Public Choice (2019) 180:43–55
https://doi.org/10.1007/s11127-018-0600-y

RegData 2.2: a panel dataset on US federal regulations

Patrick A. McLaughlin[1] · **Oliver Sherouse**[1]

Received: 1 August 2018 / Accepted: 21 August 2018 / Published online: 6 September 2018

Abstract
How much regulation exists? Can short- and long-term growth trends in regulation be identified? Which agencies produce the most regulation? Are some sectors of the economy more regulated than others, and how big are the differences? RegData 2.2, a recent panel dataset from the RegData Project at George Mason University's Mercatus Center, offers answers to these questions and more. RegData 2.2 quantifies various aspects of US federal regulations by industry, by agency, and over time. The resulting datasets include metrics on volumes, restrictiveness, and relevance of federal regulations to different economic sectors and industries. RegData datasets are publicly released at http://quantgov.org. We explain the features of and methodology underlying RegData 2.2.

Keywords RegData · Regulation · Policy analytics · QuantGov · Machine learning

1 Introduction

Regulations are an important and widely used policy tool, but empirical analysis of regulations' actual effects has been hampered by a paucity of data (Al-Ubaydli and McLaughlin 2015). The RegData Project exists in part to fill that gap by creating and publicly releasing data that facilitate previously infeasible research. The RegData series of datasets from the RegData Project supply several decades of annual panel data on US federal regulations, including variables that measure regulatory quantities, origins, and applicability. Several other datasets, including preceding and subsequent versions of RegData as well as State RegData, which quantifies state-level regulations, and RegPulse, which quantifies features of pending, rather than existing, regulations, also are available at http://quantgov.org/data.

RegData is both a methodology and a database that quantifies regulations by industry, by regulatory agency, and over time. The RegData Project, originally launched in 2012, involves developing custom-made computer programs to perform text analytics and apply machine learning algorithms designed to quantify several features of regulation. The resulting statistics, which include metrics on the volume, restrictiveness, and relevance of federal regulations to different sectors and industries, are then released publicly at http://quantgov.

✉ Patrick A. McLaughlin
pmclaughlin@mercatus.gmu.edu

[1] Mercatus Center at George Mason University, Arlington, USA

org/data so that other researchers and interested parties can use the data in whatever ways they see fit.

While originally focused on the quantification of United States federal regulation, the RegData Project has expanded in scope to apply RegData methodology and corresponding technology to other jurisdictions' regulations as well as other policy-relevant documents besides regulations. To that end, we created an open-source, generalized version of the programs developed for RegData. This generalized approach, which is called QuantGov because the goal is to "quantify governance", permits the application of the same methodology to other bodies of policy-relevant text, including subnational regulatory jurisdictions (e.g., US states), other countries' regulations, and other types of policy-relevant documents, like executive orders, trade agreements, bills, or even transcriptions of policymakers' speeches.[1]

In this article, we explain RegData 2.2, one of the more recent datasets released in the RegData series, which spans 1975–2014 and covers hundreds of different industries. In Sect. 2, we summarize the primary features of and variables included in RegData 2.2. Section 3 supplies more details on the methodology followed to create the metrics of regulation included in RegData 2.2. Finally, we conclude in Sect. 4 by offering some examples of applications of RegData and briefly discussing the future of the RegData and QuantGov projects.

2 Features of RegData

Broadly speaking, regulation refers to a body of law known as administrative law. Administrative law comes about when a legislature delegates lawmaking authorities and obligations—often jointly referred to as statutory mandates—to one or more regulatory agencies, which then create and administer regulations to fulfill the mandate. In reference to the federal government of the United States, the term regulation refers to the administrative rules created by the executive-branch agencies, such as the Department of Transportation and the Department of Labor, as well as the rules promulgated by so-called independent regulatory agencies, such as the Securities and Exchange Commission and the Federal Trade Commission. These administrative rules, written under authority delegated by Congress and published in the *Code of Federal Regulations* (CFR), have the full force of law; failure to comply can result in fines or even incarceration.

Since its inception, the RegData Project has quantified key features of regulatory texts found in the CFR. The resulting datasets are given a version number (e.g., RegData 1, RegData 2.0, RegData 2.1) and released for public use. All RegData datasets capture the restrictiveness of regulations by counting words and phrases indicating a specific prohibited or required activity, called regulatory restrictions. RegData datasets also report the word counts, the originating agencies or departments, and an estimate of industry relevance for all regulations contained in the CFR. Each of these concepts is explained in more detail below.

