Regulation & Governance

Regulation & Governance (2023) doi:10.1111/rego.12543

Concepts and measures of bureaucratic constraints in European Union laws from hand-coding to machine-learning

Fabio Franchino

Department of Social and Political Studies, Università degli Studi di Milano, Milan, Italy

Marta Migliorati

Institute for European Studies, University of Malta, Msida, Malta

Giovanni Pagano and Valerio Vignoli

Department of Social and Political Studies, Università degli Studi di Milano, Milan, Italy

Abstract

Scholars employ two main measures of the executive constraints embedded in European Union laws: one is based on the *variation* in the use of different types of restrictions, and the second is based on the *frequency* of such use. They reflect two alternative conceptualizations of bureaucratic control. We label them, respectively, as the "toolbox perspective" and the "design perspective". We illustrate that the constraint frequency measure poses fewer validity problems in estimating legislators' intent to constrain implementation and tends to produce less severe measurement errors. We then evaluate the performance in estimating constraint variation of a recent computational application and identify potential drawbacks of automated learning from hand-coded provisions. We lastly introduce a skeletal framework for a machine-learning approach based on the syntactic structures employed by legislators that could improve the performance of this innovative technique.

Keywords: bureaucratic constraints, conceptualization, EU law, machine-learning, measurement.

1. Introduction

Pick up any law and you read a procedure or a prescription of some sort. They are ubiquitous. These constraints are the nuts and bolts of legislative design, and they are core components of the formal executive discretion of implementing agents. Their use reflects the important trade-offs legislators make between the benefits of delegation and the risks of agency drift. Yet, measuring the legislative intent to delimit the room for maneuver of implementers is arduous, and doing so for thousands of pieces of legislation is treacherous. Scholars of legislative politics in the European Union (EU) have adopted different measurement strategies, some of them inspired by the literature on congressional politics in the United States. Yet, the choices scholars make and the assumptions that underlie these decisions are hardly discussed.

In this article, we carry out such an examination, especially in light of ever more sophisticated procedures to extract estimates of constraints from legal texts that can scale up to large corpora in an efficient and (presumably) valid manner. We first review the different procedures scholars have employed to measure the executive constraints embedded in EU laws. Next, we argue that these procedures reflect two different conceptualizations which we call the toolbox perspective and the design perspective on bureaucratic control. Scholars that adopt the former employ measures that are based on the *variation* in the use of different types of restrictions. Those that prefer the latter use measures that are based on the *frequency* of use of such constraints. Which of the two provides a more valid estimation of legislators' intent to constrain implementation? Employing the data from

Correspondence: Fabio Franchino, Department of Social and Political Studies, Università degli Studi di Milano, Via Conservatorio 7, 20122 Milan, Italy. Email: fabio.franchino@unimi.it

Conflict of interest: None declared.

Accepted for publication 9 June 2023.

Franchino (2007) and Migliorati (2021), we show how the constraint variation measure may exhibit greater problems of validity than the frequency one. Next, we examine Anastasopoulos and Bertelli's (2020) machine-learning method for estimating constraint variation. This innovation has considerable promise because of its flexibility and scalability, but, as we shall show, its performance invites caution and careful design. In the last section, we lay out the skeleton of a machine-learning approach that is based on the syntactic structures adopted by legislators to constrain implementation. An initial validity assessment indicates that this approach could improve the performance of this innovative technique.

2. The (lack of) debate on concepts and measures of bureaucratic control

Although several researchers have sought to meaningfully identify constraints imposed on implementers through legislation, there is hardly a consensus on the best way of measuring them, let alone a debate on the conceptual assumptions underlying the choice for a specific measurement strategy.[1] In a landmark study of delegation in the United States, Epstein and O'Halloran (1999: 99–106) canvass the extensive literature on congressional oversight and argue that US legislators have at their disposal 14 different categories of procedural constraints. They then construct a constraint ratio which is obtained by dividing the number of different *types* of constraints that are present in a given legislative act by 14.[2] In other words, the greater the *variety* of the types of constraints employed by the legislators in an act, the tighter the limitations imposed on implementers. This measure has been applied to EU law by Franchino (2004, 2007), and more recently by Ershova (2019) and Migliorati (2021)—the only difference being 12, rather than 14, possible categories of constraints. And, in perhaps the most innovative recent contribution to the field, it has been used by Anastasopoulos and Bertelli (2020) for training a supervised classifier and extracting a machine-learning measure of constraint from a legal text. Despite its popularity, is this constraint ratio a valid measure of legislators' intent to set the boundaries of implementation? This issue is hardly addressed and the assumptions that underlie these choices are not made explicit.

As we will show below, its validity should not be taken for granted. Indeed, scholars have veered away from this measurement strategy. Gastinger and Heldt (2022) initially follow the approach of Epstein and O'Halloran (1999) and Franchino (2004, 2007) but then they "distinguish between 'universal' constraints affecting the act of delegation as a whole […] and 'specific' constraints tied to corresponding provisions delegating authority" (Gastinger & Heldt, 2022: 545). For instance, the development cooperation instrument established by Regulation 233/2014 terminates in 2020. This time limit is relevant for the whole act, so accounting for its presence would be enough to gauge the legislative desire for control of implementation. On the other hand, the European Parliament and the Council can revoke the Commission's power to adopt delegated acts on four specific areas of cooperation in this law. Had the legislators used this constraint only for two areas, it would have meant a desire to give the bureaucracy greater room for maneuver—the implication being that we should also account for the *frequency* of use of constraints, even if they are of the same type. Similarly, Zbiral et al. (2022) prefer to distinguish between only five categories of constraints and count the *total* number of constraints in each legislative text, as do Franchino and Mariotto (2020).

Other scholars use different measurements, which however follow similar conceptual premises. Steunenberg and Toshkov (2009: 958) distinguish between closed statements in EU directives that prescribe "what a member state has to do (order) or cannot do (ban)" and open statements that "allow for a choice by the implementing authorities in the member states." Gastinger and Dür (2021) create an index of the discretion granted to the Commission in EU external relations (see also Gastinger & Adriaensen, 2019). After determining five dimensions influencing its discretion, they assign weights to the five indicators according to the frequency with which they appear in the index.[3]

These scholars focus to a greater extent on the *frequency* of use of constraints and they do not assume that there is a pre-determined toolbox to which legislators recur when they design a law. Rather, legislators have carte blanche as to the imposition of limitations to any delegating provision, should they find it necessary. Is this however a better conceptualization of control? Is this a more valid measure of legislative intentions? These scholars too hardly address these questions.

The absence of consensus on how constraints should be measured reflects implicit differences on how best to conceptualize the control problem legislators face. Some scholars find it reasonable to think that legislators rely on a predetermined set of control tools. Yet, it is also plausible to argue that they are instead free to constrain the executive actors as much (or little) as they want, depending on the features of the measure at hand. Without a comprehensive discussion and a thorough comparison of these concepts and measures, even estimates obtained from the analysis of a vast number of legislative acts through machine-learning techniques may suffer from serious validity problems. We begin this discussion in the next section.

3. Two concepts of bureaucratic control

Most laws contain a set of provisions that confer substantive policy authority to an implementing institution, subject to some constraints. For instance, in the early days of the European Economic Community, the Council of the EU adopted a regulation that abolished discriminatory practices.[4] Carriers used to charge different transport rates and set different conditions for the carriage of the same goods over the same route on the grounds of the country of origin or destination of the goods. To facilitate inspection, the Council established that a transport document had to be attached to each consignment of goods by the first of July 1961. It then gave the Commission the power to issue a regulation to postpone this deadline, subject to consulting the Council. In order words, legislators attached a requirement of consultation to the delegation of a policy prerogative. This process of constraining executive action can be conceptualized in two different ways, which we call the toolbox perspective and the design perspective on bureaucratic control.

