RegData 3.2 User's Guide

Patrick McLaughlin, Kofi Ampaabeng, Ethan Greist, Jonathan Nelson, Thurston Powers, and Stephen Strosko

March 23, 2020

1 Introduction

RegData is both a methodology and a database focused on the quantification of various dimensions of regulation. We use custom-made text analysis and machine-learning algorithms to create statistics designed to measure several features of regulation, including volume, restrictiveness, complexity, and relevance to different sectors and industries. The RegData Project was launched in 2012 with the express purpose of facilitating research that was previously infeasible. Regulations have been an important and widely used policy tool for decades, but empirical analysis of the actual effects of regulation has historically been hampered by a paucity of data.

RegData was designed to solve this data problem. RegData 3.2 maps United States federal regulations to the sectors and industries affected by them. This opens the way for new research about the impact of regulation. RegData captures the restrictiveness of various regulations by counting words and phrases that indicate a specific prohibited or required activity; these words and phrases are called regulatory restrictions. RegData also reports the word count, complexity, agency, department, and an estimate of industry relevance for all regulations contained in the Code of Federal Regulations.

RegData is produced with the open-source QuantGov policy analytics platform. Because of its open-source nature, QuantGov can be used to produce modified versions of the RegData dataset or datasets based on other documents, such as guidance documents or regulations from other jurisdictions. The technical details on using the QuantGov platform are available on the qunatgov library docs page. Data from other jurisdictions (including national and subnational jurisdictions in the United States, Canada, and Australia) are also available on the download data page.

2 New Features Included in Version 3.2

RegData 3.2 introduces several new features. These include new metrics, additional years of data, and improved accuracy. The new metrics are complexity

MERCATUS CENTER George Mason University

QuantGov

and the last updated date of the regulations. These new features are described below.

2.1 Complexity Metrics

RegData 3.2 features a new suite of metrics that measure the complexity of a piece of regulatory text. Each year, title, part combination is associated with three different measures of complexity: Shannon Entropy, Sentence Length, and Conditional Counts. Whether examined individually or combined to create a custom metric, these three measures of complexity can be used to surface regulations that may be difficult to read or comprehend from an objective, statistical standpoint. Complexity can also be associated with compliance costs: regulations that are difficult to comprehend are liable to demand additional time for regulated entities to achieve complete compliance. These metrics can also be combined with other RegData metrics including restriction counts by industry and agency. It is recommended that complexity metrics not be summed or averaged; they are best used as a relative comparison at the part level.

Shannon Entropy: This metric measures the likelihood of encountering new words and concepts in a given body of text. The average Shannon Entropy score at the part level in the Code of Federal Regulations (CFR) for 2019 is 7.86. Removing outliers above or below two standard deviations, the average increases to 8.24. For the sake of comparison, compositions by Shakespeare tend to have an entropy score between 9.3 and 9.7.

Conditional Counts: This metric counts the number of branching words such as "if", "but", and "provided" found in any given part in the CFR. These words identify separate logical branches in a body of text. From a computer programming standpoint, this is similar to the concept of cyclomatic complexity. The average conditional count for a part in the CFR for 2019 is 83.35 conditional words.

Sentence Length: This metric is relatively straightforward and measures the mean sentence length across any given part in the CFR. The average sentence length in the CFR for 2019 is 21.62 words. While this metric is certainly useful, it should be noted that regulatory text is not always conducive to analysis by sentence length. Some regulations include tables, long lists, and appendices that may skew the effectiveness of a sentence length metric. Therefore, it is recommended to use this metric as a surfacing tool or in combination with other complexity metrics.

2.2 Last Updated Metric

Last Updated: This metric attempts to capture the "staleness" of a regulation and can be used to surface regulations that have not been updated for a long period of time, or are perhaps being updated too often. The metric is calculated

by looking for year to year variations in word count at the part level of the CFR. A part's staleness date is the last time that the individual part saw a 1% change in word count from the previous year. This change is calculated in both the positive and negative directions.

2.3 New Years of Data

RegData 3.2 expends the date range of RegData, including analysis through the 2019 version of the CFR. The two additional years, 2018 and 2019, both use text made available through the eCFR (the electronic version of the CFR) which is published daily. Later sections in this User Guide will go into the specifics of how this text is chosen and cleaned for RegData 3.2.