RegData 2.2 and its successors, RegData 3.0 and 3.1, rely on machine learning algorithms to map federal regulations to the sectors and industries, as defined by the North

[1] For more on the QuantGov Project, visit http://quantgov.org.

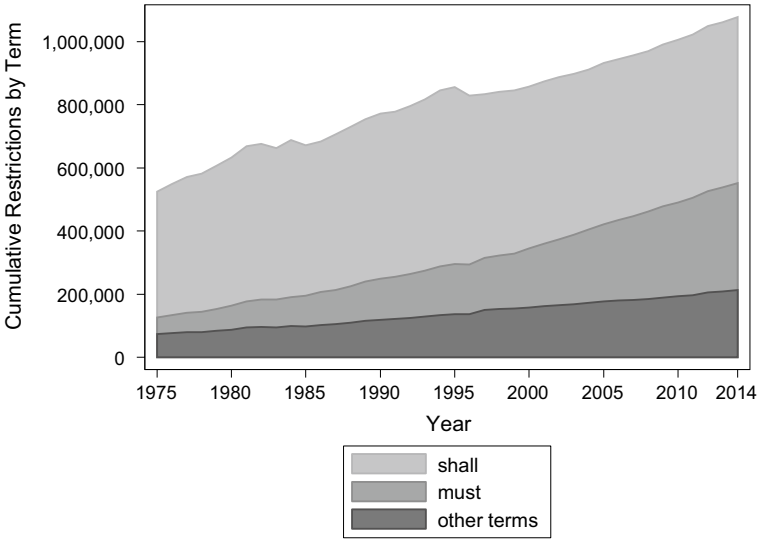

Fig. 1 Regulatory accumulation, 1975–2014

American Industry Classification System (NAICS), that are affected by those regulations.[2] NAICS classifications commonly are used in a wide variety of economic datasets, permitting users to merge RegData with other governmental information sources that may reflect the results or the causes of regulatory policies. In the United States, the Bureau of Economic Analysis and Bureau of Labor Statistics are just two examples of information sources that publish datasets organized according to the NAICS. Our intent was to facilitate research by designing around a commonly used system of industry classifications, at least for North America.

2.1 Primary metrics of regulation

The primary metrics in RegData 2.2 are *restrictions* and *industry relevance*. *Restrictions* is a cardinal proxy for the number of regulatory restrictions contained in regulatory texts, devised by counting specific words and phrases, such as "shall" or "must," that typically are used in legal language to create binding obligations or prohibitions. The database also includes a secondary measure of volume—the total *word counts*—as an alternative measure of the volume of regulations over time. Figure 1 shows the growth of *restrictions* from 1975 to 2014 by frequency of "shall," "must," and "other terms." The final category, "other terms," includes three terms: "may not," "prohibited," and "required."

As a way of measuring obligations and prohibitions, the RegData methodology of quantifying federal regulation contrasts starkly with some rougher proxies used in research preceding RegData's creation (see, for examples, Coglianese 2002; Mulligan and Shleifer

[2] Earlier versions of RegData also mapped regulations to NAICS-defined industries, but they used a human-assisted algorithm to achieve the mapping, rather than machine learning algorithms. The human-assisted algorithm used in the first two versions of RegData (1 and 2.0) is explained in great detail in Al-Ubaydli and McLaughlin (2015).

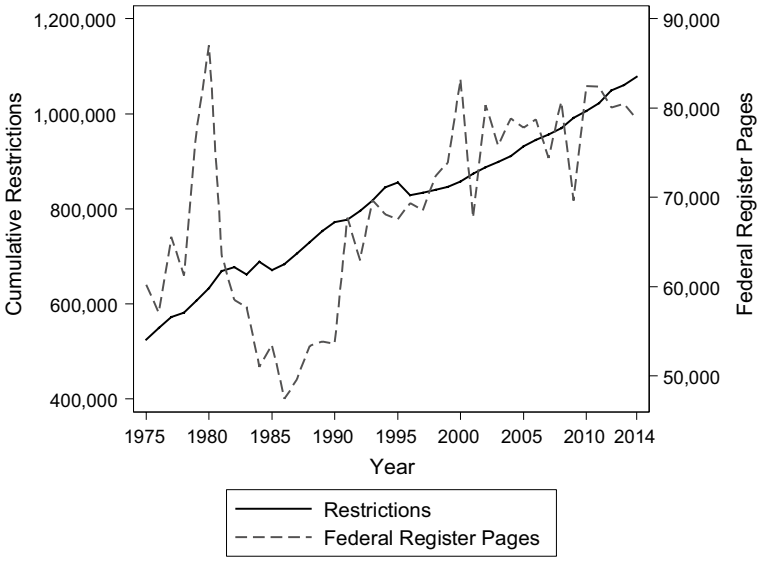

Fig. 2 Regulatory restrictions in the CFR and *Federal Register* pages, 1975–2014