3.1. Control toolbox and the constraint variation ratio

According to this view, control over the implementers is a process whereby legislators have at their disposal a limited set of predefined and fixed control tools, that is, *types* of constraints. When a tool is used (a constraint is inserted into a law), it is assumed to be applied to the entire law even though the specific constraint may be associated with only one policy prerogative granted to the implementer. The greater the variety of tools used, the greater the control exercised on the bureaucracy. If legislators empty their toolbox and use, even sparingly, the entire set of tools, there is full control and the implementer has no policy leeway. In the context of EU law, this concept is operationalized by a measure of the *variation* in the use of constraints. We could call it a constraint variation ratio: the sum of the *types* of constraints that are present in a given law divided by 12, that is, $\frac{T_i}{12}$ where T_i is the number of types of constraint used in law i and 12 is the total number of predefined categories of constraint (Franchino, 2004, 2007).[5] A law with a constraint ratio of 0.5 indicates that legislators have laid out six different *types* of constraint, one with a ratio of 1 contains all 12 categories. Accordingly, the executive agent is more constrained in the latter circumstance.

Is this ratio a valid measure of legislators' intent to set the boundaries of implementation? Consider the example above about transport regulation. This law also conferred upon the supranational executive the power to impose penalties on carriers that maintained discriminatory practices. However, before taking such a decision, the Commission had to consult all the member states concerned. States had to reply within 2 months and could seek the opinion of an independent national body. This is a second consultation requirement, attached to a different substantive policy authority. However, Epstein and O'Halloran's (1999) constraint ratio would fail to capture this additional use because the measure records only the presence of a *type* of constraint rather than the frequency of its use. In other words, this ratio underestimates the employment of procedural constraints by legislators and, plausibly, their intent to restrain executive discretion.[6]

Imagine another regulation on discriminatory practices that is identical to the one we just discussed, except for the fact that the Commission does *not* need to consult the member states before imposing penalties. This second measure gives the Commission greater leeway but this difference would not be captured by this constraint ratio because the *presence* of a consultation requirement has already been recorded by the obligation to consult the Council before deciding to postpone the deadline for the adoption of the transport document. Ultimately, two laws that exhibit the same variety of constraints but employ them with different intensities are undistinguishable, even though legislators have displayed different preferences for control in the *design* of the two acts. This conceptualization measures the presence of bureaucratic constraints at the level of the legislative act and ignores that constraints are in reality associated with specific provisions within the act. It assumes that once an obligation is present in a given law, it applies to all the provisions within.

3.2. Control design and the constraint frequency ratio

According to this perspective, control is an activity with no fixed or predefined limits to the legislators' freedom to design and utilize procedural constraints. Since constraints are associated with specific provisions within a given law, we should not assume that their use extends to the whole legal document. This conceptualization of control also accounts for the fact that legislators may associate similar constraints to substantially different policy prerogatives that are conferred upon the implementers. This concept can be operationalized by a measure of the *frequency* in the use of constraints; in other words, a constraint frequency ratio such as $\frac{C_i}{P_i}$ where C_i is the total number of constraints (of any type) used in law i and P_i is the number of major provisions in such law.

How do the two concepts of control differ? To compare them, we use the datasets of Franchino (2007) and Migliorati (2021) of major EU laws that have been adopted, respectively, between 1958 and 1993 and between 1985 and 2016. These scholars employ the toolbox perspective on control and, therefore, the constraint variation ratio $\frac{T_i}{12}$. For the design perspective, we have counted the frequency of use of constraints in these legislative acts and computed the constraint frequency ratio $\frac{C_i}{P_i}$.

In the EU, legislators tend to rely on either the Commission or national authorities for policy implementation.[7] Figure 1 presents two scatter plots of the two different constraint ratios applied to the two executive actors.[8] Figure A2 in the Supporting Information includes two additional scatter plots of the ratio between these two measures over time. These plots suggest several interesting observations. First, in the case of Commission implementation, the two constraint ratios take the same values in 41% of the laws. In the case of national implementation, the values are the same in only 16% of the laws. The Pearson's r correlation coefficients between the two measures are 0.70 for the Commission constraint ratios and 0.44 for the national constraint ratios (p-values <0.01). These coefficients are reassuringly positive but the moderate magnitude suggests that there are nontrivial differences between the two concepts, especially in the case of national implementation. And these differences appear to become more important over time (see Fig. A2). Indeed, the correlation coefficients weaken in the more recent dataset.[9]

Second, we illustrated earlier an example where the variation measure plausibly underestimates the use of constraints. These are the laws located in the bottom right corners of the plots in Figure 1. The variation ratio is lower than the frequency ratio in 3% of the laws in the case of Commission constraints and 17% in the case of national constraints, and these percentages decrease in the more recent dataset of Migliorati (2021) (see Fig. A1 and the observations below the horizontal line in Fig. A2). These laws are characterized by the extensive use of a few types of constraints. For instance, Directives 77/453 and 80/155 on the coordination of national provisions concerning the activities of nurses and midwives contained several detailed rules, such as the minimum number of years of general school education and hours of training that would be required for their qualifications to be recognized in other member states. These constraints fall under a single category of rule-making requirements. The constraint variation ratio of these measures is, therefore, only 0.08 (=1/12). However, in light of the several detailed rules inserted in these acts and their rather short length, the constraint frequency ratio is between 0.36 and 0.38.

The third interesting observation that can be derived from Figure 1 is that quite a few laws score high in terms of constraint variation ratio and low in terms of constraint frequency ratio. The former exceeds the latter in 56% of the observations in the case of Commission constraints and 67% in the case of national constraints. Moreover, the incidence of these cases is higher in the more recent dataset of Migliorati (2021) (see Fig. A1 and the observations above the horizontal line in Fig. A2). These are the laws located in the top left corners of the panels in Figure 1. Consider a case where mainly the Commission is involved in implementation. Regulation 1316/2012 establishing a fund for infrastructure investments lays out seven different types of constraints on the supranational executive, but they are used very sparingly, about once each. The constraint ratio is, therefore, 0.58 for the variation measure but, given the non-trivial length of the law, it is only 0.08 for the frequency measure (further examples of laws exhibiting large disparities between the two ratios, both in the case of national and supranational implementation, are presented in the Supporting Information).

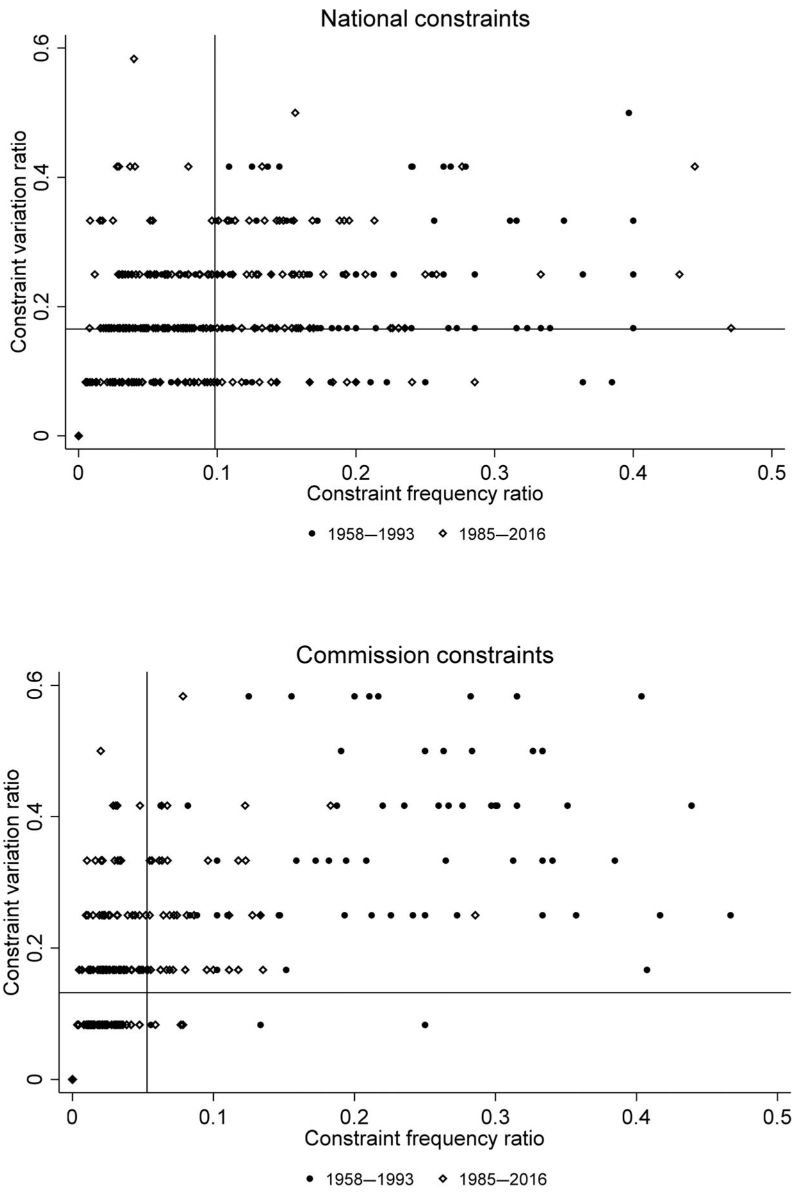


Figure 1 Variation versus frequency measures of constraints. *Note*: Franchino (2007) (158 major laws, 1958–1993) and Migliorati (2021) (276 major laws, 1985–2016). Duplicated measures are dropped. Vertical and horizontal lines at the mean values.