2.4 Improved Accuracy

With the release of RegData 3.2, we have taken some additional steps in improving the accuracy of the collected text and of the resulting data. Version 3.2 now accurately removes the majority of title, part, and chapter tables of contents from collected CFR text. Specifically, this improvement targeted the years 1997 through 2019, due to the relatively higher consistency of the text for that year range. While this is a level change across all title/parts and most years, it does marginally improve the performance accuracy of the machine learning algorithms that undergird the industry data associated with RegData 3.2.

3 Primary Features of RegData

3.1 Unit of Analysis

The CFR is divided into fifty topical titles, each published across one or more volumes. Titles are subdivided, with varying levels of consistency, into chapters, subchapters, parts, sections, subsections, paragraphs, and subparagraphs. These subdivisions generally correspond to levels of topical specificity. While the CFR is divided and subdivided into several different demarcated portions—such as title, chapter, part, subpart, section, and paragraph—all downloadable RegData 3.2 datasets use the CFR part as the unit of analysis. We analyze the CFR at the part level for several reasons. 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.

3.2 Primary Metrics of Regulation

While RegData 3.2 contains many metrics, the two primary metrics continue to be restrictions and industry relevance – which can be combined to create the industry restrictions metric.

Restrictions is a cardinal proxy of the number of regulatory restrictions contained in regulatory text, devised by counting select words and phrases that are typically used in legal language to create binding obligations or prohibitions. The current words used by RegData 3.2 are the same words used in all previous RegData versions: shall, must, may not, required, and prohibited. The database also includes a secondary measure of volume—the total word counts—as an alternative measure of the volume of regulation over time. Industry relevance is the second key variable in RegData, representing estimates of the relevance of a CFR part to the different sectors and industries in the economy.

RegData utilizes the industry definitions in 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 in the most granular version of NAICS (the six-digit version), the economy is divided into over 1,000 industries. To illustrate, NAICS code 51 signifies the "Information" industry, while NAICS code 511191 signifies a much more granular subsector of the information industry, the "Greeting Card Publishers" industry. RegData uses NAICS because it allows researchers to easily merge RegData with other datasets that contain information about the United States economy. In the United States, the Bureau of Economic Analysis and Bureau of Labor Statistics are just two examples of organizations that publish several datasets designed around NAICS.

To create the estimates of the relevance of a CFR part to a specific industry, RegData 3.2 uses custom trained machine-learning algorithms. These algorithms "learn" 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. Over the course of the RegData project, we have gathered tens of thousands of training documents from publications in the Federal Register that name the NAICS codes affected by rulemakings. Combining the probabilistic output of industry relevance metric and the restrictions metric allows the researcher to create the industry restrictions metric. This metric is an estimate of the number of restrictions that are relevant to a particular industry or set of industries in one or more CFR parts—see Al-Ubaydli and McLaughlin (2015) for a discussion and examples. The advent of an industry-specific metric of regulation that is comprehensive (i.e., inclusive of all federal regulations that are in effect in each year), replicable, and transparent has created paths to performing economic research on regulation in ways that were previously infeasible.

3.3 Summary of Data

RegData 3.2 provides the following for each part-level segment of the CFR dating from 1970 to 2019:

- The CFR publication year, title number, and part number.

- A count of regulatory restrictions, denoted by the individual phrases counted: shall, must, may not, required, and prohibited.
- A count of words.
- The authoring agency and department.
- A probability that the part is relevant to industries included in the 2007 NAICS at the 2, 3, 4, 5, and 6-digit levels.
- The three complexity metrics that are described in the new features section of this user guide.
- The last date the text was updated (only for the year 2019).

These data elements can be combined to produce the following metrics that have been commonly used by other researchers:

- Regulatory restrictions by agency and department.
- Estimates of regulatory restrictions by industry at any NAICS level.
- Estimates of regulatory restrictions by industry for each agency and department.
- Relative complexity of regulations for each industry, agency, and department.

4 Methodology

4.1 Sources of Regulatory Text

We use three sources of data to construct a plain-text historical series of the Code of Federal Regulations. For the years 1970–1995, we use scans of book pages that have been processed with Optical Character Recognition (OCR) using the Tesseract-OCR engine. For 1996–2016, we use the HTML versions of the historical CFR made available by the Government Publishing Office (GPO). While GPO does make an XML version of the historical CFR available, this version is produced automatically from the language used to typeset the CFR and is not reliable for our purposes. Finally, the XML version of the Electronic CFR, a current version of the CFR produced by the Office of the Federal Register and GPO, is annualized and used for 2017-2019.