2005; Coffey et al. 2012; Dawson and Seater 2013). Al-Ubaydli and McLaughlin (2015) offer a critique of several of those previously used metrics, such as counting pages in the *Federal Register*. They note, for example, that raw page counts from the *Federal Register* "may measure bureaucratic activity more than regulatory growth" because of the multitude of documents published there that do not add, and sometimes may even subtract, regulatory texts to the existing stock of regulations.

Figure 2 compares cumulative restrictions to *Federal Register* pages over time. Cumulative restrictions capture both the existing stock of regulations in each year and the annual change in that stock from new regulatory actions. In contrast, using *Federal Register* page counts, even as a measure of new regulatory activity, has several drawbacks. First, the *Federal Register* includes documentation only of changes to the stock of regulations, and says nothing about the existing stock. Second, a regulatory action published in the *Federal Register* historically is likely to add to the stock, but a significant number of those actions remove portions or the entirety of existing regulations—thus conflating a positive page count with a growth in regulation. Finally, many agency actions besides rulemaking activity are published in the *Federal Register*, making it a relatively noisy metric of regulatory activity.

While we direct the reader to Al-Ubaydli and McLaughlin (2015) for further details on alternative metrics of regulation and their drawbacks, we note one final point from that study. Previous attempts to quantify regulation, regardless of their relative merits and drawbacks, suffer from the major limitation of containing only longitudinal variation (i.e., time series) in total regulation for a given jurisdiction. By quantifying cross-sectional variation (i.e., variation across industries), RegData greatly enhances the possibility of better understanding the causes and effects of regulation. That goal is reached by estimating the relevance of regulatory texts to specific industries, as we explain below.

The second key variable in RegData is *industry relevance*, representing estimates of the relevance of a CFR entry to the different sectors and industries in the US economy.

RegData utilizes the industry definitions of the North American Industry Classification System (NAICS), which categorizes all economic activity into different industries. For example, in one version of NAICS (the two-digit version), the US economy is divided into approximately 20 industries, whereas the most granular version of NAICS (the six-digit version), divides the economy into more than 1000 industries. To illustrate, NAICS code 51 signifies the "Information" industry, while NAICS code 511191 signifies a much narrower segment of the information industry, "Greeting Card Publishers."

RegData 2.2 uses machine-learning algorithms to assess the probability that a unit of regulatory text targets a specific NAICS industry. That assessment requires two steps. First, the program "learns" what words, phrases, and other features can best identify when a unit of text is relevant to a specific industry by analyzing our compilation of training documents. Training documents are documents that are known to be relevant to one or more explicitly named industries—in the case of RegData, we gathered tens of thousands of training documents from publications in the *Federal Register* that name the NAICS codes affected by rulemakings. We explain the procedure and examine our machine learning algorithm's performance in some detail in Sect. 3 of this article.

2.2 The industry regulation index

With some simple calculations, the user can combine *restrictions* and *industry relevance* into a single variable to create an estimate of the number of restrictions that are relevant to a particular industry or set of industries in one or more parts of the CFR. The most commonly used combination of *restrictions* and *industry relevance* is the *industry regulation index* introduced by Al-Ubaydli and McLaughlin (2015). Although users easily can build different industry-specific regulatory indexes using different weightings or combinations of *restrictions* and *industry relevance*, Al-Ubaydli and McLaughlin's (2015) *industry regulation index* is available as a pre-constructed variable in RegData 2.2, which has been used in a broad range of studies.

