Regulation Complexity Measurements: Methods and Patterns across Time and Industry

Tran Hoang Phuong Linh Supervisor: Prof. John Galbraith

October 3, 2025

Abstract

How do measures, especially complexity measures, based directly on regulation texts relate to economic figures? Can they explain things that conventional numerical variables cannot? To answer this question, we first need to examine and present all possible ways of extracting information directly from legal texts, including those that already exist and newly proposed methods. Regarding the complexity aspect, we found that the new and traditional methods can differ drastically over time and across industries. The new method measurements can also potentially explain some variation in economic figures not reflected by the traditional counterparts. Finally, we also present methods for extracting various textual features that are not complexity for a broader view of legal text patterns.

Keywords: complexity measurements, regulation texts, industry

1 Introduction

Laws and regulations have always played an indispensable role in maintaining social order and ensuring the stability of growth. For example, stronger property rights laws encourage the creation of products or innovations by guaranteeing the security of ownership and minimizing the uncertainty of expected profits. In this way, economic activity and investment grow strongly, pushing the economy in a positive direction. However, we also know that too many laws - in their complexity - can hinder economic activities by creating barriers to entry, vulnerability to lawsuits due to lack of knowledge, and excessive regulatory costs. For example, Gallini (2017) mentions the possibility of increasing patent trolling and patent thickets due to stronger intellectual property laws, which could lead to increased litigation costs and deter new technologies in "complex" technology areas.

In the case of U.S. regulation, one of the most important legal documents is the Code of Federal Regulation (CFR), which is also our main object of study due to its size and importance. Over the past 56 years, this system has grown at an incredible rate in terms of both size and complexity. Specifically, according to Davis (2015), the CFR has up to 180,000 pages, containing "as many words as 133 copies of the King James Bible!" Because these documents are extremely important to many aspects of the economy, such as business operations (intellectual property, patents, and administrative documents) and workers' rights and responsibilities (tax code and employment contracts), a proper understanding of the CFR is required regardless of one's background. This comes at an enormous cost in time, effort and money. For example, the National Taxpayer Advocate (2014) mentioned that approximately 94% of individual taxpayers used a preparer or tax software for 2013 because they were overwhelmed. Another example comes from the article by Hadfield (2000), where the author shows how expensive lawyers are and why this is the case, thus reflecting the cost of doing business that increases as the complexity of the law increases. On the other hand, in some cases the expansion of laws is necessary. The "loss" from complexity, such as GDP, usually does not take into account the benefits to society, such as positive environmental externalities, job stability and security.

Studies in this area mostly rely on the "complexity" of the regulation and examine its impact on

the economy. This measurement has been proxied by after-effect variables with intuitively obvious impacts on the economy. For example, a clear after-effect of additional regulations on business entry could be reflected in the time or number of procedures required to obtain legal documents for new businesses. There are many others, such as the cost of starting a new business, the cost of operating a business, the degree of technology restriction, and so on. Most of these measures can be associated with an extremely diverse set of model structures (perhaps even uniquely associated with some measures) and require a deep understanding of different domains. On the other hand, measures of complexity can be derived directly from the regulatory texts themselves. This is usually in the form of the number of words/pages - basically the volume of the text. The latest movement in this type of measurement is to count the number of rules instead. Overall, this type of method tries to estimate complexity based on the text itself, and its effect on the economy can be ambiguous, unlike its after-effect counterparts.

In this research, we will focus on measures derived directly from legislative texts. Specifically, we want to examine these complexity measures in more detail and introduce a new method that has not yet been used in the economic literature. Is the traditional way of measuring regulatory complexity sufficient? How do the old and new ways of measuring complexity differ over time, across industries, and across external events such as changes in the composition of Congress? On the other hand, complexity is only one specific aspect of laws and regulations. Are there other characteristics we can derive? And again, how do they behave over time and across industries? Finally, how well does the new way of measuring complexity explain the variation in the performance of some industries? Is there a difference from the traditional way?

Due to time constraints, we will specifically examine Title 40 - Environmental Protection - of the CFR, which is only one of 50 available titles (although this title is the largest, spanning the length of several titles combined).

In the next section, we will review the literature on three aspects: the impact of legislation, complexity measures, and current text mining methods. In section 3, we will introduce the three datasets containing the raw text and industry information. This is followed by the methodology for complexity specifically, for other potential features, and for simple regressions using the complexity

measures on economic outcomes. Section 5 discusses the results and section 6 concludes with some limitations and future research directions.

2 Literature Review

This paper potentially involves three different areas of study: (i) the impact of legislation on the economy, specifically for the industry level, (ii) the measurement of regulatory complexity, and (iii) text mining methodology.

2.1 The effect of legislation on the economy

There have been many studies examining the effect of regulations on the economy at the aggregate level. Persistent within the literature is a long-standing debate on how regulations affect the economy's functionality. Mixed results are reported theoretically and empirically. Supporters of increasing regulations suggest its role in decreasing legal uncertainty, including enhancing property rights and adjusting incomplete laws (thus, reducing transaction costs due to market flaws). By strengthening economic stability, additional regulations stimulate investment decisions and economic activities. For instance, Ash et al. (2022) shows that new regulations are welcome - increase growth in this case - under certain situations such as during times of economic uncertainty, when the initial stock of laws is low, and when the newly added law is in the form of contingent clauses.

At the other end of the spectrum, unnecessarily complicated regulations can discourage economic agents from contributing to the economy and distort their behaviour in negative ways. Much of the literature focuses on the effect of fixed costs - either time or money - of regulations on markets. Some articles proxy the complexity of market entry by the time it takes to obtain legal documents and the number of procedures entrepreneurs have to go through before opening their businesses. (Fonseca et al., 2001; Ciccone and Papaioannou, 2007; Braunerhjelm and Eklund, 2014). Other empirical studies also support the negative relationship between a country's economic growth rate and its legislative output. (Kirchner, 2012). For this reason, most theories are strongly tailored to firms and innovation. Other theories rely on the interaction between the

two rival policymakers (Kawai et al., 2018), thus creating complex regulation that is suboptimal to cause problems for the other party. Similarly, Foarta and Morelli (2020) (and in a similar sense, Gratton et al. (2021)) build a model to find equilibrium in a game where decision makers accept or reject a regulation from a proposal based on the state of the economy with asymmetric information. They find that a less competent proposer tends to increase complexity in the long run, which reduces the efficiency of the policy enacted later.

However, examining this relationship - between the legislative output and the economic GDP or growth - at the country level can be misleading. Different industries/sectors can vary significantly from one another, and the need for regulations and rules for each market to function effectively can be vastly different: a set of regulations from one market is usually not transferable to another. Some industries require more attention to the competition level and collusion while others require more regulation on workers' conditions and job security. This can be particularly helpful when, in some economic situations, different sectors are hit differently than others, and the need for policy change should not be made aggregately. Thus knowledge from heterogeneous sectors can come in handy, and this is what I want to focus on. Coffey et al. (2020) also discusses the effect of regulatory complexity on industry growth where the author uses linear specification derived from closed-form solutions of a multi-sector Schumpeterian model of endogenous growth. Ideally, when a firm decides to invest, this leads to productivity improvement and hence growth. However, in the presence of regulation, investment behaviours are distorted, thus affecting growth. However, similar to previous literature, the definition of regulatory complexity does not fully represent the concept as a whole. As a result, while most industries' growth is negatively correlated with this complexity, there are some industries that have the exact opposite sign. This is also what our paper aims to improve by introducing a finer and more complicated indicator of regulatory complexity. This could potentially explain or incorporate the mixed results in the literature.

2.2 Measurement of regulation complexity

Most of the proxies for regulatory complexity are intuitively uncontroversial or only partially reflect the meaning of regulatory complexity. We consider most of them to be aftereffect variables.

As noted above, an increase in measures such as the time required to obtain legal documents and the cost of production (either in labour or capital) will intuitively reduce output and increase barriers to entry. But this is not the only thing that laws have to offer. Another example comes from Di Vita (2018), where regulatory complexity is measured by the unweighted sum of annual flows from four sources of regulation. Since there is little or no formal or informal agreement between the different bodies that issue laws, overlap and confusion are obvious, so negative relationships are expected. Since we treated the data as if they came from a single source, this is not taken into account.

On the other hand, there has been an emergence of using the raw legislative text directly and using some method to approximate the complexity. The most popular idea is that the more regulations are written, the more complex the law is. This is reflected by variables such as the total number of words (Gratton et al., 2021), the number of pages in total/per acts (Kirchner, 2012), the number of lines (Kawai et al., 2018), the number of bills/laws newly introduced (Gratton et al., 2021) etc. Recently, the measurement has been refined to be more accurate and tailored to count the number of rules rather than the volume of text. For example, the RegData series (2.2 specifically - McLaughlin and Sherouse (2019)) contains an indicator for regulation complexity which is expressed as total counts on specific words and phrases such as "shall", "must" and "other terms" (including "may not", "prohibited" and "required"). The authors argue that raw page counts may measure bureaucratic activity rather than regulatory growth. In Ash et al. (2022), the author also counts the number of times certain kinds of words occur in the "right" context: strict modals (e.g. "shall", "must"), permissive modals (e.g. "may"), delegation verbs (e.g. "require"), constraint verbs (e.g. "prohibit"), and permission verbs ("allow"). However sophisticated these extractions are, they stop at measuring the volume of rules contained in the legislative text. Does a greater number of rules imply greater complexity? What if the extra text is added to clarify a situation that is considered too vague and is causing losses due to unresolved litigation? If so, the number of rules remains the same, but the complexity does not. We want to propose measures of regulatory complexity other than the number of rules.

These potential features can be extracted in the context of legal writing patterns or in the

context of general human reading comprehension. In the former case, some papers below mention useful patterns that could be used:

- In Gratton et al. (2021), the author's comments on the incompetence of the proposer can be observed by counting the number of syntax and spelling errors, incomplete or inconsistent sentences, and corrections by subsequent legislative amendments. Furthermore, Normattiva (2016) shows that there are 4 indicators that measure the quality of individual laws and the complexity that the law injects into the legal system: the average length of sentences in characters, the number of gerunds per 1000 words in the law, the presence of a preamble in the law, and the number of references to other laws in the main body of the law per 1000 words in the law. The first two indicate clarity; the last two indicate legal complexity and accessibility of the law to laymen.

- In Nicoletti et al. (2003), the patterns are measured specifically for what extent competition and firm choices are restricted in industries/areas where no prior reasons for government inferences or regulatory goals could plausibly be achieved by less coercive means. Although constructed for the OECD, it can be applied to the CFR. There are 4 sets of regulatory indicators used in this paper: economy-wide regulation, industry-level regulation, regulatory reform, and privatization. Each of the figures is determined manually according to specific charts and different aspects of "industry restrictions".

Finally, according to Katz and Bommarito (2014), looking at complexity in the view of human comprehension, the measurement is based on knowledge acquisition theory: how a human goes through a document and learns from it. The goal of this paper is to rank and identify which title of the CFR is the most complicated, and to discover the connections between different parts of these regulations. This is not an economically oriented result, as the research comes from computer science and legal disciplines. However, the methodology is extremely helpful and close to what we want to achieve. So I will use this theory to build the new complexity measure. More details will be presented in the Methodology section.

Outside of the legal context, the complexity of understanding a text can be estimated. A great example is the application Grammarly, where the machine evaluates a piece of text and

assigns a degree of clarity to it. According to their website (gra, 2019), their trained model analyzes "millions of sentences from research corpora. (A corpus is a large collection of text that has been organized and tagged for research and development purposes.)". This might suggest that the model is trained on labelled corpora, which means that if we want to match their method, we need to obtain legal text data with each of the sections/sentences/statutes/etc. labelled with a level of complexity that is not the CFR. We then train different machine learning models on this dataset and apply them to the CFR. This could work if we could find some manual pre-determined complexity of the legal text, train on it, and apply it here for the CFR. A good candidate is the data from the OECD laws in Nicoletti et al. (2003). Another potential dataset is from "Doing Business 2017: Equal Opportunities for All". However, due to time constraints, we will leave it open for future work if possible, and it is not included in this paper.

2.3 Text mining literature

Complexity is not the only thing legislation data contain. Focusing only on measuring complexity may not paint the full picture of how regulation affects the economy. For example, the recent emergence of online marketing and services raises many concerns about privacy and personal data collection. These are not necessarily bad things, and they are not concerns in industries like manufacturing. So we want to look for other features, and this can be treated either as unsupervised text learning (in this case specifically topic modelling) or as a prediction from trained models.

First of all, we need to look at a typical text data usage structure in economics. This is taken from Gentzkow et al. (2019):

1. Raw text D to numerical array C. Typically, C is extremely high-dimension, thus the need for text preprocessing.

2. use C to predict value $\hat{V}$ of the outcome V

    - Columns of C used as features in linear regression for GDP prediction.
    - Use C to get the topic proportions for each document - topic modelling (Section methodology)

3. Use $\hat{V}$ for subsequent descriptive or causal analysis

    - Topics can be used as features

Other literature in different disciplines has confirmed the crucial influence of the first step - text pre-processing - on the final results of unsupervised text learning methods. However, in economic literature, there is little to no guidance on why each technique is used, only "to the best of your judgment". Furthermore, most of the reasons for using these methods are not tailored to the characteristics of the document.

- Pejić Bach et al. (2019) and Gentzkow et al. (2019) list out all available techniques but there is little to no guidance
- Benchimol et al. (2022) does not account for contraction, abbreviation, spelling error/extra white space.
- Aruoba and Drechsel (2022) constructing all possible words/phrases of length 1 to 3, and manually picking the most representative ones (18,000; 450,000 and 600,000 terms for length 1 to 3)

We will be using most of the algorithm from Hickman et al. (2022) (Organizational research). However, we will also be reviewing the pros and cons of each text pre-processing method in other disciplines to create a more general algorithm for dealing with text data, including Grishman (2015) (Artificial intelligence), Yang et al. (2022) (Causality extraction), Pejić Bach et al. (2019) (Stock/Finance), and their relevant articles. This will be in more detail in Section 4.

In the second step, since we are dealing with unsupervised learning, one of the most popular methods is topic modelling using Latent Dirichlet Allocation (LDA). We will use this to discover whether the diversity of topics discussed within Title 40 changes over time and changes across industries.

On the other hand, using pre-trained models can also help extract information from the text. This requires little to no expertise as you only need to feed the machine the raw text itself and it will give you the figures that the model was trained for. For example, there is a pre-trained

model that gives out a sentimental score - a number that signifies how positive/negative a piece of text sounds. With this, we can pass on a piece of document and receive this number. An excellent analogy for this is a simple linear model with pre-estimated coefficients. Passing the raw text to the model is equivalent to passing the value for the independent variables and getting the predicted output. More details are discussed in the 4 section.