The CFR is divided into titles, which are subdivided into chapters, subchapters, parts, and so forth. We analyze at the part level for several reasons. Parts are present in every CFR title, unlike some division types, and are generally concerned with only one relatively specific topic, without being so specific as to lose important context. Moreover, it is at the part level that authoring agencies and authorizing legislation are identified in the CFR indices. Parts are parsed out of 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. This procedure affected less than 5 percent of all CFR parts analyzed.

4.2 Source of Industry Relevance Estimates

For each unit of analysis (i.e., for each CFR part), we estimate the probability of relevance to industries as defined by the 2007 NAICS. NAICS specifies a mutually exclusive and collectively exhaustive set of industry definitions at levels of specificity, with 2-digit codes corresponding to the most general industries, and 6-digit codes corresponding to the most specific. More specific industries are subdivisions of more general ones, so that industries have exactly one parent and one or more children.

Our training data comes from the XML version of the historical Federal Register. In some proposed and final rules, agencies use the NAICS codes and descriptions to identify the industries to which their rules are expected to apply. We searched all 106,966 proposed and final rules published in the Federal Register from 2000 to 2017 for exact matches of the full NAICS industry name, the name of a parent industry, or the name of a child industry as indicators of direct relevance to an industry. Matches must be non-overlapping, with the longest match taking precedence; for example, an occurrence of the term "Basic Chemical Manufacturing" will match for the industry of that name, but not for its parent industry named simply "Chemical Manufacturing."

While most industry names are only meaningful in the context of the industry, a few (such as industry 51, "Information") have more generally used names and are therefore blacklisted as matches, although child and parent industries of these are still used. Two agencies—the Small Business Administration and the Office of Personnel Management—both frequently publish rules about the formal definition of NAICS industries that are not actually restrictions on those industries. Rules from these two agencies are therefore excluded. Additionally, any rules that matched more than 2.5 standard deviations above the mean number of matched documents were assumed not to be industry-specific and were dropped.

The full final training set for each NAICS-defined industry consists of all rules that are labeled positive for at least one industry, provided that there are also a minimum five individual documents that are positively labeled for that industry. The training documents were vectorized using bigram counts. Words for vectorization were defined as sequences of two or more alphabetic characters, with English stop words excluded. The vocabulary was limited to bigrams occurring in at least 0.5 percent but not more than 50 percent of trainers.

Because a CFR part can and often will be relevant to more than one industry, we use a multilabel approach for classification. We tested one parametric and one nonparametric model: Logistic regression (Logit) and Random Forests, respectively. In our Logit model, we used a lasso penalty and implemented multilabeling using a one-vs.-rest strategy. In both models, we employed a Term

Frequency–Inverse Document Frequency preprocessor to normalize document length and, in the case of the Logit model, to normalize coefficients for the purposes of calculating the penalty. The models were tuned and compared using fivefold cross-validation using the average F1 score across all classes. For each level of NAICS, the Logit model was the superior classifier. The smallest regularization parameter that was within one standard deviation of the top score was selected for training the model on the full training set for each level.

For the model selected using cross-validation, we additionally estimate a variety of performance metrics for each class individually. Table 1 explains these performance metrics and their definitions. Since these performance metrics are primarily useful for comparing models, we produced a normalized score that represents the percentage of possible improvement over the baseline actually seen in the trained classifier. Because the training documents are overwhelmingly true negatives, the baseline classifier for accuracy is an all-negative classifier. For all other metrics, the baseline classifier is one that randomly classifies as positive or negative.

Table 1: Metrics of Performance

Metric	Description Definition
F1	Balances recall and precision in a combined score. Geometric mean of precision and recall.
Precision	Measures resistance to false positives. Percentage of positive-classified documents that are true positives.
Recall	Measures detection ability. Percentage of true positive documents that are classified positive.
Accuracy	Measures exact correctness but is subject to inflated scores when most observations are false. Percentage of documents correctly classified.
Receiver Operator Characteristic (probability)	Measures that true positives generally have a higher predicted probability than true negatives. Area under a curve plotting the false positive rate (percentage of true negative documents classified positive) against the recall at every probability threshold from 0 to 1.
Receiver Operator Characteristic (binary)	Balances recall against resistance to false positives True positives rate (recall) multiplied by one minus the false positive rate.