This *industry regulation index* is designed to measure regulations relevant to industry i in CFR Part p in year y. Adopting the notation of Al-Ubaydli and McLaughlin (2015), industry relevant restrictions, r_{pyi}, is a function of the restrictiveness of a CFR Part, R_{py}, and the applicability of the Part, a_{pyi} (i.e., the probability that the restrictions apply to industry i):

$$r_{pyi} = f\left(a_{pyi}, R_{py}\right),$$

where the partial derivatives f_a, f_R, and their cross-partial are positive. Al-Ubaydli and McLaughlin (2015) operationalize the function simply by multiplying restrictions by probability,

$$f\left(a_{pyi}, R_{py}\right) = a_{pyi}R_{py},$$

and summing across all CFR Parts in each year,

$$r_{yi} = \sum_{p} r_{pyi}.$$

This construction has been applied in several contexts, including economic growth, industrial organization, law and economics, banking and finance, and public choice.

2.3 Summary of the core features of RegData 2.2

We provide a short summary of the core features of RegData 2.2 below.

For each Part-level segment of the CFR's annual editions from 1975 to 2014, RegData 2.2 provides:

- The CFR publication year, title number, and part number.
- A count of regulatory *restrictions*, denoted by the strings, "shall," "must," "may not," "required," and "prohibited," both individually and in total.
- A total *word count*.
- The authoring agency and department.
- *Industry relevance*, or the probability that the part is relevant to industries included in the 2007 North American Industry Classification System (NAICS) at the 2-, 3-, and 4-digit levels.[3] A CFR Part may be relevant to any number of industries, or relevant to no industry.
- The *industry regulation index* suggested by Al-Ubaydli and McLaughlin (2015).

The foregoing data elements can be or already are combined to produce annual series of:

- Regulatory restrictions by agency and department;
- Estimates of regulatory restrictions by industry at any NAICS level;
- Estimates of regulations on a NAICS industry by agency, by department or a combination thereof; and
- Total regulatory restrictions across all agencies for each year.

3 Methodology

3.1 Sources of regulatory text and the unit of analysis

All federal regulations are collected and published annually in the CFR. We thus gather and use the CFR as our primary source of regulatory texts. We draw from two sources for historical CFR texts. From 1996 to 2014, the Government Printing Office (GPO) makes the CFR available electronically in PDF, HTML, and XML formats. We use the HTML-formatted files for our analysis, because they are more complete and contain fewer errors than the XML or PDF versions.

For the years 1975–1996, we use text based on Optical Character Recognition (OCR) applied to scanned PDFs of the CFR. OCR is an inherently imperfect process, and the texts contain a variety of errors. The errors range from misidentification of certain letters to complete obliteration of certain portions of the text.

The CFR is divided into 50 topical titles, each published across one or more volumes. Titles are divided and subdivided, with varying levels of consistency, into chapters,

[3] If classifications for a given industry are not sufficiently reliable, that industry is included only in a supplemental, unfiltered dataset. For some industries, it is not possible to produce classifications at all because of the small number of example documents. See Sect. 3.3 below.

subchapters, parts, sections, subsections, paragraphs, and subparagraphs. The divisions and subdivisions generally correspond to levels of topical specificity.

While any of the divisions and subdivisions could serve as a unit of analysis, we analyze the CFR at the part level for several reasons. First, part-level divisions are 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 related set of issues that are likely to have similar relevance to industries throughout. Third, the CFR contains an index that attributes each CFR part to both authoring agencies and authorizing statutes—information we extract and use to create our datasets.

To identify individual CFR parts for analysis, we parsed each volume-level text to identify section-header lines. Section numbers contain the part to which the section belongs (e.g., Section 103.15 belongs to Part 103), and section text was appended to the appropriate part-level file. Parts are parsed from the raw text using regular expressions and extracted into individual files for analysis. Because both the OCR-processed documents and the HTML volumes occasionally suffer from missing data or difficult-to-parse formatting, we employ an error detection and smoothing algorithm to produce the final series.[4] That procedure affected less than 5% of all CFR parts analyzed.

3.2 Metrics of volume

The most basic information of interest about a CFR part is how much regulation it contains. A word count gives a rough approximation under the assumption that all CFR text is uniformly dense with regulation. A more precise measure of regulatory stringency, the regulatory restriction count, was introduced by Al-Ubaydli and McLaughlin (2015). Regulatory restrictions are words phrases that indicate a specific mandatory or forbidden activity: *shall*, *must*, *may not*, *prohibited* and *required*. While a regulatory restriction count has its own limitations, most notably the inability to distinguish between more and less costly restrictions, it improves both on the word count measure and on the traditional page count measure by allowing the regulatory density of CFR text to vary. Both total word counts and regulatory restriction counts are included in RegData for each CFR part.