3 Data

3.1 Code of Federal Regulation Raw text Data

The Code of Federal Regulations raw texts from 1997 to 2021 are available as XML files. Any raw text prior to 1996 is available only as scanned images, which requires the use of optical character recognition (OCR). According to Ash et al. (2022), this method comes with a significant amount of misspellings. This type of data will also deviate greatly from the format of XML-extracted data. The plan, for now, is to ignore this part of the CFR and focus more on the more recent regulations, which are from 1997 to 2021.

Some general information about the CFR:

- There are a total of 50 titles each year. Each title contains one or more individual volumes that are updated once each calendar year. Title 40 - our object of interest - is called Environmental Protection.

- Within each Title, the hierarchy elements are listed as follows (in XML-files structure): Title, Chapter, Part and Section. Sections are the highest level where substantial bodies of texts can be found (higher elements only contain headings and table of contents). Since some of the sections only have 1-2 small paragraphs, we extract the data at part-level instead, which is convenient for matching with the RegData dataset (details in the next section) where the authors give additional reasons why they choose part-level: "First, part-level division is present in every title of the CFR, and the parts in a title collectively contain all non-appendix regulatory text. Second, parts tend to focus on a set of related issues that are likely to have

similar relevance to industries throughout. Third, sources of regulatory authority are cited at the part level."

From the XML files, I manage to extract the raw texts at part-level (each observation is a piece of document d(i,y) - all text from part i of the title 40 in year y - for convenience purpose, title subscript is dropped since we only consider 1 title), the part number, the part heading, the number of sections within the part, the title number (which is only 40 for this research), and the year. A snippet example can be found in Appendix A, figures 12 and 13. Notice that some chapters/parts/sections are branded with "[Reserved]", which is used to indicate that a portion of the CFR was intentionally left empty and not accidentally dropped due to a printing or computer error. For this reason, these elements are as good as empty, hence omitted from the body of texts (Although we utilized this information later). Furthermore, some of the appendices are added separately away from the main texts, which is also omitted. The reasons are that these are usually quite standard procedures and are determined by available equipment. Including them could distort the complexity measurement since they are not really adding significant layers of complexity.

3.2 Industry Information

The RegData dataset is obtained from QuantGov. Each observation (document) is identified by the year, title number, and part number, which allows us to link to the main dataset from the previous section. The main information from this dataset is the probability that a document is related to a certain industry, which ranges from a 2-digit to a 6-digit NAICS industry code. We decided to use both the 2-digit and 3-digit codes for the research. Specifically, both are used for regression purposes to maximize the number of observations available; the 2-digit codes are used for graphical purposes, and some of the 3-digit codes are also used when the categories in the 2-digit codes are too broad to make much sense of the complexity measure. For example, industry codes 31-33 are all labelled "manufacturing" and include both "food manufacturing" and "computer and electronic product manufacturing", which require drastically different types of regulation. On the other hand, 4-digit codes or higher are not used for several reasons: the

probability relevance can be very noisy and less accurate since the probability is obtained by using a trained classification model (more categories mean less reliable estimates); they are also extremely heavy to run correctly.

Data on potential economic impacts are obtained from the Statistics of U.S. Businesses (SUSB) Annual Datasets by Establishment Industry. These data are linked to the data in the previous section using the NAICS industry code. They include information on firm size, number of firms, number of establishments, number of employees, and annual payroll. They are currently available from 1998 to 2020. More detailed descriptions of each variable and other issues with the data are provided in Appendix A.

4 Methodology

In this section, we will first present the main complexity measures (both state-of-the-art and new methods). We will then look at other potential measures. Depending on the goal and the nature of each measure, different methods are needed. Note that each of these measures is estimated for each piece of document d(i,y) (all text from part i of Title 40 in year y).

4.1 Complexity measurement

4.1.1 Volume-based measurement

The nature of these methods is simple: larger volumes of text require more time and effort to review and understand. Although, in general, two documents of the same length can have drastically different levels of complexity, there may be reasons to justify these measurements to some extent. Regulatory texts are official documents, which means that the style and language used should be fairly consistent (in the sense that it is based on an official language). In addition, a larger volume of text simply requires a longer attention span, and this could be especially true for regulatory texts, as there will be numerous details due to the rigorous nature of the laws. More things to keep in mind can increase the chance of making mistakes or forgetting details.

How to measure this is perhaps not so straightforward. Here we will use three different

proxies that are quite similar:

1. The simplest way is to count the number of "words" (**num_words**). Raw texts are split by whitespace (including the conventional spacebar, newline notation and other kinds of whitespaces), and individual items after the split are considered words.

2. The number of tokens (**num_tokens**) and sentences (**num_sents**) are also estimated. These are derived using the SpaCy package for natural language pre-processing. Intuitively, tokens are derived more intricately than simple splits in the previous measurement: it counts components of the text differently even when they are not separated by whitespaces. For instance, special characters such as commas or exclamation marks are counted separately as tokens instead of "sticking" with the word right in front of it (no whitespace). Hence, the number of tokens is larger than simple word counts.

These are widely utilized in the economic literature since they are easy, convenient and fast to obtain. Notice that these account for everything in the texts. As mentioned before, McLaughlin and Sherouse (2019) argue that the growth in text size could be solely due to the growth in bureaucratic activity rather than the actual restrictions growth in the regulations. This leads to the second category of complexity estimates.

4.1.2 Rule-based methods

The number of rules may remain the same even though more texts are added (this could be due to the need for clarification, changes in standard presentations, etc.). The main idea is that more rules mean more things to consider (very similar to the previous section). However, the extra information that is not considered as rules might not add to the complexity if it is straightforward (at least to the targeted people) or just added for the sake of completeness. Methods in this section are relatively new and the goal is to extract the number of rules only and try to be as precise as possible. The method we tried to replicate here comes from McLaughlin and Sherouse (2019), which we refer to as RegData. The details are as follows:

1. The basis is to detect rules: we will count on specific words and phrases that intuitively

"generate" rules. Here the authors considered "shall", "must" and "other terms" (including "may not", "prohibited" and "required"). These are counted as **numShall**, **numMust**, **numMayN**, **numProh** and **numReq**

2. They then proposed 2 different ways to finalize the complexity measurements:

   - The early version only sums up the total number of occurrences of all the above words - **resWordv1**
   - The later version counts an appearance multiple times if the paragraph they lie in ends with the colons (":") and follow by a list. This is called **resWordv2**. For instance, consider the following piece of text:

     "The business **must** satisfy the following conditions:

     (a) ...

     (b) ..."

     Here, the word "must" appears once, which is counted as one in the first measurement but is counted as 2 in the second measurement (2 bullet points).

There are other papers heading in the same direction. A notable example is from Ash et al. (2022) where the authors count a much more complicated system of keywords (rather than just five like in RegData) and are restricted to specific sentence structures (not just the appearance of the rule-indicator words). However different they may be, in the end, the goal is similar: they are trying to measure the stock of rules. Once again, this is not error-free. Two different pieces of documents with the same number of rules might have different comprehensive timing due to the way they are written. Considering the three following texts: (i) "You cannot smoke inside the building except the empty room on the 4^{th} floor"; (ii) "You cannot smoke in the 1^{st} floor, 2^{nd} floor, 3^{th} floor, the dining hall and the common room in the 4^{th} floor, ... [listing all possible rooms]"; (iii) "You can only smoke in the special room, to see the list of special rooms, please refer to < Section 1.5>". Intuitively, they are all trying to achieve the same thing. However, the way they are written can hinder one's understanding, leading to different complexity levels. We move on to the next section.

4.1.3 New methods

The idea is to mimic how a human with a question goes into this body of text to answer that question. Based on the knowledge acquisition theory, the complexity of regulatory texts can be categorized into 3 different aspects: structure, language and interdependence (Katz and Bommarito, 2014). The first kind relies on a larger structure (either implying the vast and deep knowledge to be conveyed or born from the need to present a large amount of knowledge) contributing to the complexity of the texts. The second element measures the diversity of the language and concepts within the text - "it is more difficult for an individual to assimilate information in a corpus with high concept variance than one comprised of largely homogeneous material" (Katz and Bommarito, 2014). Finally, the interdependence can be accessed by the number of citations, either to itself (since the text itself is very large) or to other titles. The idea is that more references to other parts increase the complexity since the searcher needs the knowledge of not only the current part they reading but also the part to which the text refers.

To be more specific, the following shows a detailed description of how we define and measure each of the three aspects:

1. **Structure aspect.** This is measured based on the hierarchy of the regulatory texts.

    - Number of section (subpart) in each part of the title: **numSubPart**.
    - Number of elements below the section level: **levelBelowSection**. These are generally in the form of alphabetical letters (lower or upper cases) within the parenthesis (e.g. (a), (bb), etc.), numbers within the parenthesis (e.g. (1), (15), etc.), and Roman numbers within the parenthesis (e.g. (i), (xi), etc.). There are occasionally format errors but are accounted for (e.g. (a) but there is a new line and white space between the letter "a" and the parenthesis, which is not obvious on the webpage showcase but exists in XML files).
    - Adding these two numbers up, we obtain the total figures for the Structure complexity -**hierarchyCount**.

The higher these numbers are, the more complicated the text should be according to the

previous intuition.

2. **Language aspect.** We have the following measurements:

   - Shannon entropy: it is introduced for dataset creation in McLaughlin and Sherouse (2019) (although unofficially, very recent and not mentioned in any of their articles) but I have not seen it used in any empirical works. It is calculated as **shannonEntropy** using the following formula:

   $$Entropy_{d(i,y)} = -\sum_{w \in d} p_w log_2(p_w)$$

   Where $Entropy_{d(i,y)}$ is the entropy for document d(i,y), p_w is the probability the token w appears in the corpus (here, the corpus includes all parts of Title 40 in year y). The higher this number is, the more complicated the text seems to be (higher diversity in language)

   - Number of gerunds (**gerund** as the variable) - a verb as a noun (e.g. managing, persuading, etc.), tagged as "VGB" in SpaCy. The number of words/tokens per sentence: **num_words_sent** and **num_tokens_sent**. Normattiva (2016) suggests that these two are elements that make the text more complicated since it decreases the text's clarity. Longer sentences also tend to be more difficult to follow and harder to comprehend. The higher these numbers are, the higher the complexity should be.

3. **Interdependence aspect.** There are two main types of citations: internal (within the title) and external (to other titles). Note that there are also two different formats of citations within the title:

   - Those with an informal format where the reference target is within its current chapter/subchapter/volume or even part. These are usually in the form of "in accordance with § 51 of this chapter" or "refer to §§ 51.3 to 51.8". We omit any other informal citations to the same sections since the distance is too short to make any difference in the complexity. This is called **citInternal1**.

- Those that are formal - far away from each other and are in the form $< title >$ CFR $< part >$, which is officially mentioned on government websites. These are referred to as **citCFRInternal** and **citCFRExternal** depending on whether the $< title >$ part is equal to 40 or not. The total of these two is simply **citCFR**. Intuitively, these should all increase as the complexity increases.

4.1.4 Combined features using principal components

Since there are many measurements available, we will group them appropriately using principal components. It is more convenient to work with a few features, especially when it comes to graphing. Although there may be discrepancies between measurements, the principal component can group them nicely and give a reasonably consistent estimate of complexity, which can then be conveniently used for graphs and regressions. I will use four different sets of principal components based on the original variables:

- PCVol - including all features of the Volume-based measurements.
- PCRule - including all features of the RegData rule-based measurements.
- PCNew - including all features of the New methods.
- PCAll - including all of the above.

4.2 Other measurements

Complexity is not the only characteristic that we could learn from the regulatory text. As mentioned in section 2, there are two main ways of extracting other characteristics from a piece of text.

4.2.1 Pre-trained models

Due to time constraints, we will use only one model as a demonstration - the sentiment analysis. As mentioned before, as long as the pre-trained model is available, we could simply feed the raw text into this system and obtain the desired characteristics of the text.

The sentimental model we used here is from **nltk** package under the name **SentimentIntensityAnalyzer()**. We use the polarity score (compound) to obtain the degree of positive/negative sounding of the text. This sentiment score increases when the text sounds more positive and decreases (can go negative) when it has a negative tone. Using this feature, we could see, across years and industries, if there are some industries that receive more negative attention than others. This could potentially identify industries that have recently been more focused on the prohibitive side of regulations.

There is another method falling into the same category called text entailment - fill in the blank: Zero-Shot Classification. For more details on the mechanism and how the model was built, see Yin et al. (2019). It works similarly to the previous model, except this time it requires additional inputs. Specifically, three inputs are needed: the piece of text (in our case, d(i,y)), the set of keywords and the hypothesis chosen by the users. The output is how likely the keyword matches the hypothesis (i.e., fill in the blank) based on the document. If multi_label is true, the proportion will not add up to 1 (but individually, they must be between 0 and 1) and is simply the degree of likeliness for each keyword. Otherwise, if multi_label is false, all these proportions will add up to 1, analogous to a classification problem. For instance, if I want to know how much of the text is relevant to air, water or land aspects of the environment, I will have the following input:

- The document d(i,y)
- The hypothesis "This text is about {}"
- The set of keywords ["air", "water", "land"]
- multi_label = True

The output will be a set of 3 probabilities [80%, 40%, 5%], meaning that the hypothesis "This text is about air" is 80% likely to be true. This implies that this document d(i,y) is likely (80% certainty) to talk about the air aspect of the environment. Notice that the proportions are not adding up to 1. This setup results will be presented in the Result section as an example.

In general, any keywords can be selected. Keywords can be either based on intuitive hypothesis or topic modelling top keywords for each topic. This is purely based on the research questions.

There are two main limitations to this type of method (especially when dealing with text). First, the model itself is highly dependent on the data set on which it was trained. If there is a large discrepancy between it and our data (the CFR Title 40), the inference it makes may be misleading and untrustworthy. This is analogous to the extrapolation problem. Second, the time it takes to run just one year's worth of text for just Title 40 is extremely long (up to 4 hours). Great care must be taken when approaching the problem with these pre-trained models.

4.2.2 Unsupervised learning - Topic modelling

For text pre-processing, we will reference from Hickman et al. (2022), with an additional step of entity recognition - recognizing whether or not a certain phrase represents an organization, an event, a person, etc.