Lastly, if we subtract the frequency ratio from the variation ratio, the correlation between this difference and the number of major provisions in a given measure is 0.66 for national constraints and 0.59 for Commission constraints (p-value <0.01), indicating that variation ratios tend to exceed frequency ratios in longer laws, and vice versa.

In sum, the two ratios differ under four main circumstances. First, when legislators employ sparingly several types of constraints, the variation ratio tends to exceed the frequency ratio, other things being equal. Second, the opposite occurs when legislators employ intensively few types of constraints. The frequency ratio tends to exceed the variation ratio. Third, the higher the number of major provisions (i.e., the longer the law) the more likely the variation ratio exceeds the frequency ratio. The reason is straightforward. As laws become longer, legislators employ more constraints and a greater variety of constraints (see more on this below), but this activity has a larger impact on the variation measure because its denominator is fixed by the assumed upper limit of 12 control tools at the disposal of legislators, while in the frequency measure, the denominator increases with P_i, the length of the act. Lastly, this dynamic engenders the specular effect: the lower the number of major provisions (i.e., the shorter the law) the more likely the frequency ratio exceeds the variation ratio.

4. Toolbox versus design concepts and measures: A validity assessment

Against this backdrop, should we assume that there is a fixed set of control tools at the disposal of legislators, or should we account for the legislators' freedom to design and employ constraints at their discretion? Most importantly, which of these two measures better gauges the intent of legislators to limit bureaucratic authority?

Consider first the circumstances where the frequency measure exceeds the variation measure as in the bottom right corners of the panels in Figure 1. This occurs when few constraints are intensively used, particularly in short laws. In our view, the inability of the variation measure to capture the legislative desire to curb executive leeway when similar constraints are repeatedly employed in these short acts is problematic. Assume, for instance, that the Commission has been delegated the power to take substantive decisions in two provisions of a given law, subject to consent from a regulatory committee. This design is different from one where the Commission needs consent in only one provision since legislators have decided *not* to restrict the Commission's prerogative in the other circumstance. Because the toolbox conceptualization ignores the provision-specific use of constraints and pays attention only to the *types* of constraints that are assumed to apply to the whole legal document, it fails to capture this important difference. If the frequency measure is more valid in capturing legislative intent in these circumstances, this implies that current studies have underestimated especially the extent to which EU laws constrain national authorities (or at least legislators' desire to constrain them) given the higher proportion of laws where the variation ratio is lower than the frequency ratio.

Consider now the circumstances where the variation ratio exceeds the frequency ratio as in the top left corners of the panels in Figure 1: this happens when several *types* of constraints are sparingly used, particularly in long laws. Does the variation ratio overestimate legislators' intent to limit executive action in these circumstances? Or is it the frequency ratio that underestimates such intent? With laws that are made of hundreds of provisions, one could expect that it would be easier to come across a greater variety of constraints so that relatively high values of the variation ratio can be reached fast, given that only 12 types of constraints are categorized. The question remains as to whether the *single* use of a constraint gauges the increase in control across the *whole* act. Take, for instance, Directive 2013/36, discussed in the Supporting Information, which has 551 provisions. Should the fact that 7 out of the 12 possible types of constraints are present in this measure be the determinant of the intent to limit national discretion, or should it be the fact that they are used only 22 times across this very long directive? In other words, is legislative intent to constraint better captured by a ratio of 0.58—the highest value in the dataset—or by a ratio of 0.04, which is less than a standard deviation below the mean of the frequency ratio?

We are inclined to argue that the overestimation of legislative intent by the variation ratio is more likely than the underestimation by the frequency ratio. For two related reasons: First, longer laws are positively and significantly associated with both a greater variety and a greater use of constraints, for both the Commission and national administrations.[10] Second, the measurement of the variation ratio is coarser than that of the frequency ratio since it can take up to a maximum of 13 values.[11] In other words, the toolbox approach restricts the choice of legislators to 13 levels of control for *any* law (i.e., T_i from 0 to 12). It is like having a screwdriver that can only be used on a fixed set of screw sizes. Finding an additional constraint type, as is more probable in longer laws, produces a predetermined step increase of 0.08 (i.e., 1/12) for *any* law, regardless of the length of the document and the powers conferred upon the implementers. This marginal effect is not only fixed and uncalibrated to the length of the document but most of the time is also larger than the marginal effect on the frequency ratio of an additional constraint.

This latter ratio allows instead for more fine-grained and flexible measurement. Legislators have $P_i + 1$ levels of control they can choose from, since C_i can take any value between zero to P_i (the number of major provisions). This is not inconsequential given that P_i is greater than 12 in 92% of the laws in our dataset. Correspondingly, 9 times out of 10 the marginal effect on this ratio of adding a constraint (i.e., $1/P_i$) is smaller than 0.08. Moreover, P_i is, of course, *also* a matter of legislative choice, which makes the number of control levels to choose from infinite and the step increases highly fine-grained. In other words, legislators can *design* the screwdriver (i.e., control procedure) that best fits the screw size (i.e., policy problem) they are dealing with.

To sum up, the design perspective on control accounts for the fact that policymakers choose both C_i and P_i. In other words, it recognizes that the size of the increment in control from employing an additional constraint is a matter of legislative choice. Since these increments are most of the time smaller and more fine-grained than the fixed step increase of 0.08, we can plausibly argue that it is the variation ratio that is likely to overestimate the marginal increase in legislators' desire to exercise control in these circumstances.

This line of reasoning leads us to a final observation: measurement errors have more severe consequences for the variation ratio than the frequency one because of the lower granularity. If a coder incorrectly classifies a new type of constraint in a given act, the ratio increases by 0.08, regardless of the length of the law. For the frequency ratio, a wrong classification produces most of the time a smaller change in the value of the ratio.[12]

5. An assessment of the machine-learning method to estimating constraint variation from labeled provisions

The most innovative application to estimate the constraint variation ratio in EU laws is a supervised machine-learning method recently proposed by Anastasopoulos and Bertelli (2020). Machine-learning offers promising opportunities for developing methodologies that leverage data to improve the classification of legal provisions. The scalability of these techniques—the cost-effective and time-saving handling of large amounts of data—is a clear advantage over manual coding. In this section, we critically assess the performance of Anastasopoulos and Bertelli's (2020) method. In the next section, we propose a machine-learning approach that could address some of its drawbacks.

Anastasopoulos and Bertelli (2020) have trained, on Franchino's (2004, 2007) hand-coded provisions, 16 gradient-boosted decision tree classifiers for 10 categories of constraints.[13]

To put it simply, these classifiers estimate the probability that a specific constraint is present in the legal provisions. This inference is made by examining the distribution of words in a training set, which consists of a sample of hand-coded provisions from Franchino's dataset. The validity of this estimation is then assessed using a test set, which is a smaller subset of provisions in the same dataset, and includes the associated labels for the various categories of constraints.