Additionally, restriction and word counts are included in the dataset for the relatively small portions of text that cannot be assigned to any CFR part by the parsers. Those numbers are calculated by subtracting the number of words and restrictions in the parsed parts from the number in the unparsed source files. The same smoothing algorithm is used on this residual count as on the individual parts. The counts are especially useful when measuring the total number of words or restrictions in the CFR over time.

Another basic but readily available piece of information about a CFR part is the agency that authored it, and the department within which that agency resides. Such information is

[4] For RegData 2.2, our error detection and smoothing process proceeded in two main steps. First, if the section-level number could not be parsed because of OCR errors, a rolling-plurality vote for the previous 10 sections was used as the appropriate section. That approach was taken to localize errors to a single section rather than an entire part and to ensure that one-off errors were not carried forward to other sections.

Second, after the initial parsing, the file size for each part was analyzed for every year it was present in the CFR. Parts present in a single year, but not the year before or the year after were dropped. Parts missing in a single year, but present the previous year or the following year, were filled in using the text from the part in the preceding year. Because part size generally follows a smooth trend, we also corrected for outlier discontinuities. If a part's file size was not within 15% of either the previous or following year, the text from the previous year was used.

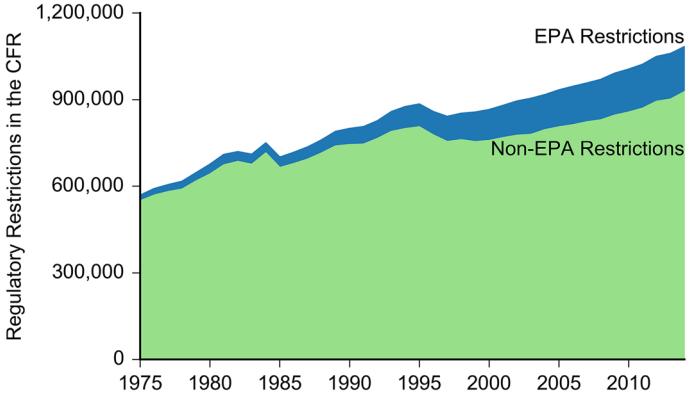

Fig. 3 EPA and non-EPA restrictions, 1975–2014

given in the "Index and Finding Aids" of the CFR at the part level starting in 1975. Agency and department names, along with unique IDs, also are included in RegData.

This basic dataset by itself reveals interesting trends over the past 40 years of regulation. The total number of restrictions, shown in Fig. 1, has increased steadily over the entire period, except for slowdowns in the mid-1980s and late 1990s. The total number of restrictions passed one million in 2012, and for 2014 the total stands at 1.04 million.[5]

Looking into the agency data, one particular agency, the Environmental Protection Agency (EPA), dominates the growth trend since 1975. Over that period, EPA-authored CFR parts have comprised 26.5% of the total growth in regulations, as shown in Fig. 3. While the non-EPA parts of the CFR have grown at a compound annual restriction rate of 1.4% from 1975 to 2014, the EPA-authored parts have grown at a compound annual rate of 5.5%.

3.3 Classification of regulations by industry

A more difficult problem than quantifying the amount of regulation is identifying the applicability of regulations. In producing RegData 2.0, Al-Ubaydli and McLaughlin (2015) classified CFR parts according to industry by counting the frequency of a specified set of search terms for each industry. RegData 2.2 improves on that approach by incorporating machine learning.

We classify industries according to the North American Industry Classification System (NAICS). NAICS classifies businesses and other organizations by their production methods. The system identifies a set of codes at increasingly detailed levels of differentiation that are intended to be mutually exclusive and collectively exhaustive. Table 1 lists NAICS's 11 2-digit industries with sub-industries at the 3-digit and 4-digit levels, along with their descriptions.

NAICS data for training purposes were obtained from the *Federal Register*, a daily publication of the federal government which includes rules, proposed rules, presidential

[5] Subsequent versions of RegData have added additional years of coverage. RegData 3.0 spans 1970–2016, while 3.1 covers 1970–2017. These datasets also are available at http://quantgov.org/data.