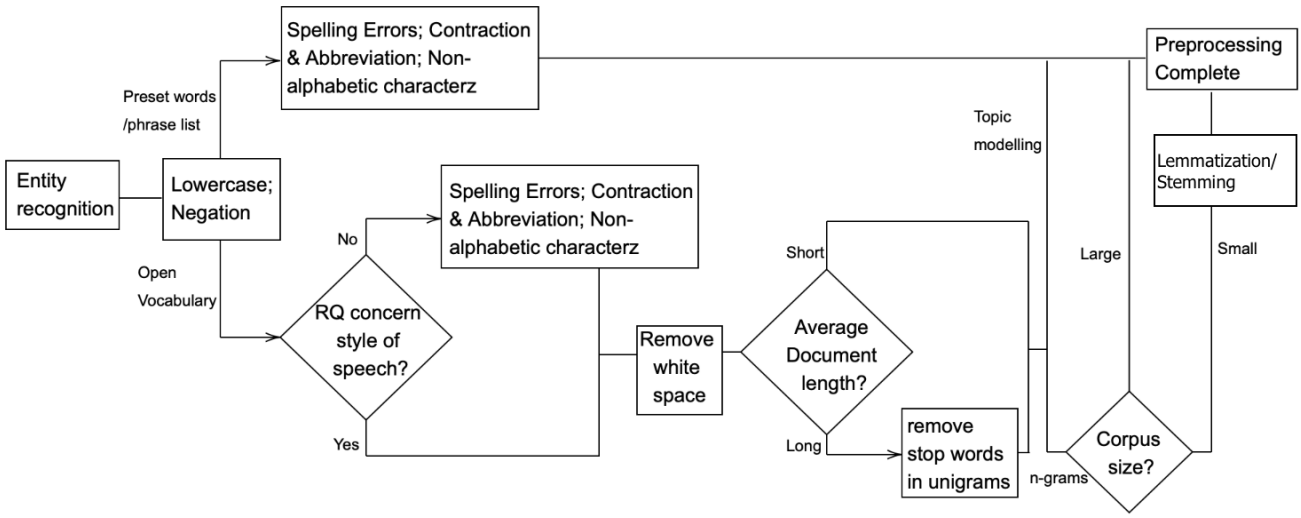


Figure 1: Text pre-processing diagrams

For this exercise, I will test different text pre-processing setups to answer the following question: Is this step really important for our particular data set? That is, is there a significant difference in the final results between these setups? The setups include

1. Guideline setup - includes steps that are for topic modelling. Since our research question does not concern with the style of speech, involves the use of topic modelling and the size of the texts is extremely large, we will include everything except the last two steps - n-grams

and lemmatization.

2. No 2-gram setup - includes all of those in the guideline setup but with lemmatization.

3. Everything setup - includes all of those in the guideline setup but with 2-gram and lemmatization.

4. Basic setup - is the everything setup but only includes tokens (words/phrases) that are nouns, verbs, adjectives, and adverbs (proper nouns such as entities are excluded)

A quick note on the lemmatization vs stemming: both are used to reduce inflectional forms and reduce the number of unique words in a text. The difference is that stemming a single word required no knowledge of the context, and thus, can produce erroneous results or miss out on those that are supposed to be similar. However, the upside of stemming is that it is easier to implement and runs faster, and the reduced accuracy may not matter for some applications. However, it is recognized in Hickman et al. (2022) that lemmatization is superior to stemming in most cases. Here are some examples taken from Wikipedia as to why:

- The word "better" has "good" as its lemma. This link is missed by stemming, as it requires a dictionary look-up.

- If you lemmatize the word "Caring", it would return "Care". If you stem, it would return "Car" and this is erroneous.

For the second step, we will use Latent Dirichlet allocation (LDA) as a popular method to address the topic modelling problem. The detailed model (math and technical mechanism) can be found in Blei et al. (2003). Essentially, what it does is take the pre-processed version of a raw text (obtained from the first step), and with the number of potential topics K (chosen by the researcher - a hyperparameter), output two matrices: a topic-term matrix and a document-topic matrix (see figure 2). The former shows which terms are most dominant for each topic, allowing the researcher to manually label the topic based on this information. Automatic labelling using word embedding and related methods is also emerging, but due to time constraints, we will not use this matrix. Our main focus is on the document-topic matrix where the topic proportion for each

document d(i,y) can be used to construct meaningful features - topic variance, higher numbers mean that topics are discussed in multiple sections of the texts rather than only concentrated on a few sections.


THE WALL STREET JOURNAL.

TERRORISTS DESTROY WORLD TRADE CENTER, HIT PENTAGON IN RAID WITH HIJACKED JETS

Nation Stands In Disbelief And Horror

What's News—

Business and Finance

World-Wide

Death Toll, Source of Devastating Attacks Remain Unclear; U.S. Vows Retaliation as Attention Focuses on bin Laden

Streets of Manhattan Resemble War Zone Amid Clouds of Ash

Hour of Horror Forever Alters American Lives

Attacks Will Force People To Make Adjustments In Ways Large and Small

Document-term matrix

Document i	Term j			
	Accord	Activ	Averag	...
News 1	3	9	4	...
News 2	6	4	16	...
...	...	...	...	...

topic modelling (LDA)

		Topics		
		Topic 1	...	Topic K
Terms	bank	0.14	...	0.02
	increas	0.34	...	0.26
	...	...	...	...

and

		Documents		
		News 1	...	News N
Topics	Topic 1	0.36	...	0.02
	...	...	...	...
	Topic 2	0.14	...	0.52

Figure 2: From text pre-processing to LDA

This feature is constructed similarly to how the Shannon entropy was constructed:

$$TopVar_{d(i,y)} = -\sum_{k=0}^{K} a_{k,d(i,y)} log_2(a_{k,d(i,y)}) \times numTokenConsidered \tag{1}$$

where $TopVar_{d(i,y)}$ is the topic variance for the document d(i,y), $a_{k,d(i,y)}$ is the probability relevant of document d(i,y) to topic k. Note that we need to weight the measure with the number of tokens considered for a specific document d(i,y) since (i) different text pre-processing will give different amounts of tokens, (ii) some documents have much more information than others, thus it is reasonable to weight them with their volumes, and (iii) this topic variance measurement without the weight does not take into account the volumes of text since all topic proportions must add up to 1, unlike other measures we have before.

The question is how to pick the optimal number of topics K. For this, we use the criteria called "coherence score". This score measures the degree of semantic similarity between high-scoring words in the topic. There is another popular criterion called "perplexity score". It captures how

surprised a model is by new data and is measured as the normalized log-likelihood of a held-out test set. However, recent studies have shown that predictive likelihood (perplexity) and human judgment are often uncorrelated, and sometimes even slightly anticorrelated. Therefore, we will use the coherence score for the rest of the analysis. Basically, we ran the model with the grid of $K = [2, 4, 6, \ldots, 30]$, picked the one with the highest coherence score, checked the neighbour K (± 1), and finalized the results.

There are other hyper-parameters. α represents document-topic density, and we choose $\alpha = 1/K$ (asymmetric option). η represents topic-word density, and we choose 0.01. Chunk size is the number of documents to be used in each training chunk, update_every determines how often the model parameters should be updated and passes is the number of training passes. The optimal K can be refined further by fine-tuning α and η specifically. Notice that, ideally, we would like to run the method using all texts from Title 40 from all year at once. However, due to computational constraints, we will only run each year at once, and obtain the optimal topic numbers K_y for each year y. This method also introduces the randomness element: Different runs will give different sets of models, thus different coherence scores and possibly different optimal K_y. Therefore, we set the seed to 100. Another fine-tuning would be running this method for a few different seeds and averaging the results to get less noisy features.

Another option that helps to choose the optimal K when the output variable is available is to use simple OLS. This is presented in more detail in the next section. Essentially, we will use industrial performances as outputs, topic proportion as independent variables and choose the number of topics based on information criteria (AIC in this case).

In the result section, we would like to see the following:

- We want to see if the optimal number of topics chosen by the coherence score exhibits any peculiar patterns (and if different text pre-processing gives different answers).
- How does the topic variance behave across years and industries (similar to how complexity measures are analyzed)?

4.2.3 Others

We also do the following as a means to better understand some aspects of the regulatory texts.

1. We look at word clouds using the same text pre-processing with an additional step of scaling accounting for the word's rarity across all documents. The size of a word represents its importance. This should give us a general idea of the most important keywords in Title 40 for each year. This will also give us an idea of the differences between each setup, which may give us some insight as to which text pre-processing is more appropriate for our research question.

2. We also count the number of [Reserved] portions - either appearing within the texts or being branded for the entire sections.

3. Finally, we also look at the distribution of external citations made in Title 40 - Environmental Protection. This should give us a good look at which other titles this Title 40 depends the most on.

4.3 Simple prediction model

Before addressing the regression model utilizing the complexity measurement, we need to obtain the industry-specific measurements.

As mentioned before, for each piece of document d(i,y), we obtain a certain measurement called m(i,y) (this could be the number of words, the number of external citations, etc.). We also obtain the industry proportion for each document Ind(p,i,y) with p as the industry code. Thus, the industry-specific measurements will simply be estimated as:

$$m(p, y) = \sum_{i \in y} m(i, y) \times Ind(p, i, y) \tag{2}$$

4.3.1 Simple Fixed-Effect model

We deal with panel data with each industry (p) as entities and year (y) as time:

$$Output_{py} = \beta_0 + \beta_1 PCVol_{1py} + \beta_1 PCRule_{1py} + \beta_1 PCNew_{1py} + \alpha_p + \epsilon_{py} \tag{3}$$

where the sets of PC here are the four main ones discussed before. In this case, we only use the first principal component for each set. Even when other principal components (second to fifth) are included, the results do not change. This is the same for the following regression:

$$Output_{py} = \beta_0 + \beta_1 PCAll_{1py} + \alpha_p + \epsilon_{py} \tag{4}$$

The reason (which will be discussed in more detail in the Results section) is that the majority of the correlation to the original variables is from the first principal components. Hence, it is sufficient to include only the first PC. As a robust check, including other components does not change the final conclusions.

4.3.2 Choosing number of topics with outputs

Different from the previous section where we choose the optimal number of topics only for each year at once, we will just use very simple OLS regression and choose models with the highest AIC.

$$Output_p = \beta_0 + \sum_{k=1}^{K} \beta_k * a_{pk} + \epsilon_p \tag{5}$$

5 Results, discussion and limitation

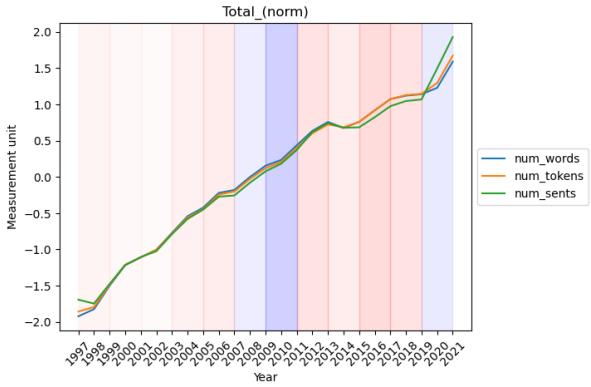
5.1 Complexity measurement

There are a few notes about the graphs. First, when the line graphs have a blue and red background, it represents the party divisions of the U.S. House of Representatives. Red is when the Republicans dominate, and the degree of redness also illustrates how unbalanced it is. Blue is the

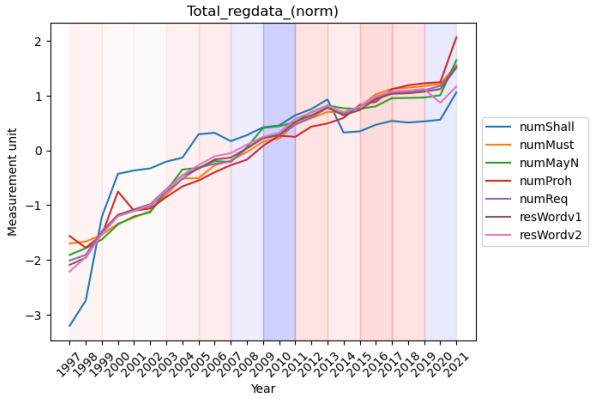
opposite - this is when the Democrats dominate. The numbers are taken from par. Second, if the measurements are fundamentally different in units, the graph will show the normalized version instead for better visualization.

5.1.1 Measurements across years and industries

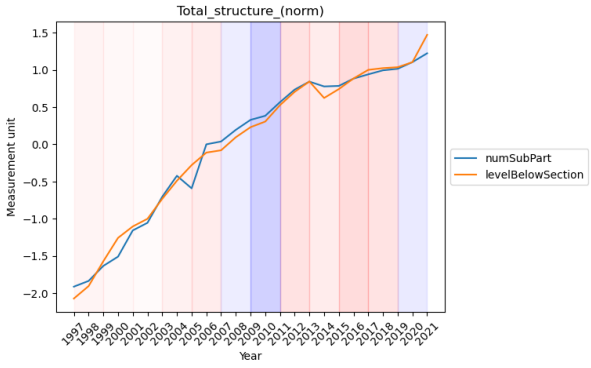
First of all, we would like to see how different complexity measurements behave when industries are not of interest.



