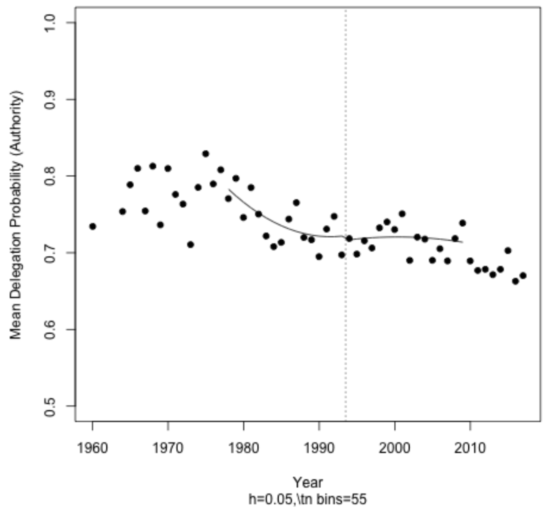
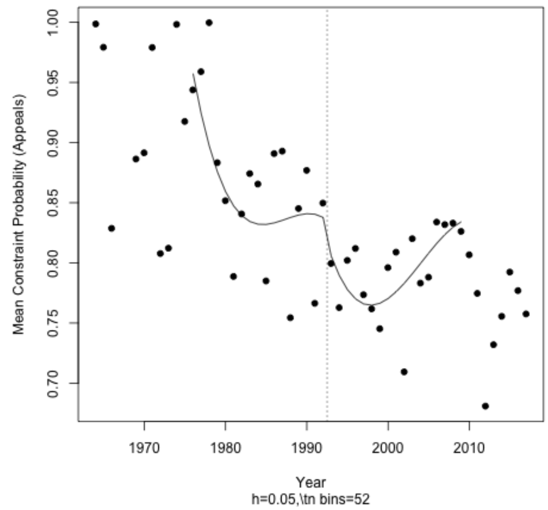


Figure 11: European commission discontinuity tests for each of the delegation and constraint categories