Table 1 Sample NAICS hierarchy

NAICS code	Industry name
11	Agriculture, forestry, fishing and hunting
111	Crop production
1111	Oilseed and grain farming
1112	Vegetable and melon farming
1113	Fruit and tree nut farming
1114	Greenhouse, nursery, and floriculture production
1119	Other crop farming
112	Animal production
1121	Cattle ranching and farming
1122	Hog and pig farming
1123	Poultry and egg production
1124	Sheep and goat farming
1125	Aquaculture
1129	Other animal production
113	Forestry and logging
1131	Timber tract operations
1132	Forest nurseries and gathering of forest products
1133	Logging
114	Fishing, hunting and trapping
1141	Fishing
1142	Hunting and trapping
115	Support activities for agriculture and forestry
1151	Support activities for crop production
1152	Support activities for animal production
1153	Support activities for forestry

documents, and a variety of notices of current or planned government activity. Some of those documents are specifically labeled with relevant NAICS codes, and the language they employ is similar to that of the CFR.

Training documents for each NAICS industry were obtained by searching the *Federal Register* using its official Application Procedure Interface (API). Taking industry 111 as an illustrative example, we searched first for exact, case-insensitive matches of "NAICS 111", "NAICS code 111", and "NAICS category 111". In addition, we searched for all documents containing both the exact name of the industry and the industry code. Those searches were repeated for all sub-industries in the NAICS hierarchy. Finally, for sub-industries at the six-digit level, we searched for the NAICS code alone. Because our goal was to identify industry-specific documents, we excluded trainers that were associated with more than six industries. Only industries with more than five industry-specific positive trainers were considered trainable, because five-fold cross-validation was used for per-industry evaluation.

Table 2 Best performing classifiers

Classifier	Parameters	Weighted average label F1	SD
Regularized logistic regression	C = 1000	0.8808	0.010
K-neighbors	k = 1	0.8581	0.0097
Random forests	n = 251	0.7540	0.0088

We used the scikit-learn package (Pedregosa et al. 2011) to create and evaluate our classification model.[6] We lemmatized the trainers using the WordNet lemmatizer and vectorized as bags of 1 and 2-grams, using the 10,000 most common n-grams.[7] The vector was transformed using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm.

We evaluated three different types of classification models using five-fold cross-validation: regularized logistic regression (logit) with l2-penalty, nearest neighbors, and random forests. Those models were selected because they could avoid overfitting and produce probability scores. For the logit and nearest neighbors, we used a one-versus-rest strategy for multilabel classification (random forests classifiers are inherently multilabel).

The output of each of the classification models is a probability that a unit of analysis (i.e., a CFR part) is relevant to a NAICS-defined industry. Evaluation results for the three-digit NAICS classification are laid out in Table 2, using weighted average F1 scores as our primary evaluation metric. F1 scores balance recall and precision in a combined score. Recall is the percentage of true positive documents that are classified as positive in cross-validation (i.e., out-of-sample prediction). In our case, positive classification means assigning to a CFR part a probability of 0.5 or greater for a specific NAICS-defined industry. Precision measures resistance to false positives and is calculated as the percentage of positively classified documents that are true positives. In other words, a good recall score indicates that the classifier does a good job of detecting when a document belongs to a specific class. But because a classifier could report that all documents belong to all classes and receive a perfect recall score, we want to also know how well the classifier does at avoiding false positives—which is what the precision score indicates. An F1 score balances the two and is primarily useful for comparing models on both the precision and recall dimensions simultaneously. Formally, an F1 score is the geometric mean of recall and precision.[8]

[6] Scikit-learn is an open source set of machine learning tools and algorithms for the programming language, Python, available at: http://scikit-learn.org/stable/.

[7] Lemmatization refers to an algorithmic process common to computational linguistics where a computer program identifies a word's "lemma," or dictionary form. For example, the word "environment" is the lemma for the adjective, "environmental." Lemmatization lets occurrences of different inflected forms of the same lemma (such as "environmental" in the example above) be analyzed as a single category or item. WordNet is open source Python script that performs lemmatization and is available as part of the Natural Language ToolKit (NLTK) package at https://www.nltk.org/install.html.

[8] Precision is calculated as TP/(TP + FP), where TP is *true positives* and FP is *false positives*. Recall is calculated as TP/(TP + FN), where FN is *false negatives*. In both cases, the highest possible score for a model along the single dimension equals one. The F1 score, therefore, also has a maximum possible score of one, but that is not necessarily desirable. There is usually a tradeoff between the two dimensions. A model can have very high precision *because* it creates many false negatives. F1 scores are useful comparing models for a given classification project while balancing between those two dimensions. However, the machine learning community typically cautions against comparing one project to another by using F1 scores because precision or recall may be valued in different ways in different projects.