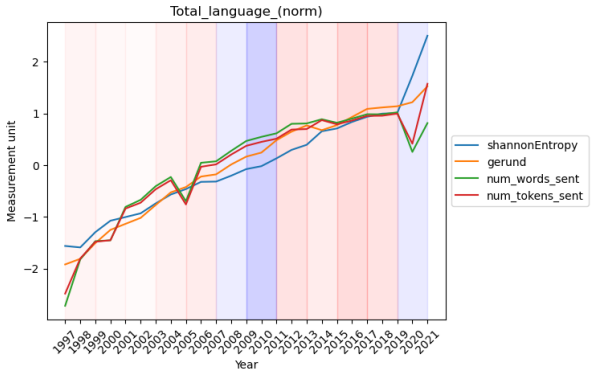
(a) Volume-based



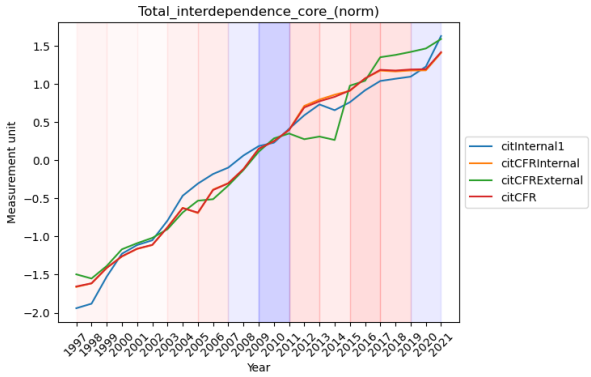
(b) Rule-based



(c) New-Structure



(d) New-Language



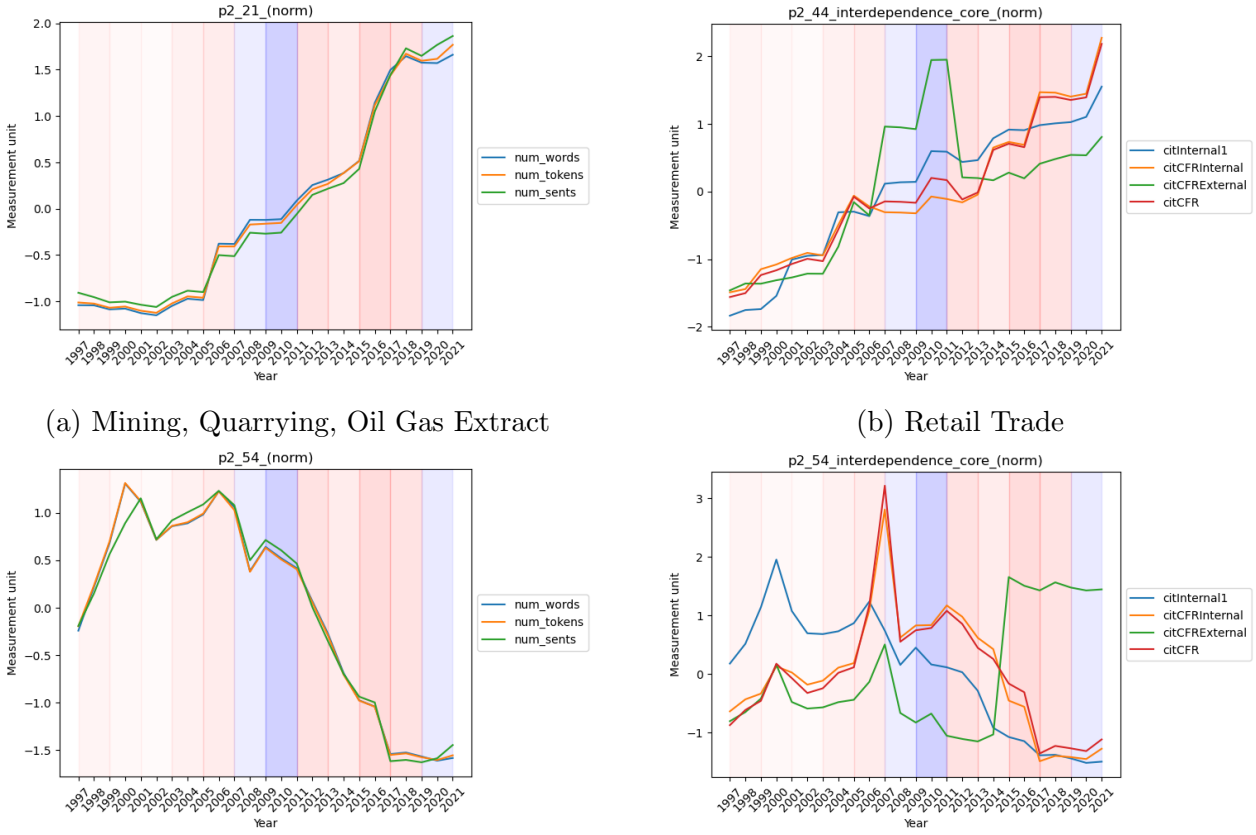
(e) New-Interdependence

Figure 3: Measurements patterns regardless of industries

In general, all measures are closely related: they all increase over time at a very similar rate. This means that if industries are not of interest, the traditional ways of measuring complexity can be quite comprehensive (relative to the new measures). Beyond the noise, there are some striking patterns. For the rule-based methods, the graph shows an outlier in the number of "shall" counted in the text. It shows that there may be a shift in the language used to describe the rules: a

decrease in the word "shall" while there is an increase in other words (especially "must" - see the non-normalized version in Appendix B, figure 14b). This also implies potential drawbacks of the rule-based method: since they rely on a set of words to detect regulatory rules, an unaccounted-for shift in language usage may slip under the radar and bias the complexity measurements.

However, the stories change when you look at specific industries rather than the overall picture. For example, the Mining, Quarrying, and Oil and Gas Extraction sector appears to be showing movement following the change in party divisions. (See figure 4a).



(a) Mining, Quarrying, Oil Gas Extract

(b) Retail Trade

(c) Professional, Scientific, Tech Service

(d) Professional, Scientific, Tech Service

Figure 4: Measurements patterns across different industries

Complexity has spiked during periods of change from Republican to Democratic dominance, but then remains slightly up or flat during Democratic dominance. Is there an underlying dynamic here that is unique to this sector? Is it possible that Democrats, who are more concerned about climate change, are able to influence regulation in this sector, signalling the coming dominance of this party? Answering these questions requires further analysis (which is not undertaken here).

Another good example of a difference that could not be captured by the traditional method is the number of citations to external titles. The Retail Trade sector experienced an abnormal increase in citations to other titles during the years 2007-2012, but then returns to the original trend and joins other measurements. The Professional, Scientific and Technical Services sector experiences what can be called a potential "migration" to other titles. There is a steady decrease in all measurements for this sector since around 2012, but the number of external citations has increased, suggesting a potential flow to another title. This might make sense, as the current title of interest is Environmental Protection, and this field - which involves the use of chemicals for research purposes - might be mainly concerned with the agreement of use, which is what Title 2 (Grants and Agreements) is about.

Other behaviours can be listed below (see Appendix B for a selected few figures 15):

- Utility sector shows a similar trend in Total but with a dip into a U-shape from 2013-2016.
- Paper Manufacturing sector follow the Total trend until 2011 when it starts to flat out and decrease (the structural measure shows the strongest patterns).
- Nonmetallic Mineral Product Manufacturing sector seems to dip down a little right before (1-2 years) the Democratic Party dominates.
- Food Manufacturing sector follows the same trend as Total except for the external citations, suggesting that some of the regulations may be brought in from other titles.
- Transportation Equipment Manufacturing sector shows two strange pits: 2006-2009 and 2013-2017.

Here are only the selected few that are looked into. It is clear that examining the complexity of law without the context of industry differences might not paint a complete picture. This could also explain why although regulatory complexity is generally on the rise, different industries react differently to this trend (mixed results).

5.1.2 First Principal components across year and industries

We first examine the principal components for all features. The figure shows that, among the theoretically relevant industries, most of the environmental regulation falls on the manufacturing sector (see figures 5). Within the manufacturing sector, Petroleum and Coal Products Manufacturing and Chemical Manufacturing receive the most attention, much more than Mining, Quarrying, and Oil and Gas Extraction. For theoretically irrelevant industries, retail trade stands out the most (see figure 6)

The first principal components can vary whether it is based on traditional or new methods. An example here is when the number of external citations alters the trend of complexity in the new method (not too drastically). The Retail Trade, as mentioned before, has a drastic change in the interdependence aspect that is not captured in the traditional methods. The change, although mitigated by other aspects of the new complexity method, is still slightly reflected in the final product (see figures 7)

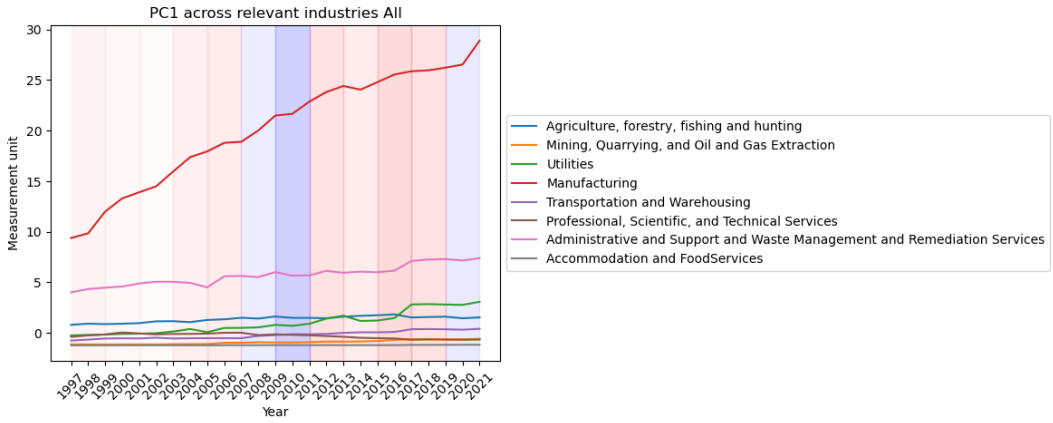
5.2 Other potential characteristics of regulation data

5.2.1 Pre-trained model

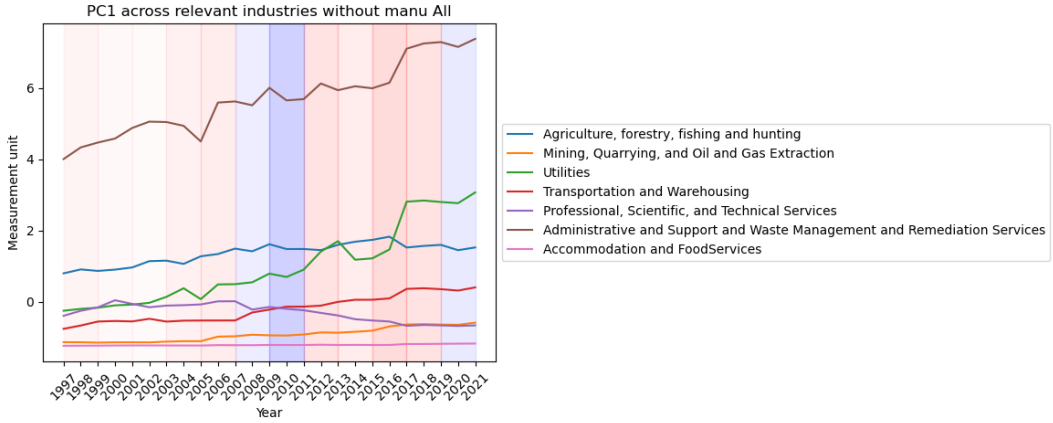
In general, the results are volatile (when normalized) but fluctuate around a constant level (applied to most industries). When viewed on an actual scale, they are relatively flat for the irrelevant industries (see figures 8). On the same set of graphs, some industries, mainly those that are the most impacted by the regulation here, have a slight increase in sentiment value over the year, suggesting that some of the volume effects may have carried over through neutral texts. The only special note is that on a sentimental level, the Waste management sector is on par with the Manufacturers, suggesting that even though the complexity may be different, the sentiment value might be the same.

However, given the shortcomings of this type of method, this shows that either little information can be extracted from it, or the model does not work well against regulatory texts such as this.

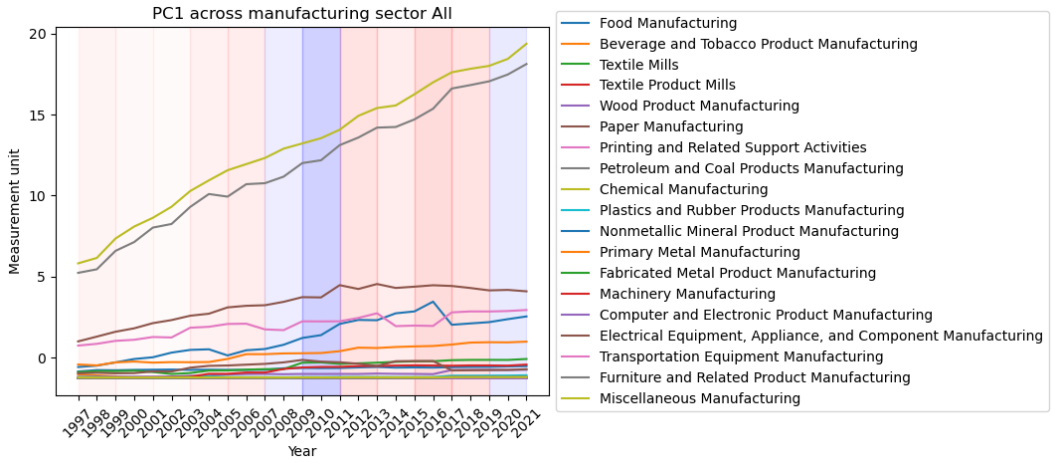
The results of using the zero-shot classification are shown in the figure 9. These themes (air,



(a) All theoretically relevant sectors



(b) All theoretically relevant sectors without Manufacture



(c) All theoretically relevant sectors within Manufacture

Figure 5: Impact of Title 40 complexity on different industries

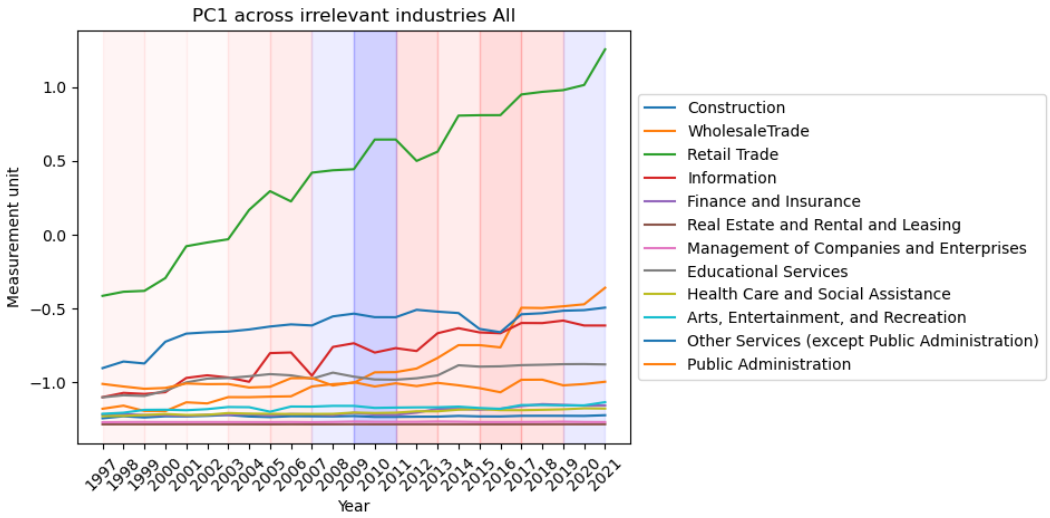


Figure 6: All theoretically irrelevant sectors

water, and land) are relatively constant over time. Air also takes the largest share, followed by water and finally land.

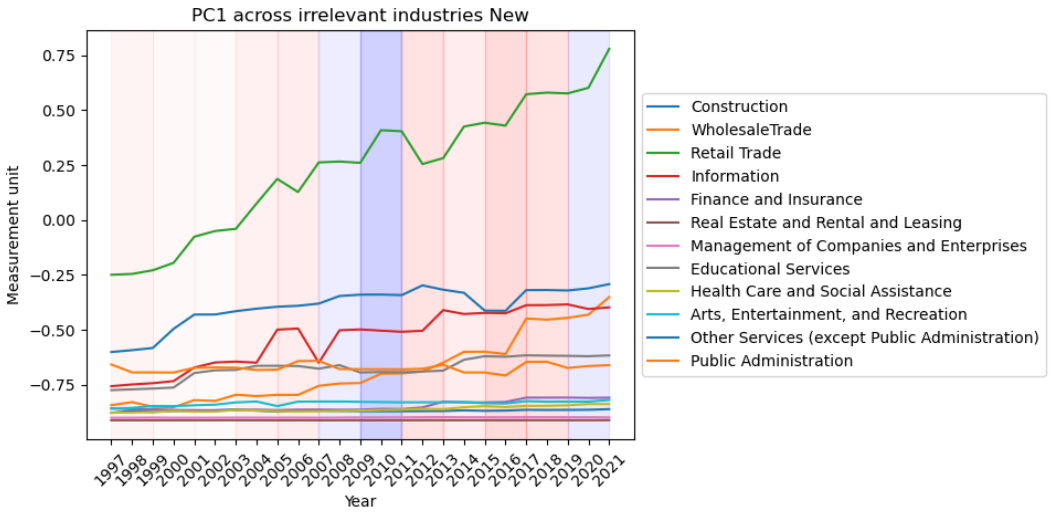
5.2.2 Topic modeling

First, we examine the optimal number of topics across the year.