Only five classifiers perform well[14]: three in the case of Commission implementation and two in the case of national implementation. The performance of three more classifiers is middling at best, while it is poor for the remaining eight. The classifiers of constraints to national implementation perform more poorly overall. We investigate these differences in detail in the Supporting Information. In a nutshell, we argue that poor performance is primarily due to (a) in some cases, the high similarity of the wording of the legal provisions that are assigned by the manual coder to *different* constraint categories; (b) in other circumstances, the high dissimilarity of the wording of legal provisions that are assigned by the manual coder to the *same* constraint category; and (c) the presence of two main implementation paths of EU law. Lastly, the effectiveness of the training process may be further impaired by (d) the limited number of examples for at least some categories of constraint among Franchino's (2004, 2007) hand-coded provisions.

For instance, the classifier of *public hearings* has correctly tagged all the provisions that include this constraint according to the manual coder, and about 7 out of 10 positively identified provisions are correctly assigned to this category. Despite its infrequent use, this constraint category has three distinct advantages: the legal phraseology is fairly consistently used across the different laws, such phraseology is distinct from the one used for other categories of constraints, and the constraint is also almost exclusively employed to limit the Commission's authority. Instead, in other circumstances, the legal provisions that the manual coder has assigned to a given constraint category (say, *executive action possible*) may be getting confused by the machine as provisions belonging to a different category (for instance, *executive action required*). As a result, the classifier of the latter category would lack precision (i.e., a high false positives count) since it would interpret a provision regulating a possible executive action as a requirement of executive action, given the high similarity in the wording. Distinguishing these two categories requires expert legal knowledge that may be hard to teach a machine using labeled provisions, especially if some constraints are not very frequent. A possible solution would be collapsing the two categories into one, which would have consequences for the measurement of the constraint variation ratio, while the measurement of the frequency ratio would be unaffected.

However, even with simpler or more common categories, the presence of two main implementation paths may create problems of precise classification in an approach based on labeled provisions. This procedure may find it hard to distinguish, for instance, a consultation requirement imposed on the Commission from one imposed on member states. The legal texts may be rather similar if both actors are referred to in both provisions. Compare, for instance, article 14.4 of Regulation 17/1962: "The Commission shall take decisions referred to in paragraph 3 after consultation with the competent authority of the Member State"; with article 14.6 of the same measure: "Member States shall, after consultation with the Commission, take the necessary measures". A machine-learning application that is based on the syntactic structure of legal provisions rather than merely word frequencies may better discriminate between these two sentences.

In case of high dissimilarity within a category, a classifier may fail to correctly tag the provisions that have been assigned by the manual coder to that category. This lack of recall (i.e., a high false negatives count) may be more severe if only a few provisions belong to a category. Surely, categories of more similarly worded constraints could be created to address this problem but such similarity may not exist because of the highly policy-specific nature of some constraints such as the *rule-making requirements*. Here too, a more fine-grained machine-learning approach may be helpful.

A final consideration pertains to the properties of the hand-coded legislation. Franchino (2007)—and Migliorati (2021), for that matter—collect only major laws. Moreover, their coding ignores supranational agencies, whose involvement in implementation has significantly increased over the past decades, and it does not distinguish between member states and national authorities in the case of national implementation. Surely, these drawbacks call for better coding but, as such, they may potentially engender bias in the out-of-sample performance of these classifiers. In the next section, we put forward a supervised machine-learning approach for identifying constraining legal provisions which could represent a first step to addressing some of the problematic issues highlighted here.

6. Machine-learning from labeled syntactic structures: A first step

Our approach leverages the syntactic structure of sentences in legal acts to identify provisions that are relevant to our research question. Relying on methods developed in natural language processing (NLP),[15] we parsed the sentences of a corpus of EU legal acts according to their lexical and syntactic structures and we extracted information on the subjects, the types of verbs used, and other relevant features. We then combined this information with that on the presence of actors and verbs of interest, obtained through two supervised machine-learning models that we developed. To identify relevant provisions, we matched the structure of each sentence against a set of "rules of extraction", which are templates that define what a constraining provision looks like. These rules are developed on the basis of expert knowledge of the linguistic expressions employed by legislators in drafting the acts and of European constitutional law (e.g., Schütze, 2018).

The need for legislators to employ highly structured, grammatically correct, and terminologically precise language to ensure the proper execution of their measure makes a syntax-based approach particularly appealing. Compared to machine-learning methods based on the so-called "bag-of-words" principle, that is, word frequencies across texts, this approach allows gaining a highly granular understanding of the patterns of constraints. In principle, it can be applied to any law, regardless of its salience, because the constraints are not classified via a model that is trained on a set of hand-coded provisions of major laws but, rather, on a set of modifiable syntactic rules. For this study, we focus only on the detection of the constraining provisions imposed on the member states and the Commission that take these specific syntactic structures. Future research could be extended to other structures that constraints take, exploiting the malleability of this approach.

6.1. A three-step procedure

Our procedure consists of three steps. First, we rely on dependency parsing, a computational linguistics method, to extract detailed lexical and syntactic information from a text corpus.[16] Dependency parsing produces annotations that describe the structure of a sentence[17] by making predictions about the role of words in the sentence and the grammatical relations between them. Parsers typically rely on grammar annotations produced by a part-of-speech tagger, which detects the grammatical function of each word in the sentence. On top of these annotations, a dependency parser links words together based on their functional dependencies. Both part-of-speech taggers and dependency parsers are supervised models of language, which are trained on large annotated corpora. These models rely on labeled data to learn how to accurately predict the part-of-speech tags and syntactic relationships between words in natural language text. The output of the dependency parser includes tags of the subject of the sentence, the object, the main verb, the auxiliary verb, and other grammatical relations. This information is useful for identifying the structure of a sentence and understanding its meaning. Figure A4 in the Supporting Information illustrates an example of how a parser identifies dependencies from a part-of-speech tagger.

Second, we have employed supervised machine-learning to recognize specific elements in the text of EU laws through so-called named entity recognition (NER) models. Unlike dictionary-based approaches, NER models can adapt and learn from examples provided in a training set, and allow for the consideration of lexical nuances in the identification of named entities within unstructured text. We trained and tested two NER models on a corpus of sentences drawn from EU legal acts,[18] and used them to locate and classify named entities into predetermined categories. The first NER model is trained to recognize four different types of executive institutions: (1) the European Commission, (2) supranational EU agencies, (3) member states, and (4) national competent authorities. In this context, we limit our attention to the Commission and the member states, but the model is trained to identify the words "European Central Bank" as a mention of a supranational agency, and the words "Central liaison office" as indicating a national competent authority. Moreover, the model also learns to manage nuances, variations, and misspellings. The second NER model classifies verbs into four categories: (1) information, (2) delegation, (3) permission, and (4) constraint verbs. Two additional categories of permissive modal and strict modal verbs are also detected. Figure 2 illustrates the combined output of the two NER models.

Third, we have combined the annotations of the syntactic dependency parser with the predictions of the two NER models. To do so, we have assembled an NLP pipeline as illustrated in Figure 3. In the last stage of the pipeline, we have put together the labels produced by the NER models as well as the dependency relations between words extracted by the dependency parser. This allows us to unearth the occurrence of specific patterns associated with constraining relations among EU and national institutions in each sentence.

The objective of this step is to establish whether a specific sentence of a legal text can be conceived as a provision constraining executive authority. For instance, the sentence "the Commission shall inform the Member States" presents a subject belonging to the category "Commission" ("the Commission"), a non-negated strict modal auxiliary verb ("shall"), followed by an information verb ("inform"). If this sentence were to pass through the last stage of our pipeline, the specific combination of syntactical roles, types of actions, and entity labels described above would be matched against "rules of extraction" to check whether it configures a case of constraint.