Table 3 Top ten most regulated industries for 2014, 3-digit NAICS

Industry name	NAICS code	Estimated regulatory restrictions
Chemical manufacturing	325	70,643
Utilities	221	58,020
Professional, scientific, and technical services	541	50,786
Transportation equipment manufacturing	336	34,812
Support activities for transportation	488	29,396
Securities, commodity contracts, and other financial Investments and related activities	523	29,155
Petroleum and coal products manufacturing	324	27,349
Food manufacturing	311	27,290
Credit intermediation and related activities	522	26,290
Air transportation	481	18,482

The highest-performing model was the regularized logit model. The one-versus-rest strategy trains one classifier per industry, and not all industry classifiers are equally strong. To guide users of the dataset, we have produced a filtered dataset that only contains industries with classifiers that surpass a minimum performance threshold. We determined a minimum performance threshold by evaluating each industry classifier in terms of the area under the receiver operating characteristic curve (often referred to as the ROC AUC score—for Receiver Operating Characteristic curve's Area Under the Curve), a method of assessing the predictive accuracy of machine learning algorithms (Huang and Ling 2005; Fawcett 2006). A ROC AUC score measures the degree to which true positives generally have a higher predicted probability than true negatives; it is calculated as the area under a curve plotting the false positive rate (percentage of true negative documents classified positive) against the recall at every possible probability threshold from 0 to 1. For an industry to be included in the standard RegData 2.2 dataset, its classifier had to achieve a ROC AUC score of 0.75 or greater. However, all industry classifications are available in the unfiltered dataset, along with per-industry metrics, so that users can make their own choices.

Industry classification results are presented in RegData 2.2 as the probability that a CFR part for a given year belongs to a given industry. Those probabilities can be used with the basic statistics above to produce industry-specific estimates of regulation—one way to do so is explained in Sect. 2 of this article. Of course, because all of our classification data and other metrics are publicly available, plenty of other options are available to the researcher. For example, simply summing the probabilities across CFR parts for a given year yields an estimate of the number of parts that are relevant to a given industry.

The top-ten most regulated industries for 2014, using the aforementioned *industry regulation index* (Al-Ubaydli and McLaughlin 2015) as the metric of industry regulation, are reported in Table 3, along with the values of the index, which can be interpreted as the estimated number of regulatory restrictions applicable to the industry in that year.

3.4 The public law database

The *Index and Finding Aids* included in the CFR lists both the part-level authoring agency and statutory authority for the rules it contains, and they also are included in the dataset

as the public law database. While agency attribution is unique to each year-title-part and therefore relatively straightforward, attribution to authorizing laws is more complex. Agencies may list a proposed bill at any point in the three-stage legislative publishing process: by public law number, citation in the US Statutes at Large, or by the law's incorporated location in the US Code (USC). While public law numbers and Statutes at Large citations are unique and definitive, citations of the USC, which comprise the vast majority of citations, can be ambiguous.

This ambiguity arises because multiple laws can affect one cited section of the US Code. The relationship between individual laws and citations is given in Table III of the USC at the section level. The CFR also cites the USC at the section level; however, an individual law may alter less than an entire USC section. Lacking any method to determine a definitive relationship, we have included the associations for every law that affects a section of the US Code that a given CFR year-title-part lists as an authority. Agencies also may identify more than one law as an authority, either to ground an action in multiple legal theories or because multiple actions require justification. Here again, we do not attempt to resolve the ambiguity, but rather include all associations in the RegData dataset.

4 Concluding remarks

The RegData Project was designed to capture novel regulatory metrics that would advance our understanding of the causes and effects of regulation and the regulatory process in ways that previously were infeasible or impossible. Since then, several journal articles and even more working papers have featured RegData as a critical component of their analyses.[9] Because several of the studies in this special issue use RegData, this article describes RegData 2.2 and the RegData Project.

RegData 2.2 is a panel dataset containing annual observations of US federal regulation applicable to 2-, 3-, and 4-digit North American Industrial Classification System (NAICS)-defined industries from 1975 to 2014. More specifically, the dataset offers a way of quantifying the restrictiveness of regulations by counting words and phrases that indicate a specific prohibited or required activity. The words and phrases are called regulatory restrictions. RegData also reports the word count, agency, department, and an estimate of industry relevance for all regulations contained in the *Code of Federal Regulations* (CFR). In this article, we described those variables from RegData 2.2 and expounded on the methodology used to create them.