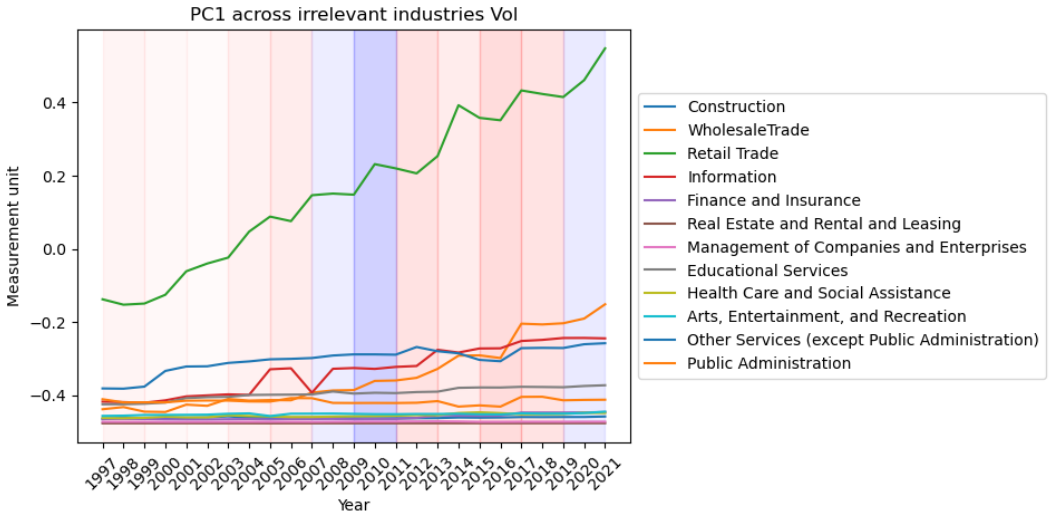
Table 1: Number of topics chosen across different methods

	Basic	Every -thing	No 2-grams	Guide -lines	Firms	Establish -ment	Employees	Payroll
count	25.00	25.00	25.00	25.00	23.00	23.00	23.00	23.00
mean	16.40	15.68	14.48	16.68	5.30	5.39	1.30	1.26
std. dev	5.28	6.24	6.20	5.38	4.72	4.58	0.56	0.54
min	8.00	5.00	4.00	8.00	1.00	1.00	1.00	1.00
25%	14.00	11.00	10.00	12.00	2.00	2.00	1.00	1.00
50%	15.00	16.00	13.00	16.00	3.00	4.00	1.00	1.00
75%	19.00	20.00	20.00	20.00	7.00	7.50	1.50	1.00
max	26.00	28.00	26.00	28.00	19.00	19.00	3.00	3.00

The first four columns of Table 1 show the optimal number of topics chosen using the coherence score for different text pre-processing setups. The average number of topics over the year is relatively similar to each other - around 14 to 17 topics. However, they are extremely volatile with a standard error of around 5-6 topics. This leads to two different conclusions: (i) the results fluctuation could be due to the randomness of the method itself, or (ii) there is some hidden

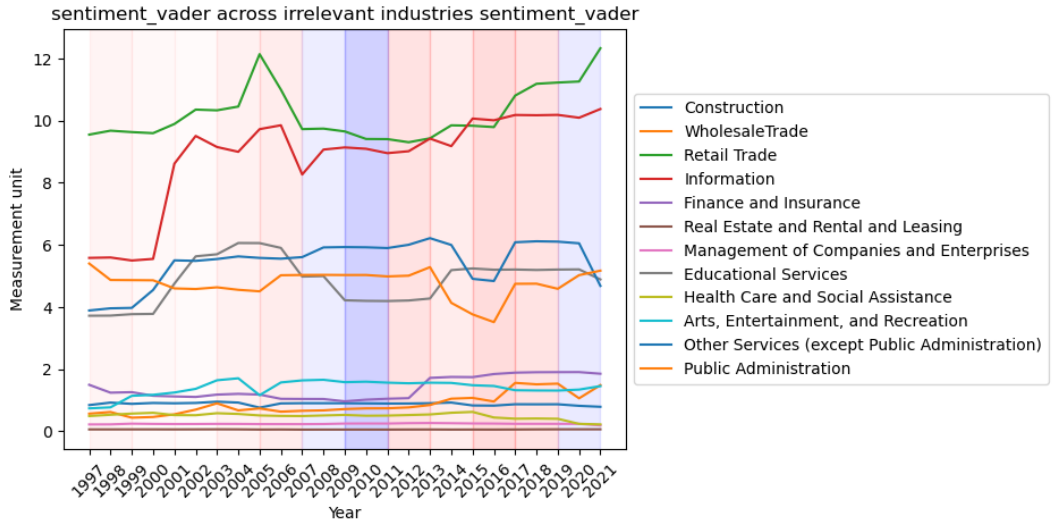


(a) New method

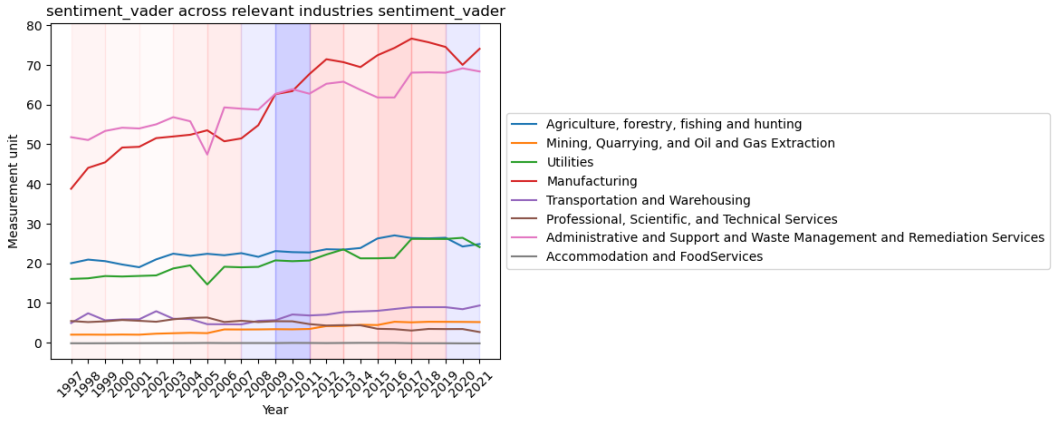


(b) Traditional method

Figure 7: Irrelevant industries New vs Traditional methods



(a) Irrelevant industries



(b) Relevant industries

Figure 8: Sentimental values across industries

mechanism that influences such changes. Personally, I lean more towards the former hypothesis as there is physically no way to change topic focuses on an annual or bi-annual basis. Such fluctuations can be mitigated by running through many different seeds and averaging up the numbers, which will have to be left to future work as it takes a lot of time to run. However, we need to acknowledge that the spikes and pits for all setups exhibit strangely consistent (to some extent) cycles, even though the length and depth are different across setups. (see Appendix B, figure 18).

Some other notes on the results using the coherence score: even though they can look very volatile and random, there are patterns here that could make up a meaningful story. Intuitively, whenever a text preprocessing method increases the number of **unique** tokens, the number of

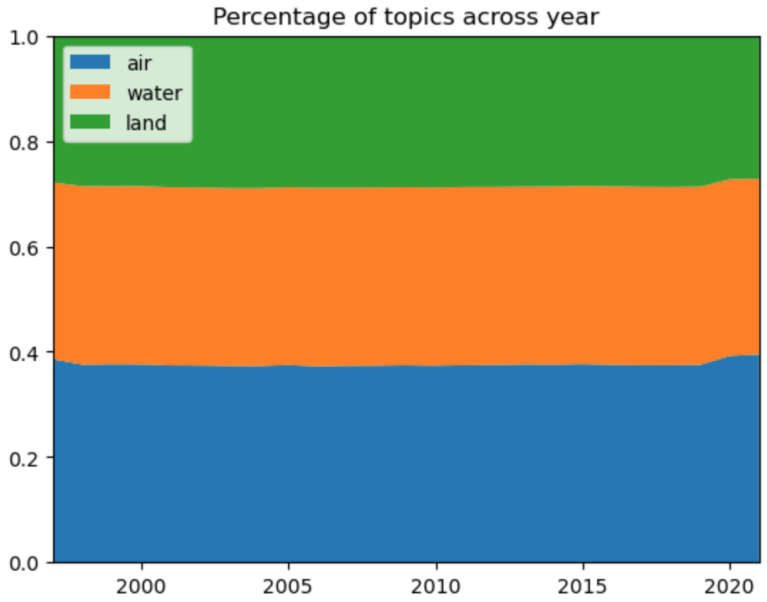


Figure 9: Proportion of air, water and land aspect over the year

topics should increase on average. This is true in the case of not applying the 2-gram procedure (so words like "economic growth" are counted as one token rather than "economic" and "growth" as 2 separate tokens, thus having 3 unique tokens instead of 2) versus those that do (average topic is 14.48 vs 15.68). This is the same for when lemmatization is not included (hence, words like "good" and "better" are two different tokens instead of both being referred to as "good" on both with lemmatization). However, the "everything" setup, which includes more tokens than those in the basic setup (thanks to the addition of extra entity nouns) gives a lower average topic number. This could be suggesting there might be a merge - a concentration - in entity over time. However, due to the volatile nature of these series (extremely high standard deviations), it is hard to say with confidence that the average reflects such conclusions.

The last four columns of the table 1 show the optimal number of topics using outputs from industry performances. The results show drastic differences compare to those using coherence score. When using the number of firms and the number of establishments, the average number of topics selected is much lower, hovering around 5 with an extremely high standard error. The patterns on these series are very similar to the counterpart using coherence scores. On the contrary, the number of employees and payroll have extremely low picks on the optimal number of topics, suggesting that the topic proportion is likely irrelevant to these outputs.

As for the topic variance measurements, these figures are just as volatile for most industries. When the topic variance is not weighted by the volumes of the texts, the behaviours, regardless of industries, are the same as those in the optimal topic over the years (See Appendix B, figure 16). This could be due to the volatility of the optimal number of topics chosen passing onto these figures or the number of topics addressed in the regulation staying consistent throughout the year.

When using the weighted measurement instead (See Appendix B, figure 17), most of the industries behave similarly to how they do with the number of words counted, thus we can say that most of the characteristics are carried over by the weight itself. However, that also shows those with higher text volumes tend to contain more topics, which is reasonable since they have a wider variety across their own part - containing a lot of sections within. Thus even though each section might be homogeneous in their reading, the diversity in the topic at part-level can still be high. Notice that diversity in topics might not reflect complexity similar to how entropy is looked at. A better look at this aspect would be seeing the figures at section-level since if they indeed contain a lot of topics, the text itself should address multiple problems or just be simply vague. There is an exception where the topic variance just hovers around a constant and not trending up: Retail trade. Although it strongly fluctuates at the end, signalling "trending" upward, it is unclear.

The weighted topic variances' patterns are surprisingly consistent across relevant industries for all text pre-processing setups. The patterns are very similar to each other despite being different overall across the years. This may be thanks to the consistency in the weight. The rest of the industries are too noisy to discern any patterns.

5.2.3 Others

There are other aspects of texts that we could look through.

Keywords - word clouds: For each setup, we will be looking at word clouds for six years (1997, 2002, 2007, 2012, 2017, 2021), which can be found in Appendix B - figures 19, 20 and 21. Since those in setup with everything and without the 2 grams are quite similar, we will only report for the former.

There are slight changes in the evolution of keywords over time. For the basic setup, keywords such as administrator, pollutant, state and effluent were important in earlier years but diminish and replaced by other emerging words with very similar functions such as emission, engine and requirement. This can be seen as a shift in the language used over the year, supporting the theory that any methods involving counting specific keywords will have to be more cautious. The story is quite similar to the setup including all text pre-processing methods except for the appearance of EPA - the U.S. Environmental Protection Agency. It is quite small in size, but gets a little bigger as the year goes on. It is especially large and even became one of the most important keywords in the setup without 2-gram and without lemmatization. For this version of text pre-processing, notice that word pairs such as (requirement, requirements) and (include, including) are counted as different instead of as one (even if they have the same meaning) since lemmatization is not used here. This means the word EPA does not become more important, but rather it is all other words that got undermined due to their other versions getting counted differently. Thus, it is recommended to use lemmatization when the research questions do not care about styles of speech, which is in contrast to Hickman et al. (2022) where their diagram recommends not using lemmatization and multi-grams when using topic modelling (2-gram text pre-processing might not be that relevant in this case). It is suggested that lemmatization should be used when topic modelling is involved while it should be used when sentiment analysis is involved instead.

The number of [Reserved] might partly reflect the uncertainty of the regulation in theory. According to figure 10, the most significant spike is from Part 49: Tribal clean air act (over 12,000 sections reserved, only less than 500 sections filled in 2021).

Finally, we want to take a closer look at citations to external titles. The citations to other titles show that there is a significant movement out, which means that considering only texts within the title may not be sufficient. This also opens up the possibility of creating another layer on top of the current new complexity measure: including the complexity of other titles and weighting it by the number of external citations made. It should be noted, however, that not all citations carry "real weight", meaning they might be there but could be ignored depending on who reads them, not the mention that some are for the sake of formality and thus do not affect the overall

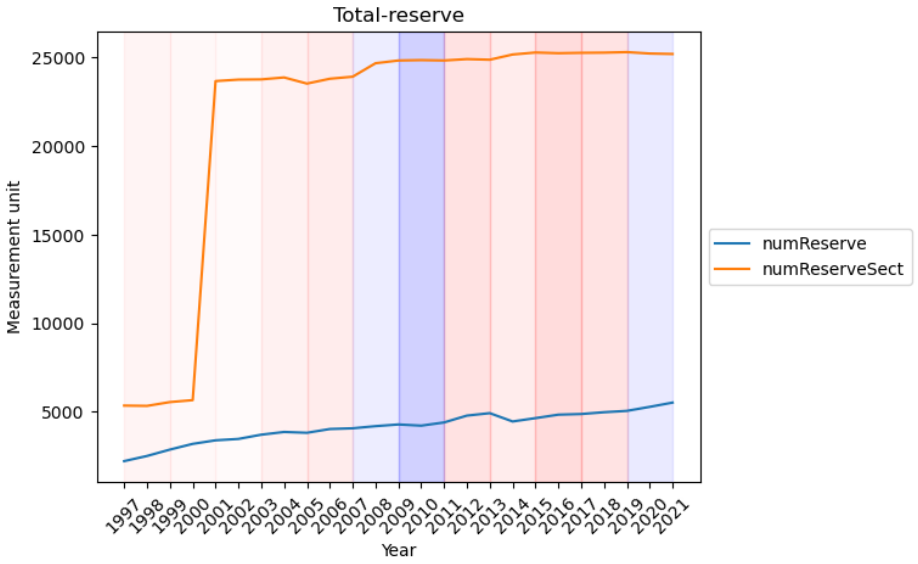


Figure 10: Number of reserve sections and subsections

complexity. Finer insights require more in-depth knowledge.