If classifications for a given industry do not exceed a minimum performance threshold (based on F1 scores), that industry is excluded from the dataset. The minimum performance threshold uses a conservative value for the normalized F1 score calculated by subtracting one standard deviation from the mean score obtained in the train-test splits. This conservative normalized F1 must be higher than 0.5 to pass the filter, yielding confidence that the classification for that

industry is at least halfway between random and perfect. Those industries for which the F1 is below the minimum performance threshold are made available in a separate, and clearly marked, unfiltered dataset for researchers wishing to use their own threshold. Median normalized score results for each NAICS level are presented in table 2, while median non-normalized score results are presented in table 3.

Table 2: Normalized Median Scores for Filtered Industries

NAICS Digit	Industries	F1	Precision	Recall	Accuracy	ROC-AUC	ROC-AUC (binary)
2	15	0.729	0.829	0.369	0.534	0.946	0.679
3	54	0.739	0.833	0.355	0.534	0.951	0.676
4	127	0.800	0.900	0.508	0.641	0.973	0.753
5	289	0.739	0.862	0.333	0.556	0.955	0.667
6	490	0.782	0.887	0.467	0.607	0.969	0.732

Table 3: Non-Normalized Median Scores for Filtered Industries

NAICS Digit	Industries	F1	Precision	Recall	Accuracy	ROC-AUC	ROC-AUC (binary)
2	15	0.746	0.831	0.684	0.992	0.973	0.840
3	54	0.742	0.837	0.677	0.996	0.976	0.838
4	127	0.804	0.901	0.754	0.997	0.987	0.877
5	289	0.742	0.864	0.667	0.998	0.977	0.833
6	490	0.784	0.890	0.733	0.997	0.984	—0.866

4.3 Obtaining Data

With the new QuantGov.org website, users can now customize data downloads by using the API or by using the interactive downloader. This means that an individual can specifically select to download the metadata for a subset of industries or agencies. In addition, the entire document-level metadata can be downloaded from 1970 to 2019. Below are the links to the current documentation for the main ways to download RegData 3.2.

- Interactive Downoald
- Main API Documentation
- Python API Documentation
- R API Documentation

Table 4 covers all of the different variables that are available in RegData 3.2 bulk dowload. For more information about these variables, how to obtain data, or RegData 3.2, please email quantgov.info@gmail.com
Table 5 covers all of the different variables that are available for RegData 3.2 through the API. For more information about these variables, how to obtain data, or RegData 3.2, please email quantgov.info@gmail.com

Table 4: Variable Descriptions

Metric	Description Definition
year	Year that the part was published in. Ranging from 1970 to 2019.
title	Title containing the CFR part. Ranging from 1 to 50.
part	Part number within the title, range caries by title.
agency	Name of the agency as given in the CFR Index.
agency-id	Budget code for the agency.
department	Name of the department associated with the agency.
department-id	Budget code for the department.
words	Total number of words.
shall	Occurrences of the word shall.
must	Occurrences of the word must.
may not	Occurrences of the words may not.
required	Occurrences of the word required.
prohibited	Occurrences of the word prohibited.
restrictions	Total number of the five restrictive terms.
naics-code	The NAICS code as defined in the 2007 NAICS.
industry-relevance	Probability that the CFR part is relevant to the industry.
industry-relevan-restrictions	Number of restrictions relevant to a specific industry. Calculated by multiplying restrictions by industry-relevance.
shannon entropy	Shannon entropy score for the given part.
conditionals	Number of conditionals for the given part.
sentence length	Average sentence length for the given part.
last-updated	Last date that the given part's wordcount changed by at least 1%.

Table 5: Variable Descriptions

Metric	Description Definition
seriesID	The ID of the series (or type of data) that is being returned.
seriesValue	The data value that the user requested through the API (restrictions, words, complexity, etc.) for that row.
documentReference	Information about how to identify the specific document.
documentSubtypeID	The type of document. For 3.2 this will always be 3, or regulations.
subtypeName	Official name of the document subtype.
documentTitle	Title of the document.
country	Country that the documents are from. For 3.2 this will always be the United States.
jurisdictionID	Jurisdiction that the documents are from. For 3.2 this will always be 38, or the United States.
endDate	The end date for the line of data.
startDate	The starte date for the line of data.
periodCode	
agencyID	The identifying number of the agency for the data contained on that row.
industryCode	The industry code for the data contained on that row.