RegData 2.2, like all data released from the RegData Project, is publicly available at http://quantgov.org/data. Several other datasets also are available on that page, and the reader may be wondering why data from the RegData project are made available at a website called QuantGov. In the process of creating and improving RegData, it became clear that the methods applied in that effort could usefully be replicated in other contexts. Other interesting bodies of law are available for study, and other interesting ways to analyze text using the advances in machine learning now have been developed. Those advances led to the creation of QuantGov, which now is the umbrella project under which RegData resides.

[9] Several of the articles in this special issue use RegData 2.2, including Bailey et al. (2018), Chambers et al. (2018a, b, c), Manish and O'Reilly (2018) and Mulholland (2018). Here is a short and by no means comprehensive list of other journal articles: Ellig and McLaughlin (2016), Bailey and Thomas (2017), Goldschlag and Tabarrok (2018) and Pizzola (2018). A more comprehensive list, including dozens of working papers, is available at: http://quantgov.org/research.

QuantGov—shorthand for "quantified governance"—represents an ongoing effort to measure and summarize the outputs of governance. In addition to the federal regulations found in the CFR, QuantGov includes other documents related to the regulatory process, such as the *Federal Register* or guidance documents, and also permits the eventual expansion of the dataset to government documents like as executive orders, bills, and legislation. QuantGov also carries RegData's methodology into other jurisdictions besides the US federal government, including US states, other countries, and other countries' states and provinces. One principal goal of the QuantGov and RegData Projects is to create a database that permits analysis of regulations and other policies not just across industries and over time, but also across governance regimes.

References

Al-Ubaydli, O., & McLaughlin, P. A. (2015). RegData: A numerical database on industry- specific regulations for all United States industries and federal regulations, 1997–2012. *Regulation & Governance, 11*(1), 109–123.

Bailey, J. B., & Thomas, D. W. (2017). Regulating away competition: The effect of regulation on entrepreneurship and employment. *Journal of Regulatory Economics, 52*(3), 237–254.

Bailey, J. B., Thomas, D. W., & Anderson, J. R. (2018). Regressive effects of regulation on wages. *Public Choice*. https://doi.org/10.1007/s11127-018-0517-5.

Chambers, D., Collins, C. A., & Krause, A. (2018a). *How do federal regulations affect consumer prices?*. An analysis of the regressive effects of regulation: Public Choice. https://doi.org/10.1007/s11127-017-0479-z.

Chambers, D., McLaughlin, P. A., & Stanley, L. (2018b). Barriers to prosperity: The harmful impact of entry regulations on income inequality. *Public Choice*. https://doi.org/10.1007/s11127-018-0498-4.

Chambers, D., McLaughlin, P. A., & Stanley, L. (2018c). Regulation and poverty. Public Choice, this issue.

Coffey, B., McLaughlin, P. A., & Tollison, R. D. (2012). Regulators and redskins. *Public Choice, 153,* 191–204.

Coglianese, C. (2002). Empirical analysis and administrative law. *University of Illinois Law Review, 4,* 1111–1138.

Dawson, J. W., & Seater, J. J. (2013). Federal regulation and aggregate economic growth. *Journal of Economic Growth, 18*(2), 137–177.

Ellig, J., & McLaughlin, P. A. (2016). The regulatory determinants of railroad safety. *Review of Industrial Organization, 49*(2), 371–398.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27–8,* 861–874.

Goldschlag, N., & Tabarrok, A. (2018). Is regulation to blame for the decline in American entrepreneurship? *Economic Policy, 33*(93), 5–44.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracty in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering, 17–3,* 299–310.

Manish, G. P., & O'Reilly, C. (2018). Banking regulation, regulatory capture, and inequality. *Public Choice*. https://doi.org/10.1007/s11127-018-0501-0.

Mulholland, S. E. (2018). Stratification by regulation. *Public Choice*, this issue. https://doi.org/10.1007/s11127-018-0597-2.

Mulligan, C., & Shleifer, A. (2005). The extent of the market and the supply of regulation. *Quarterly Journal of Economics, 120,* 1445–1473.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830.

Pizzola, B. (2018). Business regulation and business investment: Evidence from US manufacturing 1970–2009. *Journal of Regulatory Economics, 53*(3), 243–255.