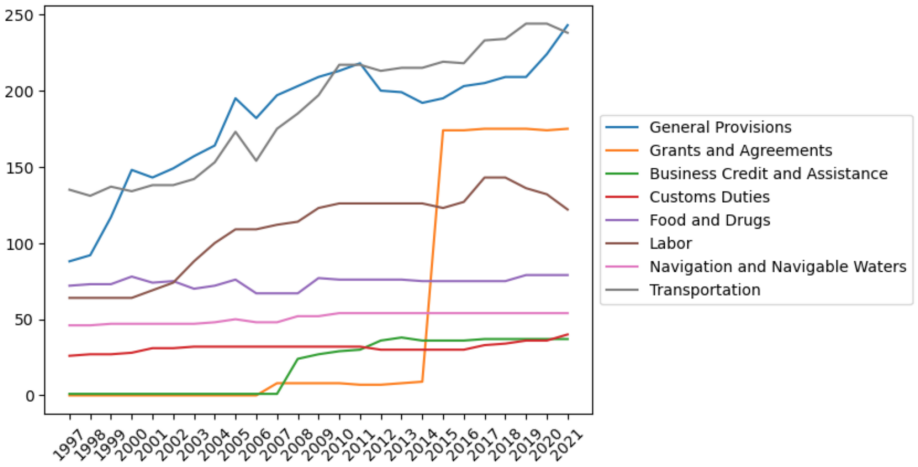


Figure 11: Number of external citations in more details

Figure 11 shows that the majority of the citations are made to the General provisions and the Transportation, meaning if the goal is to explore the complexity of environmental regulation for the transportation sectors, Title 40 might not be the best source. There is also a spike in the citations to Grants and Agreements - Title 2, which coincidentally follows the same trend as the Professional, Scientific and Tech service sectors. One can hypothesize that since most of the relevant environmental issues from this sector are applications for the use of chemicals, they are considered irrelevant to Title 40 and have been shifted to Title 2. Nevertheless, there is a change

and "migration" to other titles that should be considered.

5.3 Complexity measurements on Industry performances

Before jumping into the naive regressions, there is evidence that only considering the first principal components (PC) is sufficient. The table showing the correlation between the principal components (up to the fifth component) and the original variables can be found in Appendix C, tables 5 and 4. The result suggests that PC1 is the primary source of variation (all correlations above 0.9, some can go up to 1.0). Although for the second and third PC of the new method, their correlation can go up to 0.3, this is only for two variables involving language and the rest fall below 0.2 and are relatively insignificant compared to the first PC. Including both the second and third PC does not change the conclusions.

The results for all of the regressions can be found in tables 2 and 3. Columns 1, 3, and 5 are regressions of three variables PCVol, PCReg, and PCNew on the number of firms and establishments (Table 2), number of employees and annual payroll (Table 3). Columns 1 and 2 are restricted to small businesses while Columns 3 and 4 are restricted to medium-sized businesses. The last two columns focus on large businesses with more than 500 employees each.

Table 2: Simple regressions of different PC on different outcomes

Business size	**Small (<20 employees)**		**Medium (<500 employees)**		**Large (>500 employees)**	
Number of Firms						
const	105,496.00	104,221.30	118,524.99	117,049.78	899.97	883.18
	(517.60)	(258.66)	(592.25)	(375.46)	(7.16)	(5.06)
PCVol	-23,140.08		-22,390.26		-260.30	
	(13,075.50)		(14,277.26)		(191.91)	
PCReg	-31,685.77		-40,183.54		-448.45	
	(13,578.57)		(15,096.07)		(168.21)	
PCNew	38,272.01		43,998.55		505.74	
	(15,648.34)		(17,130.87)		(188.09)	
PCAll		-1,415.89		-2,147.57		-17.86
		(961.93)		(1,396.32)		(18.58)
Adj. R-square	0.038069	0.005668	0.043749	0.010301	0.055396	0.007621
No. of obs	2275	2275	2275	2275	2283	2283
Number of Establishment						
const	106,457.75	105,174.21	127,439.91	125,863.11	23,584.93	23,354.20
	(520.43)	(258.09)	(629.11)	(382.08)	(251.80)	(95.41)
PCVol	-23,193.36		-24,703.10		-10,466.81	
	(13,156.04)		(15,287.23)		(5,983.98)	
PCReg	-31,986.51		-42,346.31		-725.20	
	(13,750.71)		(16,294.18)		(6,630.14)	
PCNew	38,544.59		47,065.37		7,022.23	
	(15,817.92)		(18,732.96)		(7,821.54)	
PCAll		-1,425.16		-2,223.39		230.88
		(959.83)		(1,420.93)		(350.21)
Adj. R-square	0.038553	0.005737	0.041856	0.009381	0.004302	-0.000155
No. of obs	2275	2275	2275	2275	2283	2283

Table 3: Simple regressions of different PC on different outcomes (cont.)

Business size	**Small (<20 employees)**		**Medium (<500 employees)**		**Large (>500 employees)**	
Number of employees						
const	420,857.88	416,399.90	1,168,344.44	1,152,637.74	1,239,718.38	1,195,936.79
	(1,897.27)	(1,591.84)	(7,882.82)	(9,802.73)	(32,088.55)	(19,123.90)
PCVol	-33,926.38		37,051.18		-534,291.39	
	(38,281.44)		(162,847.85)		(762,016.64)	
PCReg	-148,477.64		-648,775.03		-1,285,138.95	
	(52,485.74)		(180,088.26)		(643,940.76)	
PCNew	131,093.59		446,979.28		1,302,245.79	
	(52,653.55)		(129,626.85)		(834,555.12)	
PCAll		-10,139.52		-57,277.49		-68,393.62
		(5,919.99)		(36,455.95)		(70,197.19)
Adj. R-square	0.03849	0.015576	0.070649	0.041611	0.094554	0.024633
No. of obs	2275	2275	2275	2275	2283	2283
Annual Payroll ($1,000)						
const	15,371,966.40	14,985,712.75	46,907,075.52	45,519,674.19	66,121,543.93	62,562,473.01
	(211,952.99)	(44,040.77)	(746,496.67)	(173,683.84)	(2,216,050.32)	(496,782.05)
PCVol	-13,973,303.39		-51,485,134.89		-111,559,508.81	
	(5,834,646.39)		(20,655,380.36)		(56,001,791.15)	
PCReg	-3,965,466.67		-12,263,239.86		-47,581,067.95	
	(4,439,788.04)		(15,427,028.95)		(39,903,878.39)	
PCNew	12,020,840.98		43,156,803.30		111,333,700.32	
	(6,508,582.83)		(22,335,778.91)		(61,306,197.98)	
PCAll		385,712.70		2,024,279.33		4,406,528.20
		(168,795.31)		(674,271.01)		(1,945,682.98)
Adj. R-square	0.025617	0.002792	0.033209	0.007816	0.049583	0.009441
No. of obs	2252	2252	2237	2237	2232	2232

When the number of firms is the output, both the volume-based and the rule-based principal components are negatively correlated with output (only marginally significant for the former and very significant for the latter). The new method, on the other hand, is highly positively correlated. In terms of overall complexity, the creation and destruction of small and medium firms are more strongly (negatively) affected than that of large firms, where the correlation is insignificant because two forces cancel each other out.

The same patterns can be observed for the number of establishments as output within small and medium enterprises. For large firms, however, all significant relationships are now insignificant, making the overall effect insignificant, but not for the same reason as before. The output of the number of employees generally follows the same pattern.

Finally, regarding the annual payroll behaviours, individual patterns are very much the same as in the number of firms except the significant level for volume-based and rule-based are swapped here. Furthermore, the overall impact of complexity through PC of all variables appears positive (and very significant) instead of negative across all sizes. This variable is, however, relatively noisier (measurement noises) than the rest of the outputs, which can have an impact on the final conclusions.

In conclusion, the new and traditional measures show different explanatory patterns for some industry performances across sizes of enterprises. The overall effects, in general, are more severe in small and medium-sized businesses than in large enterprises when it comes to business destruction and creation (thus the establishment and the number of employees that come with it).

6 Conclusion

6.1 Results summary

When it comes to measuring complexity directly from the regulatory text, simply counting the volume of text or the number of rules can be misleading. This is especially true when cross-citation is taken into account. We also found concrete evidence of differences between sectors. These should be taken into account, as the aggregate level at the national level does not give the

full picture.

Findings from the regressions also suggest that the principal components of the new measure could potentially explain some economic phenomena that the traditional variables such as volume-based and rule-based could not. Further investigation is needed to explore their causal ability against the traditional as well as the conventional after-effect variables.

On the technical side, different text pre-processing setups can significantly alter the final results of topic modelling. Great care must be taken when using this method.

6.2 Limitations and future research direction

From the external citation results, it is a good idea to consider the complexity measures of other titles and incorporate them into the current complexity measures.

Situations such as adding more regulations to clarify a situation are not accounted for (all of the currently considered measures will mostly all increase). To address this, we could use the litigation cases to increase/decrease the complexity measure based on the increase/decrease in cases involving specific parts of the regulations. The degree of change could depend on length, number of meetings and discussions, etc. (all provided by the potential dataset).

Potential output tied to models that allow for causal inference - there is a need to look at political/government/large organization movements. There is also a need to expand current industry performance pools to gain more insight.

It would be tremendously helpful to compare features extracted from texts themselves versus those after-effects type variables. This could answer some questions such as: can text-based features explain patterns in economic data that conventional methods cannot? If not, which after-effect variable, in particular, is the most equivalent to these text-based characteristics? This could be helpful since regulation text adjustment comes first before any after-effect variables start to change, giving advantages in terms of prediction usage and delay patterns that could be learned from.

References

Party Divisions of the House of Representatives, 1789 to Present | US House of Representatives: History, Art & Archives. URL https://history.house.gov/Institution/Party-Divisions/Party-Divisions/.

How Does Grammarly Work? | Grammarly Spotlight, December 2019. URL https://www.grammarly.com/blog/how-does-grammarly-work/.

Boragan Aruoba and Thomas Drechsel. Identifying Monetary Policy Shocks: A Natural Language Approach. CEPR Discussion Papers 17133, C.E.P.R. Discussion Papers, March 2022. URL https://ideas.repec.org/p/cpr/ceprdp/17133.html.

Elliott Ash, Massimo Morelli, and Matia Vannoni. More Laws, More Growth? Evidence from U.S. States, April 2022. URL https://papers.ssrn.com/abstract=4095044.

Jonathan Benchimol, Sophia Kazinnik, and Yossi Saadon. Text mining methodologies with r: An application to central bank texts. *Machine Learning with Applications*, 8:100286, 2022. ISSN 2666-8270. doi: https://doi.org/10.1016/j.mlwa.2022.100286. URL https://www.sciencedirect.com/science/article/pii/S2666827022000202.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar 2003. ISSN 1532-4435.

Pontus Braunerhjelm and Johan E. Eklund. Taxes, tax administrative burdens and new firm formation. *Kyklos*, 67(1):1–11, 2014. ISSN 1467-6435. doi: 10.1111/kykl.12040. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/kykl.12040. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/kykl.12040.

Antonio Ciccone and Elias Papaioannou. Red Tape and Delayed Entry. *Journal of the European Economic Association*, 5(2/3):444–458, 2007. ISSN 1542-4766. URL https://www.jstor.org/stable/40005048. Publisher: Oxford University Press.

Bentley Coffey, Patrick A. McLaughlin, and Pietro Peretto. The cumulative cost of regulations. *Review of Economic Dynamics*, 38:1–21, 2020. ISSN 1094-2025. doi: https://doi.org/10.1016/j.red.2020.03.004. URL https://www.sciencedirect.com/science/article/pii/S1094202520300223.

Steven J. Davis. Regulatory Complexity and Policy Uncertainty: Headwinds of Our Own Making. *SSRN Electronic Journal*, 2015. ISSN 1556-5068. doi: 10.2139/ssrn.2723980. URL http://www.ssrn.com/abstract=2723980.

Giuseppe Di Vita. Institutional quality and the growth rates of the Italian regions: The costs of regulatory complexity. *Papers in Regional Science*, 97(4):1057–1081, 2018. ISSN 1435-5957. doi: 10.1111/pirs.12290. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/pirs.12290. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pirs.12290.

Dana Foarta and Massimo Morelli. Complexity and the Reform Process. page 57, October 2020.

Raquel Fonseca, Paloma Lopez-Garcia, and Christopher A Pissarides. Entrepreneurship, start-up costs and employment. *European Economic Review*, 45(4):692–705, May 2001. ISSN 0014-2921. doi: 10.1016/S0014-2921(01)00131-3. URL https://www.sciencedirect.com/science/article/pii/S0014292101001313.

Nancy Gallini. Do patents work? Thickets, trolls and antibiotic resistance. *The Canadian Journal of Economics / Revue canadienne d'Economique*, 50(4):893–926, 2017. ISSN 0008-4085. URL https://www.jstor.org/stable/48581980. Publisher: [Wiley, Canadian Economics Association].

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as data. *Journal of Economic Literature*, 57(3):535–74, September 2019. doi: 10.1257/jel.20181020. URL https://www.aeaweb.org/articles?id=10.1257/jel.20181020.

Gabriele Gratton, Luigi Guiso, Claudio Michelacci, and Massimo Morelli. From Weber to Kafka: Political Instability and the Overproduction of Laws. *American Economic Review*, 111(9):

2964–3003, September 2021. ISSN 0002-8282. doi: 10.1257/aer.20190672. URL https://pubs.aeaweb.org/doi/10.1257/aer.20190672.

Ralph Grishman. Information extraction. *IEEE Intelligent Systems*, 30(5):8–15, sep 2015. ISSN 1541-1672. doi: 10.1109/MIS.2015.68. URL https://doi.org/10.1109/MIS.2015.68.

Gillian K. Hadfield. The Price of Law: How the Market for Lawyers Distorts the Justice System. *Michigan Law Review*, 98(4):953–1006, 2000. ISSN 0026-2234. doi: 10.2307/1290336. URL https://www.jstor.org/stable/1290336. Publisher: The Michigan Law Review Association.

Louis Hickman, Stuti Thapa, Louis Tay, Mengyang Cao, and Padmini Srinivasan. Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1):114–146, 2022. doi: 10.1177/1094428120971683. URL https://doi.org/10.1177/1094428120971683.

Daniel Martin Katz and M. J. Bommarito. Measuring the complexity of the law: the United States Code. *Artificial Intelligence and Law*, 22(4):337–374, December 2014. ISSN 1572-8382. doi: 10.1007/s10506-014-9160-8. URL https://doi.org/10.1007/s10506-014-9160-8.

Keiichi Kawai, Ruitian Lang, and Hongyi Li. Political Kludges. *American Economic Journal: Microeconomics*, 10(4):131–158, November 2018. ISSN 1945-7669. doi: 10.1257/mic.20150242. URL https://www.aeaweb.org/articles?id=10.1257/mic.20150242.