These extraction rules are derived from the linguistic regularities in the legal syntax that are commonly identified in human-based coding of EU legal texts (e.g., Franchino, 2004; Migliorati, 2021; Thomson & Torenvlied, 2011).[19] The use of phrases such as "the Commission shall inform" is rooted in both the standard practice of legal drafting and the attribution of competencies as specified in the EU treaties (e.g., Schütze, 2018). Each syntactic component of

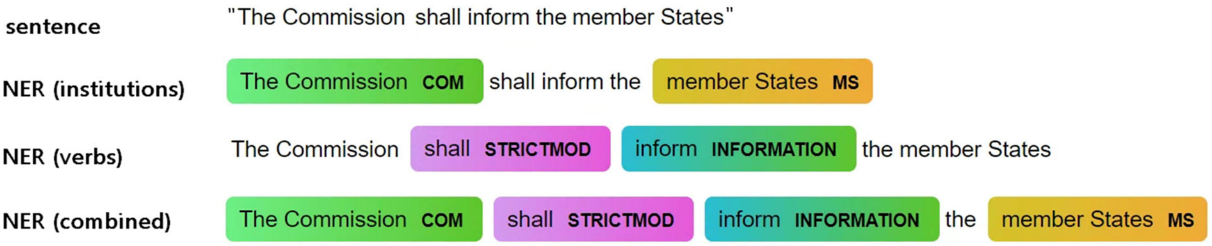


Figure 2 Token boundaries as predicted by named entity recognition (NER) models.

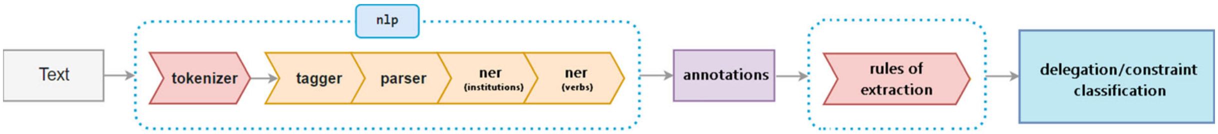


Figure 3 The natural language processing (NLP) pipeline and its components.

these expressions is essential to convey the meaning of the intentions of EU legislators, whether it is to empower or constrain executive actors. For instance, the modal verb "shall" configures a constraint imposed on the Commission as long as it is associated with an information verb such as "inform" or "consult". If it were associated with a generic verb such as "apply", which belongs to our residual category of non-special verbs, or a delegation verb such as "require", the phrase would indicate a conferral of powers to the Commission.

6.2. An initial validity assessment

How effective is this machine-learning approach in detecting the frequency of the use of constraints by EU legislators? As a face validity check, we have compared the temporal patterns in the use of constraints, as predicted by

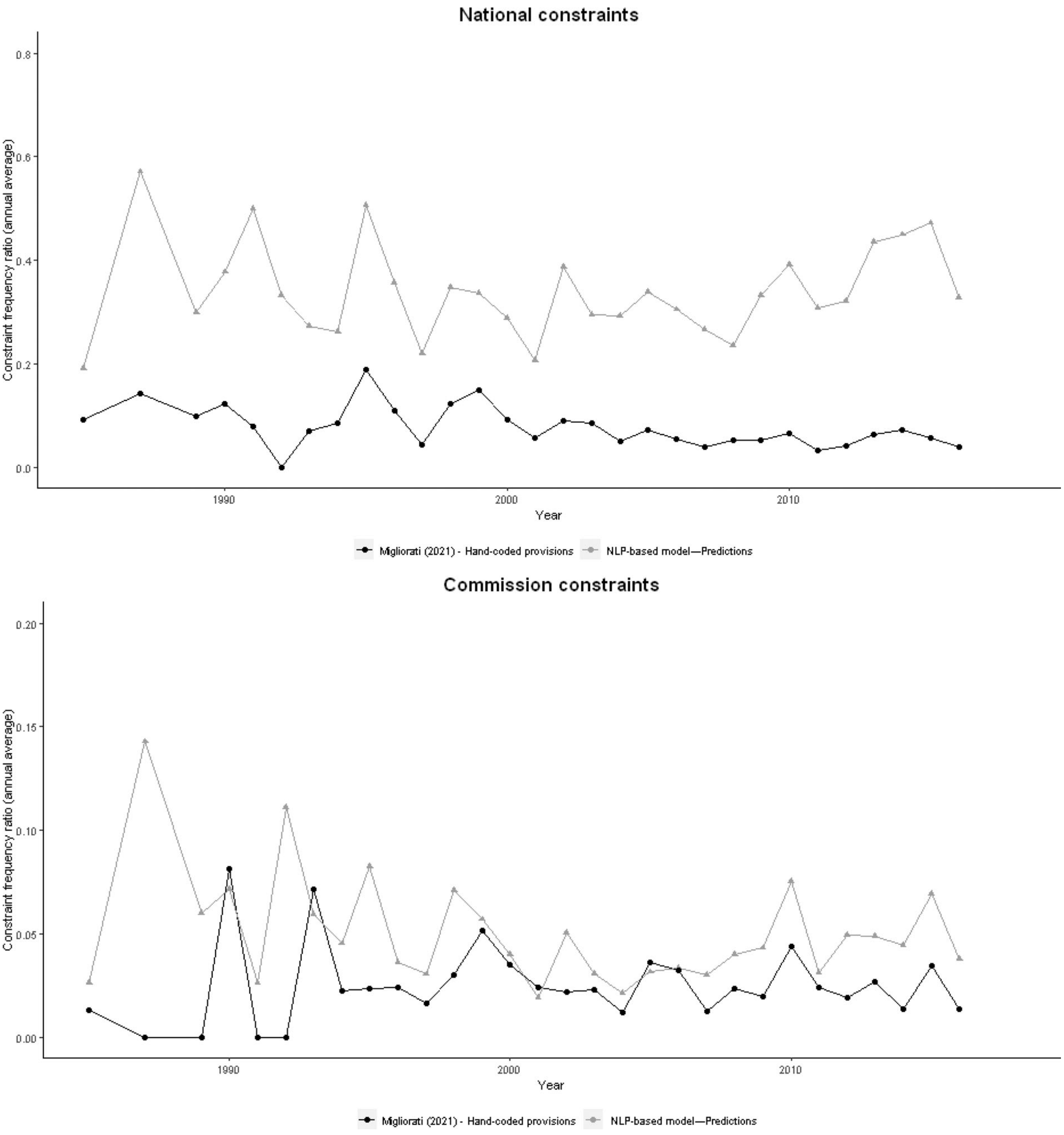
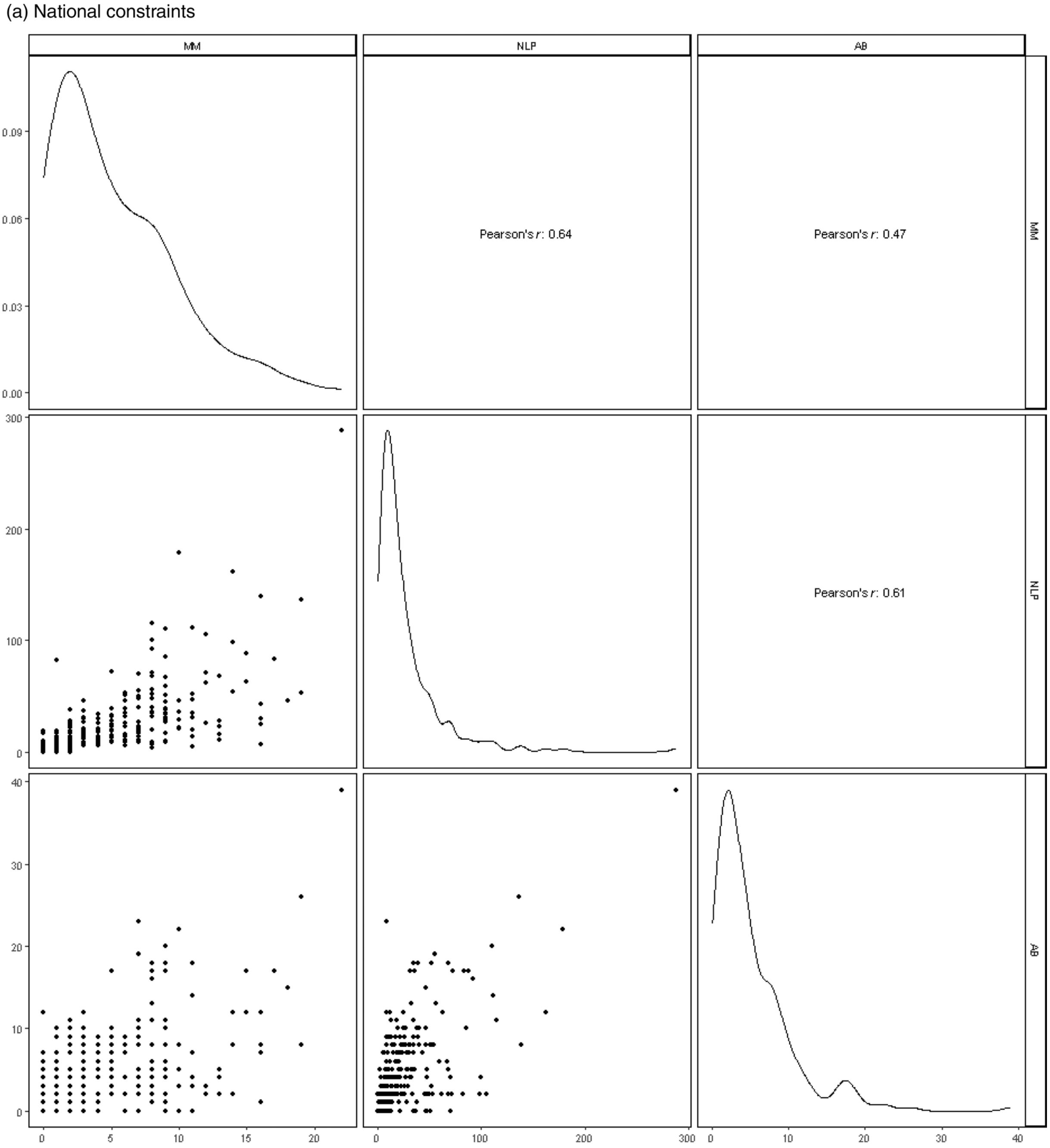