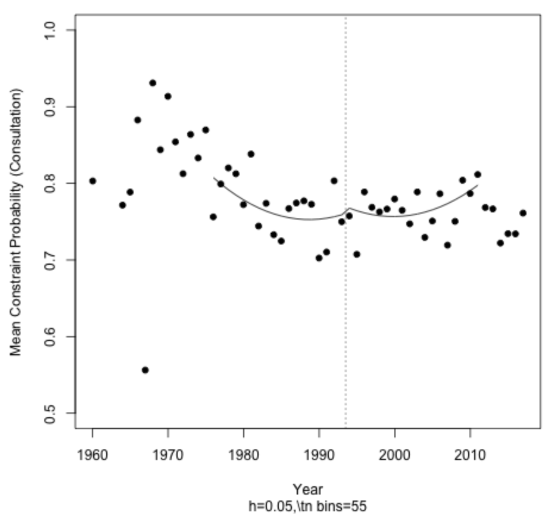
(a) Delegation



(b) Appeals Procedures



(c) Consultation Requirements



(d) Executive Action Possible

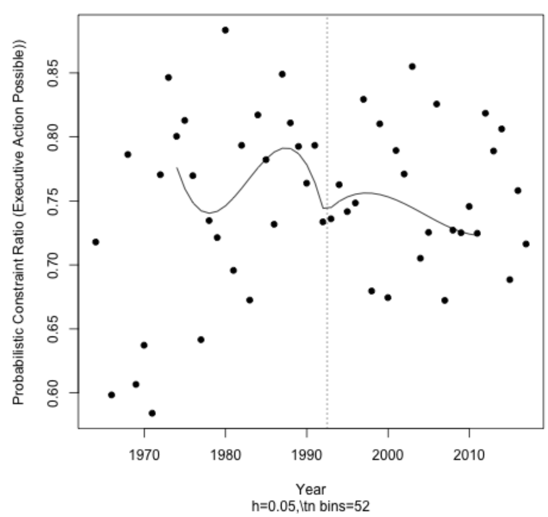
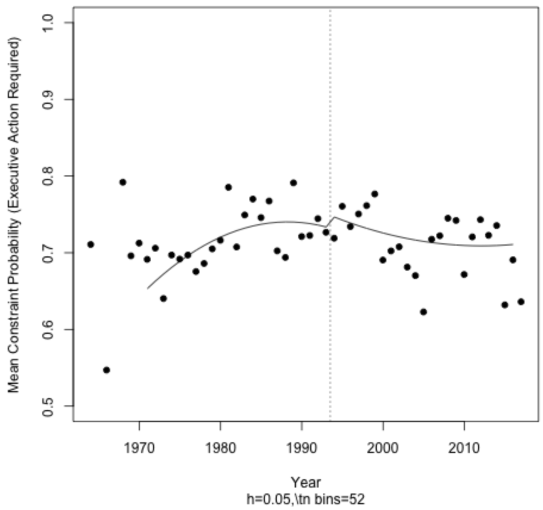
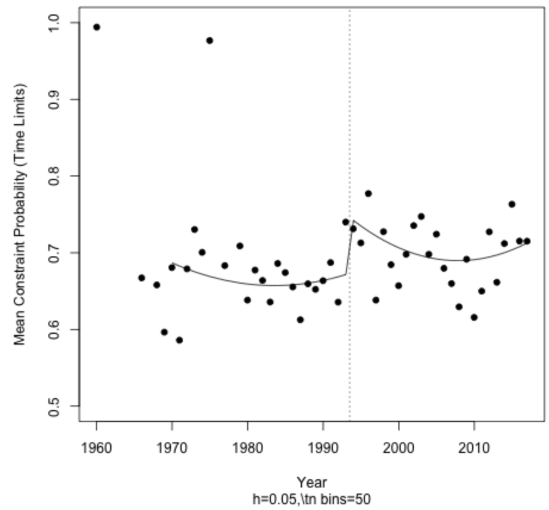


Figure 12: European commission discontinuity tests for each of the delegation and constraint categories

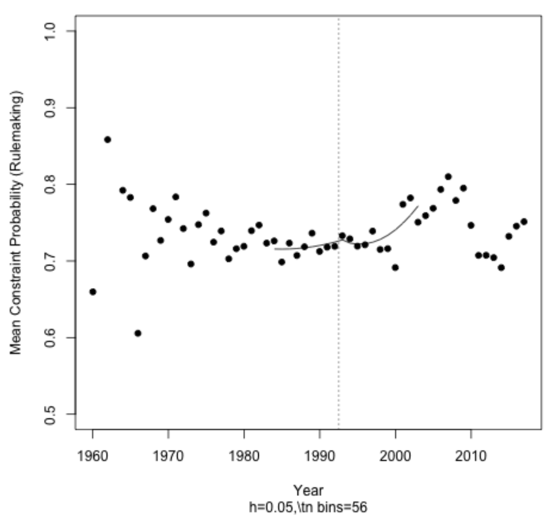
(a) Executive Action Required



(b) Time Limits



(c) Rulemaking Requirements



	Non-Amending Legislation	Amending Legislation	Difference	P Value	DF
	Mean Predicted Delegation Probabilities $E[P(Delegation = 1 \vert Text)]$				
Pr(Delegation)	0.812	0.806	0.006	0.005***	9264
	Mean Predicted Constraint Probabilities $E[P(Constraint = 1 \vert Text)]$				
Pr(Time Limit)	0.664	0.659	0.005	0.407	968
Pr(Spending)	0.733	0.749	-0.016	0.067*	508
Pr(Rulemaking)	0.719	0.717	0.001	0.697	3603
Pr(Reporting)	0.814	0.799	0.015	0.162	349
Pr(Exec. Action Req)	0.743	0.745	-0.002	0.921	124
Pr(Cons.)	0.743	0.724	0.020	0.139	209
Pr(Appeals)	0.744	0.728	0.017	0.139	315

Table 9: Member state delegation and constraint classifier confidence estimates across out of sample provisions between amending and non-amending EU legislation. Difference in p-values from a two sample t-test.

9 Consistency of predictions across amending vs. non-amending legislation

Another potential issue is related to the fact that our training data is comprised primarily of major laws is the possibility that the classifier will not be able to handle legislation which differs significantly, in terms of structure or language, from major legislation. One example of this is amending legislation, which is structured differently from non-amending legislation and as a result may yield different performance results.