Stephen Kirchner. Federal legislative activism in Australia: a new approach to testing Wagner's law. *Public Choice*, 153(3/4):375–392, 2012. ISSN 0048-5829. URL https://www.jstor.org/stable/23326376. Publisher: Springer.

Patrick A. McLaughlin and Oliver Sherouse. RegData 2.2: a panel dataset on US federal regulations. *Public Choice*, 180(1):43–55, July 2019. ISSN 1573-7101. doi: 10.1007/s11127-018-0600-y. URL https://doi.org/10.1007/s11127-018-0600-y.

Giuseppe Nicoletti, Stefano Scarpetta, and Philip R. Lane. Regulation, Productivity and Growth: OECD Evidence. *Economic Policy*, 18(36):9–72, 2003. ISSN 0266-4658. URL https://www.

jstor.org/stable/1344653. Publisher: [Center for Economic Studies, Maison des Sciences de l'Homme, Centre for Economic Policy Research, Wiley].

Mirjana Pejić Bach, Živko Krstić, Sanja Seljan, and Lejla Turulja. Text mining for big data analysis in financial sector: A literature review. *Sustainability*, 11(5), 2019. ISSN 2071-1050. doi: 10.3390/su11051277. URL https://www.mdpi.com/2071-1050/11/5/1277.

Jie Yang, Soyeon Caren Han, and Josiah Poon. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64(5):1161–1186, 2022. doi: 10.1007/s10115-022-01665-w. URL https://doi.org/10.1007/s10115-022-01665-w.

Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *CoRR*, abs/1909.00161, 2019. URL http://arxiv.org/abs/1909.00161.

Appendix A. Extra information on the datasets

SUBCHAPTER A—GENERAL

Pt. 1

PART 1—STATEMENT OF ORGANIZATION AND GENERAL INFORMATION

Subpart A—Introduction

Sec.

Subpart B—Headquarters

Subpart C—Field Installations

Authority: 5 U.S.C. 552.
Source: 50 FR 26721, June 28, 1985, unless otherwise noted.

Figure 12: Snippet of Title 40 part 1

Subpart A—Introduction

§ 1.1 Creation and authority.

Reorganization Plan 3 of 1970, established the U.S. Environmental Protection Agency (EPA) in the Executive branch as an independent Agency, effective December 2, 1970.

§ 1.3 Purpose and functions.

The U.S. Environmental Protection Agency permits coordinated and effective governmental action to assure the protection of the environment by abating and controlling pollution on a systematic basis. Reorganization Plan 3 of 1970 transferred to EPA a variety of research, monitoring, standard setting, and enforcement activities related to pollution abatement and control to provide for the treatment of the environment as a single interrelated system. Complementary to these activities are the Agency's coordination and support of research and antipollution activities carried out by State and local governments, private and public groups, individuals, and educational institutions. EPA reinforces efforts among other Federal agencies with respect to the impact of their operations on the environment.

§ 1.5 Organization and general information.

(a) The U.S. Environmental Protection Agency's basic organization consists of Headquarters and 10 Regional Offices. EPA Headquarters in Washington, DC maintains overall planning, coordination and control of EPA programs. Regional Administrators head the Regional Offices and are responsible directly to the Administrator for the execution of the Agency's programs within the boundaries of their Regions.

(b) EPA's Directives System contains definitive statements of EPA's organization, policies, procedures, assignments of responsibility, and delegations of authority. Copies are available for public inspection and copying at the Management and Organization Division, 1200 Pennsylvania Ave., NW., Washington, DC 20460. Information can be obtained from the Office of Public Affairs at all Regional Offices.

(c) EPA conducts procurement pursuant to the Federal Property and Administrative Services Act, the Federal Procurement Regulations, and implementing EPA regulations.

§ 1.7 Location of principal offices.

(a) The EPA Headquarters is in Washington, DC. The mailing address is 1200 Pennsylvania Ave., NW., Washington, DC 20460.

(b) The addresss of (and States served by) the EPA Regional Offices (see § 1.61) are:

(1) Region I, U.S. Environmental Protection Agency, room 2203, John F. Kennedy Federal Building, Boston, MA 02203. (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont.)

(2) Region II, U.S. Environmental Protection Agency, Room 900, 26 Federal Plaza, New York, NY 10278. (New Jersey, New York, Puerto Rico, and the Virgin Islands.)

(3) Region III, U.S. Environmental Protection Agency, 841 Chestnut Street, Philadelphia, PA 19107. (Delaware, Maryland, Pennsylvania, Virginia, West Virginia, and the District of Columbia.)

(4) Region IV, U.S. Environmental Protection Agency, 345 Courtland Street NE., Atlanta, GA 30365. (Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee.)

Figure 13: Snippet of Title 40 part 1 (cont)

According to the glossary, which can be found at https://www.census.gov/programs-surveys/susb/about/glossary.html, the following are some important facts:

- FIRM: number of firms/enterprises in each industry. An enterprise is a business organization consisting of one or more domestic establishments that were specified under common ownership or control. The enterprise and the establishment are the same for single-establishment firms. Each multi-establishment company forms one enterprise - the enterprise employment and annual payroll are summed from the associated establishments.

- ESTB: number of establishments in each industry. Establishment counts represent the number of locations with paid employees at any time during the year. This series excludes government establishments except for wholesale liquor establishments (NAICS 4248), retail liquor stores (NAICS 44531), tobacco stores (NAICS 453991), book publishers (511130), monetary authorities – central bank (NAICS 521110), federally-chartered savings institutions (NAICS 522120), federally-chartered credit unions (NAICS 522130), hospitals (NAICS 622), gambling industries (NAICS 7132), and casino hotels (NAICS 721120).

- EMPL: number of employments that are paid. Paid employment consists of full- and part-time employees, including salaried officers and executives of corporations, who are on the payroll in the pay period including March 12. Included are employees on paid sick leave, holidays, and vacations; not included are sole proprietors and partners of unincorporated businesses.

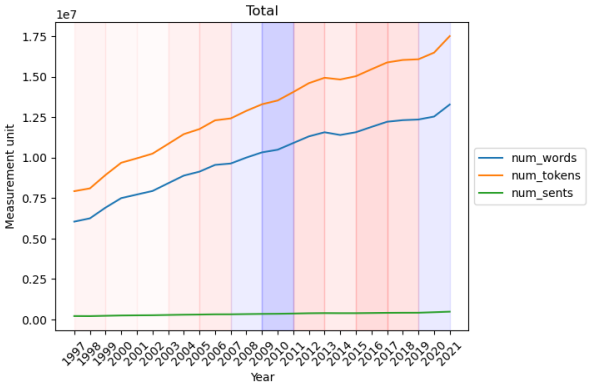
- PAYR: includes all forms of compensation, such as salaries, wages, commissions, dismissal pay, bonuses, vacation allowances, sick-leave pay, and employee contributions to qualified pension plans paid during the year to all employees.

- SIZE: enterprises' sizes, divided into 3 big categories - small (less than 20 employees), medium (less than 500 employees) and large (more than 500 employes).

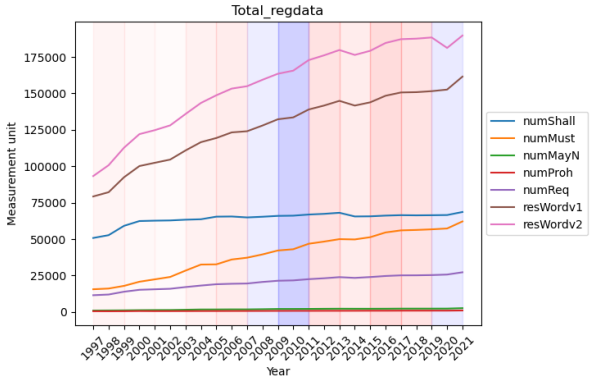
There are some issues with the industry codes:

- Among the 2-digit and 3-digit industry codes, SUSB excludes crop and animal production (NAICS 111,112), rail transportation (NAICS 482), Postal Service (NAICS 491), private households (NAICS 814), and public administration (NAICS 92).
- There are some industries that are different in code but have the same (or even identical) names while others are the combination of two different industries. We match them up manually. For instance, NAICS 513 is a combination of NAICS 515 and NAICS 517. This is fixed after some time.
- Industry codes under unclassified (99 and 95) are omitted.
- These mismatched occurs differently every year and thus, manual checking is required. The number of incompatibilities decreases drastically over the year.

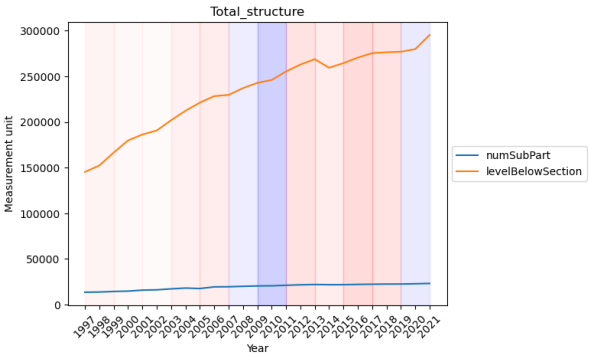
Appendix B. Extra graphs



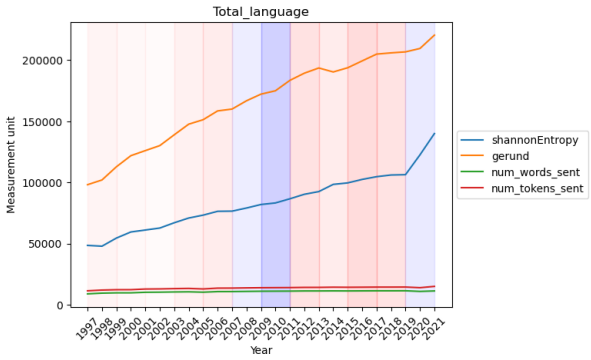
(a) Volume-based



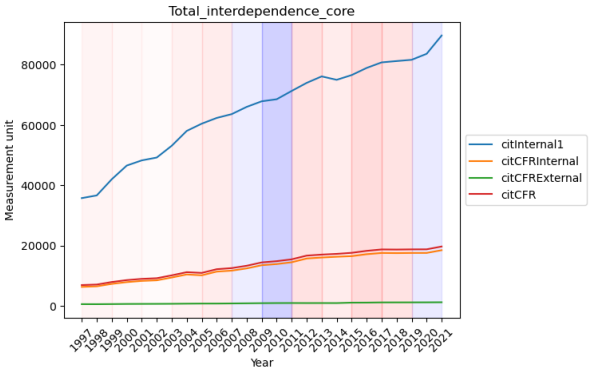
(b) Rule-based



(c) New-Structure

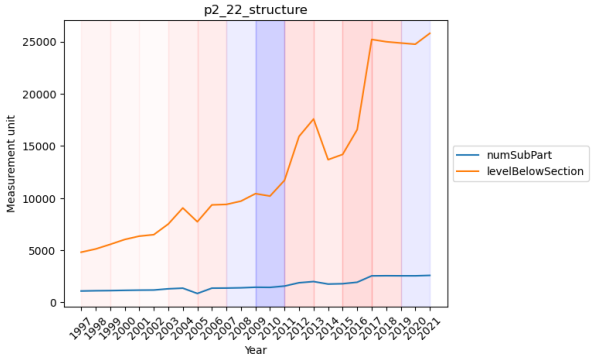


(d) New-Language

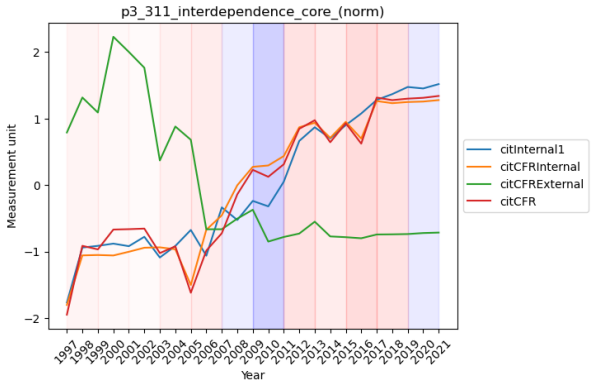


(e) New-Interdependence

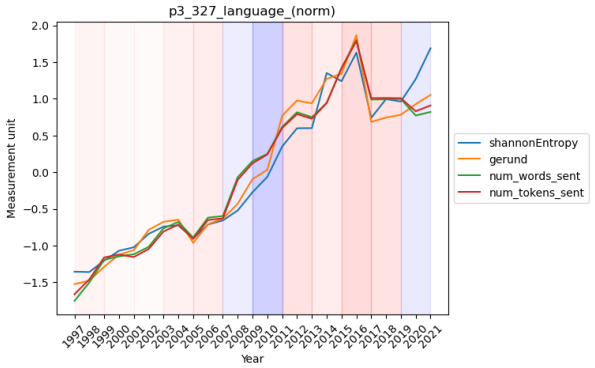
Figure 14: Measurements patterns regardless of industries not normalized