Figure 4 Trends in national and Commission constraint frequency ratios. *Note*: Migliorati's (2021) dataset of 276 major laws, 1985–2016.

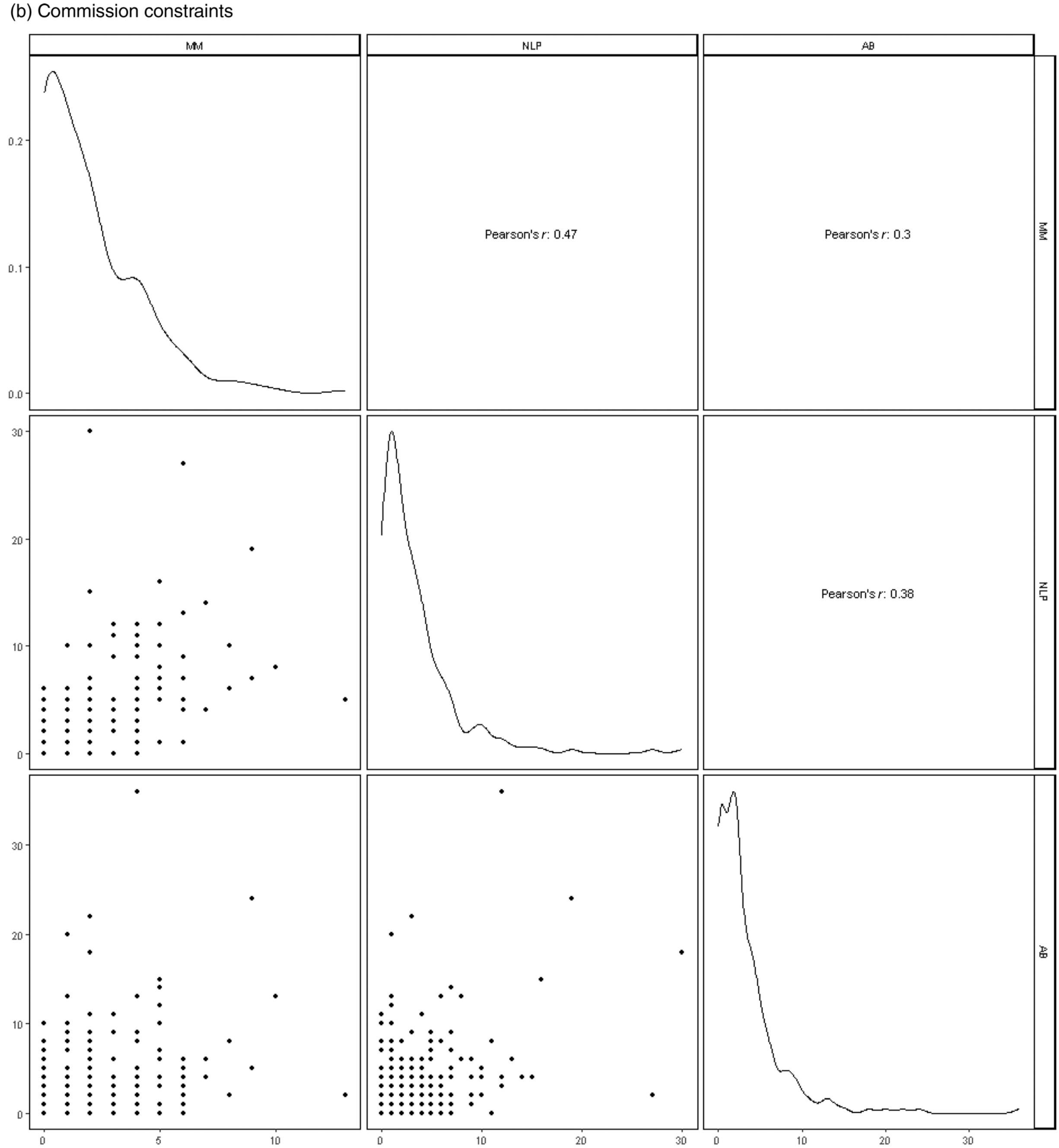
our NLP-based approach, to benchmarks based on human assessment, specifically the hand-coded legal acts in Migliorati (2021).[20] We have computed the constraint frequency ratios for the Commission and the member states by dividing the number of predicted constraining sentences by the number of major provisions in each of these laws. For the hand-coded ratios, we have followed the procedure described above. We have then averaged these ratios over each year. Figure 4 plots the trends for the member states and the Commission, respectively.

The model appears to produce estimates that accurately reflect the fluctuations in the frequency of use of constraints of the hand-coded measure, especially in the case of the Commission. As for the member states, the



(a)

Figure 5 Scatter plots, density plots, and correlation tests of constraint frequencies. *Note*: Constraint frequencies derived from the hand-coded provisions in Migliorati (2021) (MM), the predictions of the NLP-based model (NLP), and Anastasopoulos and Bertelli's (2020) (AB) models.



(b)

Figure 5 Continued

trends are similar until 2005 but then diverge somewhat, with the machine-coded measure showing an increase in the average number of constraints while the hand-coded benchmark remains relatively flat. However, the model also overestimates the presence of constraints, with the gap between machine- and hand-coded measures being wider in the case of member states. We suspect that the primary reason for this discrepancy is the fact that several sentences, especially those containing "member states" as the subject and the strict modal verb "shall", are coded as a constraint by the manual coder only when followed by detailed rules (i.e., the so-called *rulemaking requirements* constraint category). More generic sentences with a similar structure are not coded as such. This distinction between generic and specific language is so policy specific that cannot be captured by general rules of extraction. In future research, the integration of this approach with measures of policy complexity (Hurka et al., 2022; Hurka & Haag, 2020) may improve the model's accuracy.