While we cannot directly assess performance for our training sample of amending vs. non-amending legislation because is insufficient "ground truth" human coding to compare the predictions to, we can explore the extent to which the classifier is

	Non-Amending Legislation	Amending Legislation	Difference	P Value	DF
	Mean Predicted Delegation Probabilities $E[P(Delegation = 1\|Text)]$				
Pr(Delegation)	0.717	0.722	-0.004	0.220	3311
	Mean Predicted Constraint Probabilities $E[P(Constraint = 1\|Text)]$				
Pr(Time Limit Constraint)	0.688	0.712	-0.023	0.139	114
Pr(Public Hearings)	0.781	0.747	0.034	0.131	72
Pr(Rulemaking)	0.744	0.733	0.012	0.017***	1432
Pr(Reporting Constraint)	0.644	0.644	-0.000	0.153	1918
Pr(Exec. Action Possible)	0.820	0.827	-0.007	0.377	805
Pr(Exec. Action Req Constraint)	0.720	0.738	-0.018	0.099*	270
Pr(Cons. Constraint)	0.769	0.755	0.014	0.180	380
Pr(Leg. Act. Constraint)	0.782	0.732	0.050	0.003***	131

Table 10: European Commission delegation and constraint classifier confidence estimates across out of sample provisions between amending and non-amending EU legislation. Difference in p-values from a two sample t-test.

"confident" in identifying a positive class label using predicted probabilities as a proxy for out of sample classifier performance since different mean class probabilities across different types of legislation would suggest that the classifiers have greater difficulty with predicting positive class labels for one type of legislation.

Tables 9 and 10 contain estimates of the average probability of the positively labeled categories across each category among amending vs. non-amending legislation. Formally, for delegation or constraint category $C_k; k = \{1, \cdots, N\}$, amending legislation articles $i = \{1, \cdots, A_1\}$, non-amending legislation articles $j = \{1, \cdots, A_2\}$ and terms in article T_p , classifier "certainty" for **amending** legislation in each delegation and constraint category is estimated as:

$$\frac{1}{A_1}\sum_{i=1}^{A_1} P(C_k = 1|T_p \in \mathcal{L}_a) \tag{3}$$

where $\mathcal{L}_a$ is the set of all EU legislation classified by us $\mathcal{L}$ that is amending.Classifier "certainty" for **non-amending** legislation in each delegation and constraint category is estimated as:

$$\frac{1}{A_2}\sum_{i=1}^{A_2} P(C_k = 1|T_p \in \mathcal{L}_{-a}) \tag{4}$$

References

Athey, Susan, Julie Tibshirani, Stefan Wager et al. 2016. "Solving heterogeneous estimating equations with gradient forests." *arXiv preprint arXiv:1610.01271*.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. "Productivity and selection of human capital with machine learning." *American Economic Review* 106 (5): 124–27.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM pp. 785–794.

Denny, Matthew J, and Arthur Spirling. 2018. "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it." *Political Analysis* 26 (2): 168–189.

Epstein, David, and Sharyn O'Halloran. 1999. *Delegating powers: A transaction cost politics approach to policy making under separate powers*. Cambridge University Press.

Franchino, Fabio. 2004. "Delegating Powers in the European Community." *British Journal of Political Science* 34 (2): 269–293.

Franchino, Fabio. 2007. *The Powers of the Union: Delegation in the EU*. Cambridge University Press.

Gentzkow, Matthew, Bryan T Kelly, and Matt Taddy. 2017. Text as data. Technical report National Bureau of Economic Research.

Grimmer, Justin, and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21 (3): 267–297.

Hinton, Geoffrey E, and Ruslan R Salakhutdinov. 2009. "Replicated softmax: an undirected topic model." In *Advances in neural information processing systems*. pp. 1607–1614.

Ikonomakis, M, S Kotsiantis, and V Tampakas. 2005. "Text classification using machine learning techniques." *WSEAS transactions on computers* 4 (8): 966–974.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human decisions and machine predictions." *The Quarterly Journal of Economics* 133 (1): 237–293.

Korde, Vandana, and C Namrata Mahender. 2012. "Text classification and classifiers: A survey." *International Journal of Artificial Intelligence & Applications* 3 (2): 85.