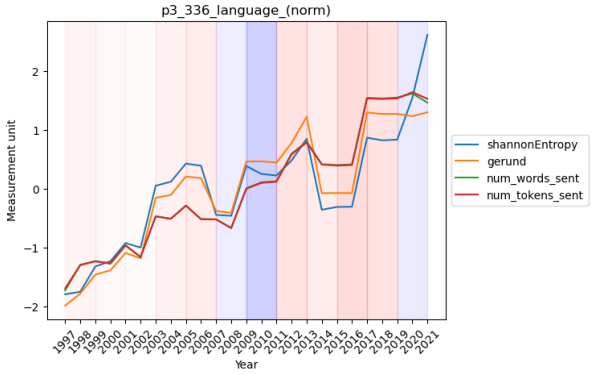
(a) Utilities



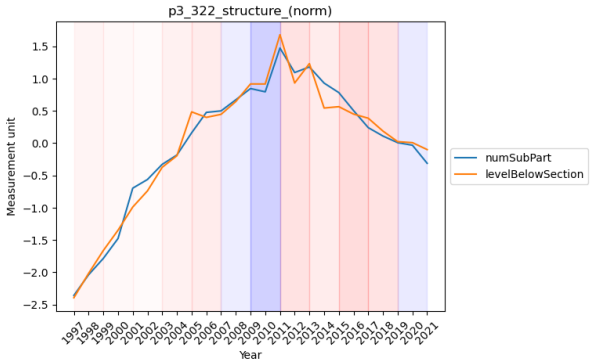
(b) Food Manufacturing



(c) Nonmetallic Mineral Product Manufacturing

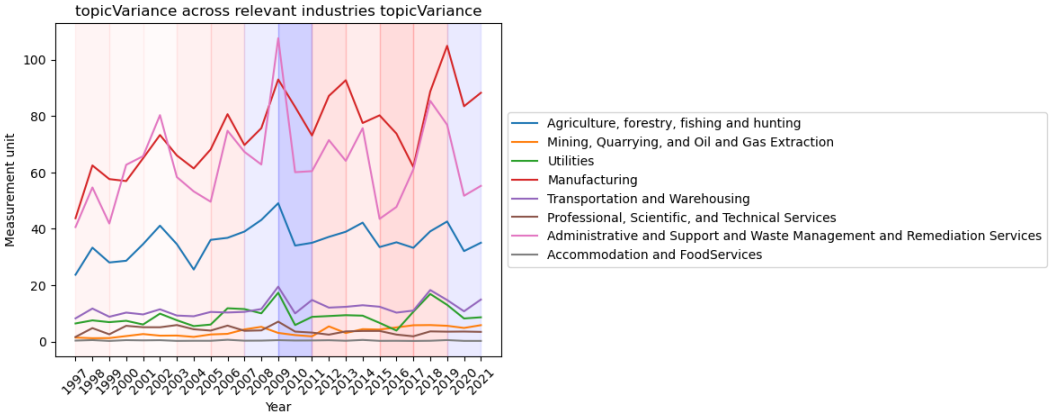


(d) Transportation Equipment Manufacturing

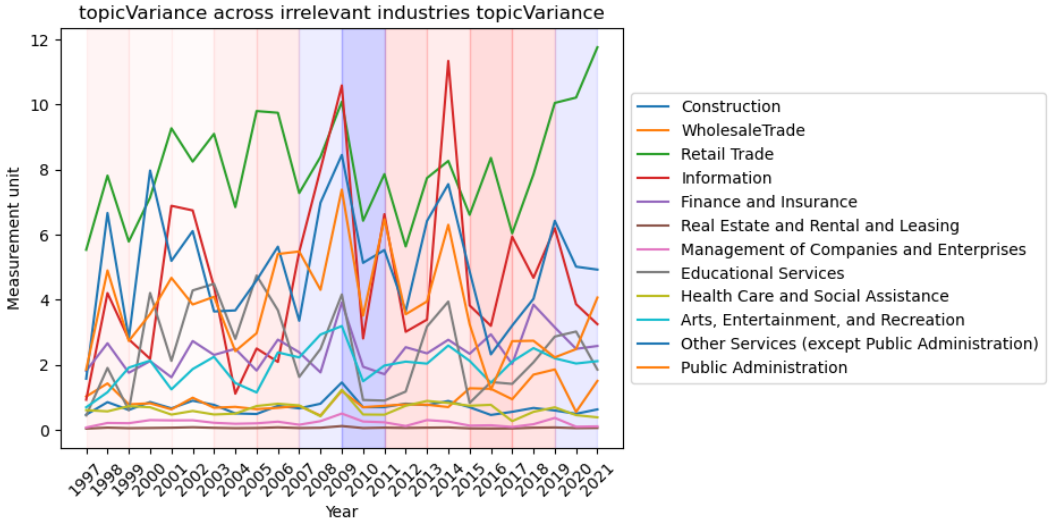


(e) Paper Manufacturing

Figure 15: Measurements patterns across different industries extra

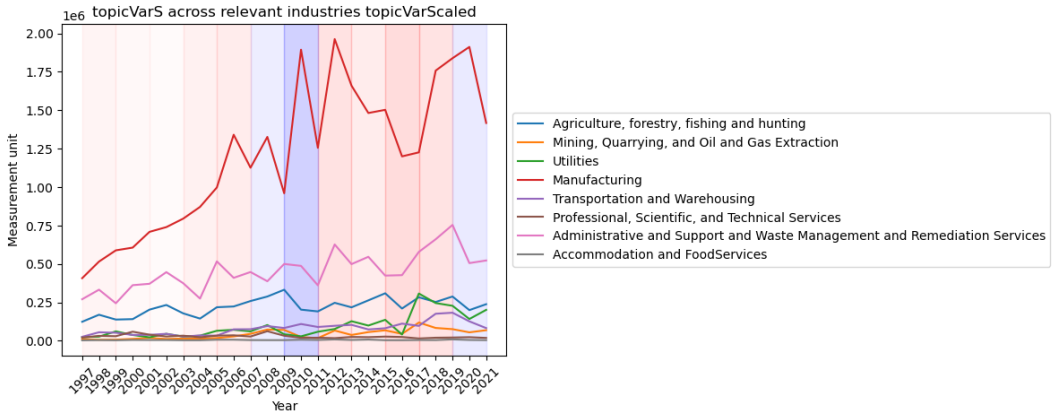


(a) Relevant industries

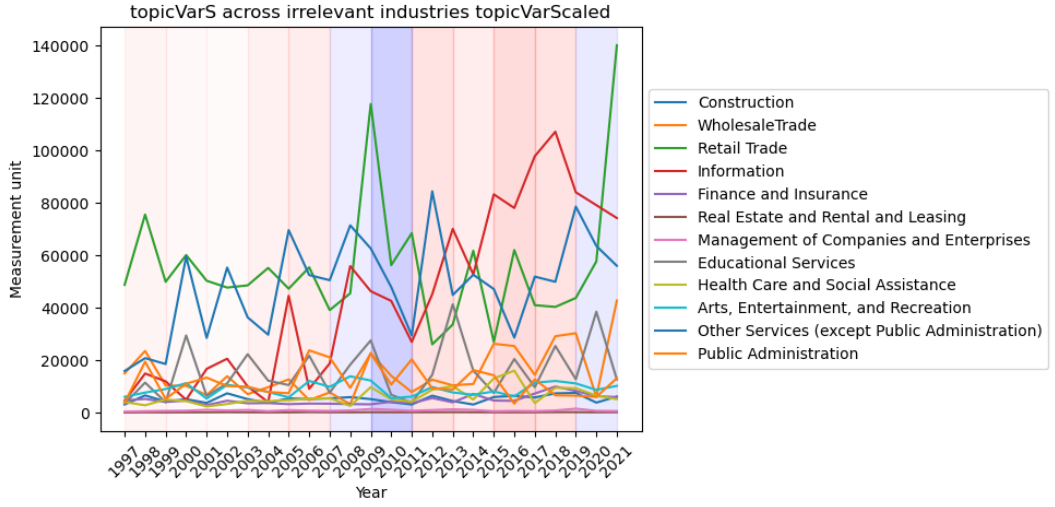


(b) Irrelevant industries

Figure 16: Topic variance without scaling



(a) Relevant industries



(b) Irrelevant industries

Figure 17: Topic variance with scaling

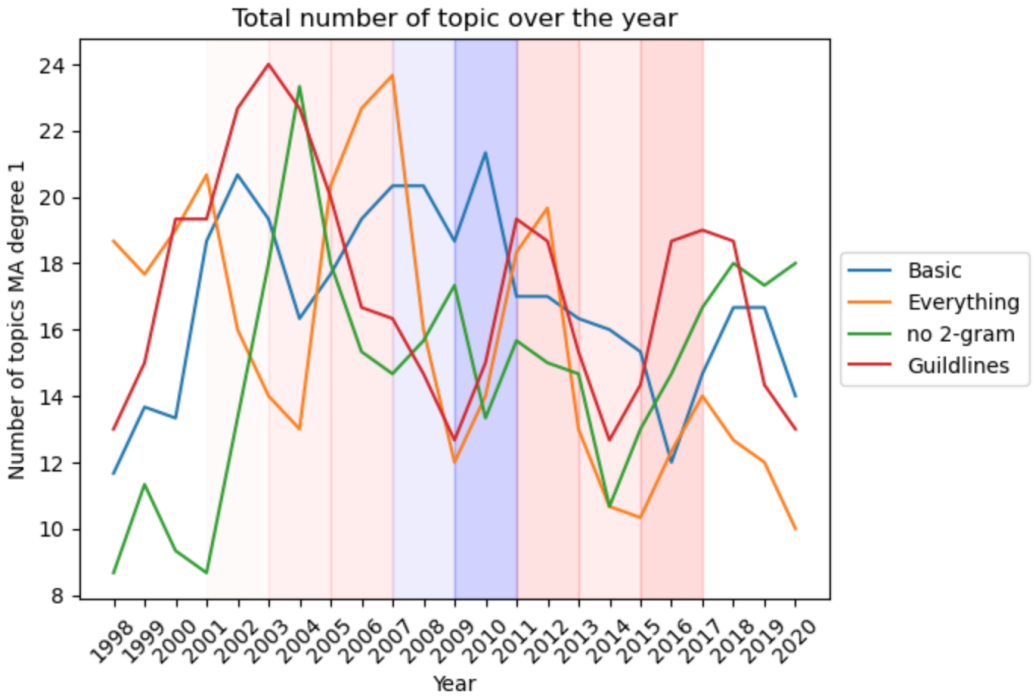
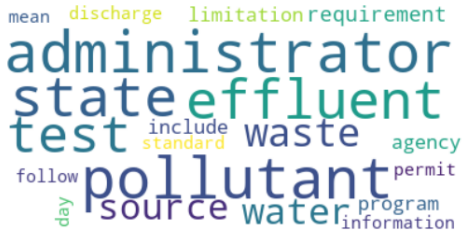


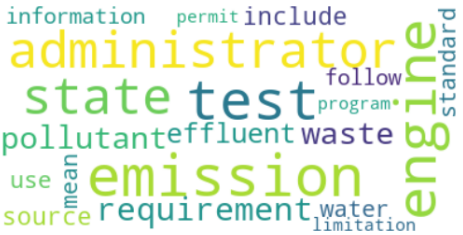
Figure 18: Number of topics chosen across different methods in Graph



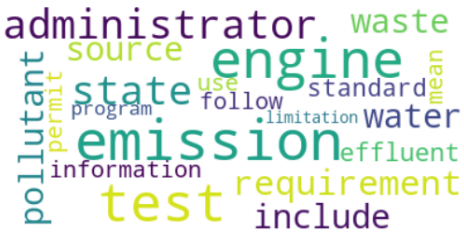
(a) 1997

(b) 2002

(c) 2007



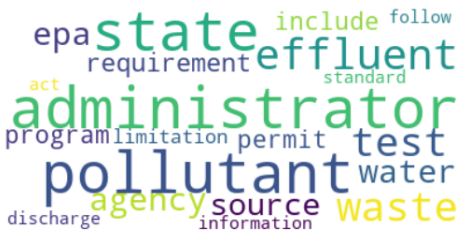
(d) 2012



(e) 2017

(f) 2021

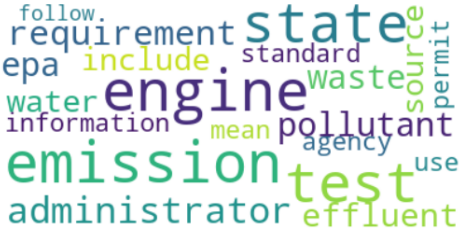
Figure 19: Keywords - word clouds for setup with no entities



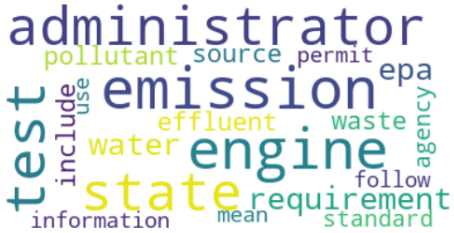
(a) 1997

(b) 2002

(c) 2007



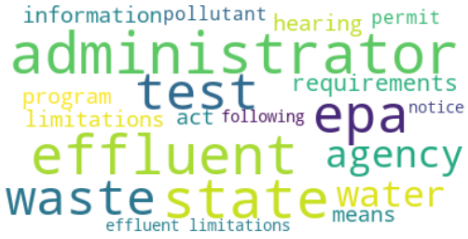
(d) 2012



(e) 2017

(f) 2021

Figure 20: Keywords - word clouds for setup with everything



(a) 1997

(b) 2002

(c) 2007

(d) 2012

(e) 2017

(f) 2021

Figure 21: Keywords - word clouds for setup with no 2-grams and no lemmatization

Appendix C. Extra tables

Table 4: Principal components and their correlation with original variables

Principle components for All variables

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
numSubPart	0.994	0.079	-0.011	-0.042	0.006
num_words	0.996	-0.087	0.027	0.001	-0.001
numShall	0.984	-0.018	0.168	-0.000	-0.019
numMust	0.979	-0.143	-0.126	-0.021	0.025
numMayN	0.966	0.049	-0.072	0.241	0.010
numReq	0.988	-0.146	0.043	0.007	0.001
numProh	0.992	0.062	-0.050	0.013	-0.026
resWordv1	0.995	-0.081	0.044	-0.003	-0.000
resWordv2	0.995	-0.080	0.042	-0.012	0.008
citInternal1	0.982	-0.175	0.034	-0.014	0.007
citCFR	0.988	-0.049	-0.125	-0.058	0.028
citCFRInternal	0.986	-0.060	-0.128	-0.060	0.037
citCFRExternal	0.977	0.132	-0.073	-0.015	-0.121
levelBelowSection	0.993	-0.097	0.058	-0.002	0.000
gerund	0.996	-0.082	0.011	-0.006	0.004
shannonEntropy	0.987	-0.138	-0.006	0.004	-0.009
num_tokens	0.995	-0.092	0.026	-0.000	-0.004
num_sents	0.994	-0.099	0.023	0.020	-0.004
num_words_sent	0.910	0.411	0.008	-0.021	-0.019
num_tokens_sent	0.913	0.402	0.004	-0.024	-0.022
hierarchyCount	0.995	-0.085	0.053	-0.005	0.001

Table 5: Principal components and their correlation with original variables (cont.)

Principle components for Volume-based variables					
	PC1	**PC2**	**PC3**	**PC4**	**PC5**
num_words	1.000	-0.015	0.005	-	-
num_tokens	1.000	-0.009	-0.005	-	-
num_sents	1.000	0.025	0.001	-	-
Principle components for Rule-based variables					
numShall	0.985	-0.079	0.154	-0.009	-0.013
numMust	0.983	-0.026	-0.179	0.010	-0.016
numMayN	0.969	0.238	0.017	-0.069	-0.002
numReq	0.994	-0.086	-0.028	-0.052	0.037
numProh	0.987	0.080	0.023	0.134	0.013
resWordv1	0.998	-0.058	0.008	-0.010	-0.005
resWordv2	0.998	-0.062	0.005	-0.005	-0.014
Principle components for New variables					
numSubPart	0.996	0.052	-0.004	-0.030	-0.043
levelBelowSection	0.989	-0.125	0.074	-0.003	-0.026
hierarchyCount	0.991	-0.113	0.068	-0.005	-0.027
gerund	0.993	-0.109	0.030	-0.002	-0.003
shannonEntropy	0.983	-0.164	0.027	0.010	0.067
num_words_sent	0.921	0.388	0.028	-0.026	0.011
num_tokens_sent	0.924	0.380	0.024	-0.025	0.017
citInternal1	0.976	-0.203	0.051	0.002	0.015
citCFR	0.989	-0.077	-0.121	-0.030	0.000
citCFRInternal	0.987	-0.088	-0.125	-0.041	0.001
citCFRExternal	0.982	0.109	-0.049	0.148	-0.010