In a second assessment, we have compared the predictions of this NLP-based method to those made using Anastasopoulos and Bertelli's (2020) validated approach. This comparison allows us to assess validity with reference to a known standard. For this purpose, we have applied Anastasopoulos and Bertelli's (2020) models to the set of Migliorati's (2021) hand-coded provisions. We have then summed up, first for each legal provision and then for each law, the predictions made across all constraint categories to obtain aggregate measures of constraint frequency for both the Commission and the member states.[21] Figure 5 illustrates the scatter plots and the density distributions of these frequencies produced from Migliorati's (2021) hand-coding (MM), our model (NLP), and Anastasopoulos and Bertelli's (2020) models (AB). And they also include their Pearson's *r* correlation coefficients.[22]

As far as national constraints are concerned, the number of Migliorati's (2021) hand-coded constraints correlates more strongly with the predicted number of the NLP-based model (0.64, *p*-value <0.01) than with that of Anastasopoulos and Bertelli's (2020) models (0.47, *p*-value <0.01). Similarly, the number of Commission hand-coded constraints correlates more strongly with the prediction of the former model (0.47, *p*-value <0.01) than with that of the latter (0.30, *p*-value <0.01). Interestingly, Anastasopoulos and Bertelli's (2020) predictions correlate more strongly with those from the NLP-based model, than with Migliorati's (2021) measure. This result could indicate that the accuracy of machine-learning from hand-coded provisions may incur problems with newer laws. Surely, one way to deal with this issue is to train the machine with more recent provisions. But, on the other hand, learning from syntactic structures does appear to hold considerable promise, although there is still room for improvement given the still moderate values of the correlations.

7. Conclusion

We have argued that the different measures scholars have employed to gauge the constraints that are embedded in EU laws reflect alternative conceptualizations of bureaucratic control. A toolbox perspective assumes that legislators have a fixed set of tools are their disposal and, once used, a constraint is applied to the whole act. The greater the variation in the use of constraint *types* the more restricted the implementing agent. A design perspective assumes no limits to the legislators' policy design capabilities and sees constraints as provision-specific rather than applied to the whole legal text. Accordingly, the more frequent the use of *any* constraint the more restricted the implementing agent.

We show that a variation-based measure may underestimate legislative preferences when, especially in short laws, few types of constraint are used frequently. And it may overestimate them when, especially in long laws, several types of constraint are used sparingly. A measure that is based on the frequency of use may instead face fewer validity problems. First, it captures the repeated use of the same type of constraint—surely, a sign of legislators' desire to restrict the executive room for maneuver. And, second, it does not tend to overestimate these preferences when legislators use different types of constraint, but only sparingly. Lastly, the consequence of a measurement error—a false positive or a false negative—tends to be smaller in the constraint frequency ratio because of the larger denominator used for this measure.

We have then assessed the performance of Anastasopoulos and Bertelli's (2020) supervised classification method to estimate constraint variation in EU laws, which is trained on Franchino's (2004, 2007) hand-coded provisions. The relatively poor performance of classifiers that are built in this way seems to have a lot to do with the patterns of similarity and dissimilarity in the wording of the constraints that are associated with the different categories, as well as with the presence of two main implementation paths of EU law. We have, therefore, introduced the skeleton of a different approach whereby learning occurs from the syntactic structures employed by legislators. The initial validity assessment seems promising. Moreover, our model can also syntactically and semantically distinguish executive actors beyond the two entities we investigated for this inquiry. This capability is particularly valuable, in our view, in light of the increasing diversity of executive actors, such as supranational agencies and specific national authorities, that are called upon by EU legislators for policy implementation.

Having said this, there is, however, no doubt that these automated methods are "no substitute for careful thought and close reading and require extensive and problem-specific validation" (Grimmer & Stewart, 2013: 267). They will probably never reach the level of accuracy of human-based coding. "Rather, computer-assisted text analysis augments our reading ability. New text analysis methods help us read differently, not avoid reading at all. This amplification of human effort improves the analyst's ability to […] measure key quantities of interest, estimate causal effects, and make predictions" (Grimmer et al., 2022: 57).

Moreover, the scalability of these techniques offers the prospect of producing a far more representative picture of the patterns of delegation and constraint in the EU. Their flexibility and affordable replicability will speed up the evaluation of our research findings. Coupled with careful human-based validation, they can contribute greatly to our understanding of policy design in the European Union.

Acknowledgments

We would like to thank Moritz Osnabrügge, A. Maurits van der Veen, and the participants at the panel on *Innovative Applications of Text Analysis to the Study of European Union Politics* of the 2022 Annual Conference of the European Political Science Association.

Data availability statement

The data that support the findings of this study will be made openly available on dataverse.org if the study is accepted for publication.

Endnotes

[1] Some scholars pay no attention to constraints either implicitly (Kaeding, 2008; Thomson, 2007; Zhelyazkova, 2013) or explicitly. Thomson and Torenvlied (2011), for instance, focus only on delegating provisions and explicitly disregard constraints.

[2] Recent applications to US law are Clouser McCann (2016), Groll et al. (2021), and Lavertu (2013, 2015).

[3] In comparative politics, Huber and Shipan (2002: 73) utilize the length of the legislative text as a proxy for control since "longer statutes on the same topic are more likely to tell agencies what to do, and shorter ones give them more leeway". On a similar note, Citi and Jensen (2022) measure the levels of precision in legislative texts which limit the executive discretion of supranational authorities, while Gastinger and Schmidtke (2022) develop a dictionary-based measure of imprecision as a proxy of discretion of international organizations.

[4] Regulation 11 concerning the abolition of discrimination in transport rates and conditions, in implementation of article 79 (3) of the Treaty establishing the European Economic Community.

[5] The 12 categories are: time limits, spending limits, reporting requirements, consultation requirements, public hearings, rule-making requirements, appeals procedures, exemptions, legislative action required, legislative action possible, executive action required, and executive action possible.

[6] Scholars of the U.S. Congress have leveled a similar criticism (Bolton & Thrower, 2019: 1271; MacDonald, 2007: 687).

[7] In this context, we set aside supranational agencies despite their increasing relevance in implementation. National competent authorities are instead considered executive actors associated with national implementation.

[8] Figure A1 in the Supporting Information illustrates separate plots for the two datasets.

[9] The coefficients for the major laws collected by Franchino (2007) are 0.55 for the national constraint ratios and 0.83 for the Commission constraint ratios, while they are only respectively 0.38 and 0.57 for the laws in Migliorati (2021), *p*-values <0.01.

[10] In the case of national implementation, the correlation between P_i (the number of major provisions in law *i*) and T_i (the number of constraint types used in law *i*) is 0.41, and between P_i and C_i (the total number of any constraint used in law *i*) is 0.44. In the case of Commission implementation, the figures are respectively 0.31 and 0.16 (*p*-value <0.01 in all cases).

[11] In practice, it takes only eight values, from 0 to 0.58 (i.e., 7 out of the 12 possible constraints), as is evident from Figure 1.

[12] A final aside on weighting: Constraints do not restrict executive actors to the same extent. A related issue is, therefore, whether we should weigh them differently and whether incorrect weighting creates greater validity problems for one of the two measures. As we discuss at length in the Supporting Information, the use of weights may be appealing on the surface but it is far from straightforward in practice. Nevertheless, incorrect weighting does not appear to have graver consequences for the validity of one constraint ratio or the other.

[13] No classifiers are trained for exemptions and requirements of legislative action which are the second and third least-popular categories, for public hearings and possible legislative involvement in the case of national implementation, and

for spending limits in the case of Commission implementation because they too are rare. Appeals procedures have not been used in the case of Commission implementation.

[14] We employ the convention that the F1-score (a measure combining precision and recall) should exceed 0.5 for the classifier performance to be considered acceptable. Precision is the share of provisions that the classifier has correctly identified as containing a given constraint, out of all the provisions identified as containing that constraint. Precision improves if the number of false positives decreases. Recall is the share of correctly identified provisions out of the sum of such correctly identified provisions and the provisions that the classifiers failed to identify as containing that constraint. Recall improves if the number of false negatives decreases. Table A4 in the Supporting Information summarizes the performance metrics of table 2 in Anastasopoulos and Bertelli (2020: 294) and adds two columns listing the number of laws in Franchino's (2004, 2007) dataset containing a given type of constraint, and the frequency of use.

[15] In so doing, we take the lead from the works of Vannoni et al. (2021) and Shaffer (2022) on US legislation.

[16] The dependencies are produced using the Python library *spaCy* (Honnibal & Johnson, 2015), a state-of-the-art NLP suite.

[17] We chose the sentence, rather than the provision, as our level of analysis for several reasons. First, the automated partition of the text into sentences is more effective compared to the partition into legal articles or provisions. Second, the resulting texts are shorter and easier to code. Third, although the wider context in which a sentence is embedded could in principle contain useful information, the sentence is the most appropriate unit of analysis for the NLP components that we employ to analyze syntactic connections between words. See the Supporting Information for a detailed account of the pre-processing steps.

[18] The models "learn" the relations between the text tokens and the entity categories from two randomly selected samples of sentences that are extracted from a pre-processed corpus and have been manually annotated using the Python-implemented platform "Prodigy" (https://prodi.gy). The text corpus consists of the legal acts in the CEPS Eur-Lex dataset (Borrett & Laurer, 2019), a comprehensive collection of more than 130 thousand laws—almost the entire corpus of digitally available acts adopted between 1952 and 2019 by the EU, the European Economic Community and the European Coal and Steel Community. On the basis of the two samples, we trained two final NER models that reported satisfactory scores in terms of precision, recall, and F1 metrics. Tables A4 and A5 in the Supporting Information report the performance metrics of the NER models.

[19] The rules of extraction are listed in Table A7 in the Supporting Information.

[20] Validation through comparison with human-coded measures is standard practice for both supervised and unsupervised methods of automated analysis of text. We expect learning from syntactic structures to perform better with the more recent laws in Migliorati's (2021) dataset because of the greater variety of actors involved in implementation.

[21] We aggregate constraints imposed on the member states with those imposed on national competent authorities for both our and Migliorati's (2021) hand-coded measure. The model developed by Anastasopoulos and Bertelli (2020) is trained on Franchino's (2007) hand-coding which subsumes member states and national competent authorities. Their model, therefore, classifies constraints imposed on both these actors as constraints on member states.

[22] See Table A8 in the Supporting Information for more details.

References

Anastasopoulos, L. J., & Bertelli, A. M. (2020). Understanding delegation through machine learning: A method and application to the European Union. *American Political Science Review*, 114(1), 291–301. https://doi.org/10.1017/S0003055419000522

Bolton, A., & Thrower, S. (2019). The constraining power of the purse: Executive discretion and legislative appropriations. *The Journal of Politics*, 81(4), 1266–1281. https://doi.org/10.1086/704330

Borrett, C., & Laurer, M. (2019). *The CEPS EurLex dataset: 142.036 EU laws from 1952–2019 with full text and 22 variables* (Centre for European Policy Studies, Ed.; H. 2020 European Union, Trans.; V2 ed.). Harvard Dataverse. https://doi.org/10.7910/DVN/0EGYWY

Citi, M., & Jensen, M. D. (2022). The effects of supranational delegation on policy development. *JCMS: Journal of Common Market Studies*, 60(2), 337–354. https://doi.org/10.1111/jcms.13228

Clouser McCann, P. J. (2016). *The federal design dilemma: Congress and intergovernmental delegation*. Cambridge University Press. https://doi.org/10.1017/CBO9781316275085

Epstein, D., & O'Halloran, S. (1999). *Delegating powers: A transaction cost politics approach to policy making under separate powers*. Cambridge University Press.

Ershova, A. (2019). The watchdog or the mandarin? Assessing the impact of the directorates general on the EU legislative process. *Journal of European Public Policy*, 26(3), 407–427. https://doi.org/10.1080/13501763.2018.1447009

Franchino, F. (2004). Delegating powers in the European Community. *British Journal of Political Science*, 34(2), 449–476.

Franchino, F. (2007). *The powers of the union: Delegation in the EU*. Cambridge University Press.

Franchino, F., & Mariotto, C. (2020). Politicisation and economic governance design. *Journal of European Public Policy*, 27(3), 460–480.

Gastinger, M., & Adriaensen, J. (2019). Of principal(s') interest? A disaggregated, multiple principals' approach to commission discretion. *Journal of Common Market Studies*, 57(2), 353–370. https://doi.org/10.1111/jcms.12801

Gastinger, M., & Dür, A. (2021). Joint bodies in the European Union's international agreements: Delegating powers to the European Commission in EU external relations. *European Union Politics*, 22(4), 611–630. https://doi.org/10.1177/14651165211027397

Gastinger, M., & Heldt, E. C. (2022). Measuring actual discretion of the European Commission: Using the discretion index to guide empirical research. *European Union Politics*, 23(3), 541–558. https://doi.org/10.1177/14651165221098487

Gastinger, M., & Schmidtke, H. (2022). Measuring precision precisely: A dictionary-based measure of imprecision. *The Review of International Organizations*. https://doi.org/10.1007/s11558-022-09476-y

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21, 267–297. https://doi.org/10.1093/pan/mps028

Groll, T., O'Halloran, S., & McAllister, G. (2021). Delegation and the regulation of U.S. financial markets. *European Journal of Political Economy*, 70, 102058. https://doi.org/10.1016/j.ejpoleco.2021.102058

Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, Lisbon, Portugal, September 17–21, 2015, 1373–1378. https://doi.org/10.18653/v1/d15-1162

Huber, J. D., & Shipan, C. R. (2002). *Deliberate discretion?: The institutional foundations of bureaucratic autonomy*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804915

Hurka, S., & Haag, M. (2020). Policy complexity and legislative duration in the European Union. *European Union Politics*, 21(1), 87–108. https://doi.org/10.1177/1465116519859431

Hurka, S., Haag, M., & Kaplaner, C. (2022). Policy complexity in the European Union, 1993-today: Introducing the EUPLEX dataset. *Journal of European Public Policy*, 29(9), 1512–1527. https://doi.org/10.1080/13501763.2021.1938174

Kaeding, M. (2008). Lost in translation or full steam ahead: The transposition of EU transport directives across member states. *European Union Politics*, 9(1), 115–143. https://doi.org/10.1177/1465116507085959

Lavertu, S. (2013). Issue-specific political uncertainty and policy insulation in US Federal Agencies. *The Journal of Law, Economics, and Organization*, 29(1), 145–177. https://doi.org/10.1093/jleo/ewr029

Lavertu, S. (2015). For fear of popular politics? Public attention and the delegation of authority to the United States executive branch. *Regulation & Governance*, 9(2), 160–177. https://doi.org/10.1111/rego.12061

MacDonald, J. A. (2007). Agency design and postlegislative influence over the bureaucracy. *Political Research Quarterly*, 60(4), 683–695. https://doi.org/10.1177/1065912907304151

Migliorati, M. (2021). Where does implementation lie? Assessing the determinants of delegation and discretion in post-Maastricht European Union. *Journal of Public Policy*, 41(3), 489–514. https://doi.org/10.1017/S0143814X20000100

Schütze, R. (2018). *European constitutional law* (2nd ed.). Cambridge University Press.

Shaffer, R. (2022). Power in text: Implementing networks and institutional complexity in American law. *The Journal of Politics*, 84(1), 86–100. https://doi.org/10.1086/714933

Steunenberg, B., & Toshkov, D. (2009). Comparing transposition in the 27 member states of the EU: The impact of discretion and legal fit. *Journal of European Public Policy*, 16(7), 951–970. https://doi.org/10.1080/13501760903226625

Thomson, R. (2007). Time to comply: National responses to six EU labour market directives revisited. *West European Politics*, 30(5), 987–1008. https://doi.org/10.1080/01402380701617407

Thomson, R., & Torenvlied, R. (2011). Information, commitment and consensus: A comparison of three perspectives on delegation in the European Union. *British Journal of Political Science*, 41(1), 139–159. https://doi.org/10.1017/S0007123410000268

Vannoni, M., Ash, E., & Morelli, M. (2021). Measuring discretion and delegation in legislative texts: Methods and application to US states. *Political Analysis*, 29(1), 43–57. https://doi.org/10.1017/pan.2020.9

Zbiral, R., Princen, S., & Smekal, H. (2022). Differentiation through flexibility in implementation: Strategic and substantive uses of discretion in EU directives. *European Union Politics*, 24, 102–120. https://doi.org/10.1177/14651165221126072

Zhelyazkova, A. (2013). Complying with EU directives' requirements: The link between EU decision-making and the correct transposition of EU provisions. *Journal of European Public Policy*, 20(5), 702–721. https://doi.org/10.1080/13501763.2012.736728

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1. Supporting Information.

Data S1. Replication_1.

Data S2. Replication_2.

Data S3. Replication_3.