MERCATUS CENTER
George Mason University

HOW-TO MANUAL

RegData 3.0 User's Guide

September 2017

1. INTRODUCTION

RegData is both a methodology and a database that quantifies United States federal regulations by industry, by regulatory agency, and over time. We use custom-made text analysis and machine-learning algorithms to create statistics designed to quantify several features of regulation, including volume, restrictiveness, 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 regulations' actual effects was historically hampered by a paucity of data. The federal government requires some regulatory agencies to perform regulatory impact analyses of significant new regulations. In theory, these analysts estimate the marginal impacts—positive or negative—that the new rule would precipitate. However, these regulatory impact analyses are performed only for a very small portion of new regulations, making it virtually impossible to use data from impact analyses to comprehensively assess the effects of regulation or the regulatory process. RegData was designed to solve this data problem.

RegData 3.0 maps federal regulations to the sectors and industries affected by them. North American Industry Classification System (NAICS) classifications are commonly used in a wide variety of economic data, permitting users to merge RegData with other datasets that may reflect the results of regulatory policies. In the United States, the Bureau of Economic Analysis and Bureau of Labor Statistics are just two examples of data sources that publish several datasets designed around NAICS. Our intent was to facilitate research by designing around a commonly used system of industry classification, at least in North America. Research that uses the database (that we are aware of) is listed on the research page of the RegData website.

RegData 3.0 is the latest iteration of the RegData series. 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**, **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. The technical details on using the QuantGov platform are available at http://docs.quantgov.org/.

2. PRIMARY FEATURES OF REGDATA

RegData quantifies key features of regulatory text found in the *Code of Federal Regulations* (CFR)—the volumes of books containing regulations in effect in each year.

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.

The unit of analysis is the regulatory text contained in a formally demarcated portion of regulatory text. While the CFR is divided and subdivided into several different demarcated portions—such as title, chapter, part, subpart, section, and paragraph—all downloadable RegData 3.0 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, the CFR indices that attribute regulatory text to their authoring agencies and authorizing statutes both do so at the part level.

Primary Metrics of Regulation

The two primary metrics in the RegData database are *restrictions* and *industry relevance*. *Restrictions* is a cardinal proxy of the number of regulatory restrictions contained in regulatory text, devised by counting select words and phrases, such as *shall* or *must*, that are typically 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 regulation over time.

The second key variable in RegData is *industry relevance*, 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 3.0 uses machine-learning algorithms to assess the probability that a unit of regulatory text targets a specific NAICS industry. This 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. 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.

Finally, some simple calculations permit the combination of *restrictions* and *industry relevance* into a single variable, *industry restrictions*, which 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.

We provide a short summary of the core features of RegData 3.0 below, while further details of each of these features, including the methodologies used to create them and several tables, are in section 3. Section 4 describes the datasets that are available for download.

Summary of the Core Features of RegData 3.0

For each part-level segment of the CFR annual editions from 1970 to 2016, RegData 3.0 provides the following:

- The CFR publication year, title number, and part number.
- A count of regulatory restrictions, denoted by the phrases *shall*, *must*, *may not*, *required*, and *prohibited*, both individually and in total.
- A word count.
- 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. A CFR part may be relevant to any number of industries or to no industry.

These data elements can be combined to produce the following annual series:

- 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 or department or a combination thereof, and
- Total regulatory restrictions.

Additional Capabilities Using QuantGov

Utilizing the open-source QuantGov framework, it is possible to both reproduce RegData and extend it. Modifications could include the following:

- A different definition of regulatory restrictions,
- A different unit of analysis than the CFR part,
- A different set of training documents to train the industry classifier, and
- Classification into a different set of categories besides the 2007 NAICS.

The QuantGov framework can also produce analysis of bodies of text unrelated to regulation. Any modification would require some amount of technical effort, but that would not be necessary to use any of the official RegData datasets.

3. METHODOLOGY

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 Code of Federal Regulations*, a current version of the CFR produced by the Office of the Federal Register and GPO, is used solely for error detection and smoothing.

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.

Analysis of Regulatory Text: Metadata

In the QuantGov platform, *metadata* refers to information obtainable from a single corpus without reference to other corpora or an estimator. The RegData metadata extracted from the CFR corpus include the word count, restriction count, count of each of the individual restriction terms, and authoring agency. Word counts are extracted using regular expressions, where a word is defined as a sequence of word characters bounded by non-word characters. Word characters are defined in the Python regular expression library to include letters, numbers, and the underscore.

Restriction counts enumerate the regulatory restriction phrases identified in Al-Ubaydli and McLaughlin (2015). Regulatory restrictions identify a specific mandated or prohibited activity, and they are thus a proxy for restrictiveness of a regulation. The specific terms counted are *shall*, *must*, *may not*, *required*, and *prohibited*. Enumerations for each term are included in the RegData metadata. Agency attribution is derived with the aid of the CFR's "Table of CFR Titles and Chapters."

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 2016 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. If classifications for a given industry do not exceed a minimum performance threshold (based on F1 scores), that industry is only included in a supplemental dataset labeled as "unfiltered." For some industries, it is not possible to produce classifications at all because of the low number of example documents.

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, Area under the Curve (probability version)	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 possible probability threshold from 0 to 1.
Receiver Operator Characteristic, Area under the Curve (binary classification version)	Receiver Operator Characteristic, Area under the Curve (probability version)	True positive rate (recall) multiplied by one minus the false positive rate.

The primary RegData 3.0 datasets are filtered to contain classifications only for those industries that pass a minimum performance threshold. 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 level	Industries	F1	Precision	Recall	Accuracy	ROC-AUC (varying threshold)	ROC-AUC (50% threshold)
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 devel	Industries	F1	Precision	Recall	Accuracy	ROC-AUC (varying threshold)	ROC-AUC (50% threshold)
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. DATASETS

Several different datasets are publicly available at RegData.org. Each RegData dataset is a zipped archive containing a bundle of .csv files. The tables below describe the variables in each of the files. Each table groups variables according to whether they are attributive (e.g., year of publication, agency, or location within the CFR) or quantitative (e.g., restrictions or industry relevance). We include NAICS codes and descriptions in the listing of quantitative variables—although they are attributive, they are attributes of industry relevance, not regulatory text.

Table 4. *metadata.csv*, Containing Variables on CFR Parts That Are Unrelated to Industry Classification

Attributive variables	
Name	Description
year	Year the part was published, from 1970 to 2016
title	Title containing the CFR part, from 1 to 50
part	Part number within the title, range varies by title
agency	Name of the agency, as given in the CFR Index
agency_id	Numerical identifier for agency, as defined by OMB budget codes. Each *agency_id* is a combination of the OMB codes for "bureau" and "agency."
department	Name of the department containing the agency, if any
department_id	Numerical identifier for the department, as defined by OMB budget codes for "bureau"
Quantitative variables	
Name	Description
words	Total number of words in the part
shall	Occurrences of the word *shall* in the part
must	Occurrences of the word *must* in the part
may not	Occurrences of the phrase *may not* in the part
required	Occurrences of the word *required* in the part
prohibited	Occurrences of the word *prohibited* in the part
restrictions	Occurrences of any regulatory restriction in the part

Table 5. *naics_Xdigit_relevance.csv*, Containing Variables Related to Industry Classification of CFR Parts

Attributive variables	
Name	Description
year	Year the part was published, from 1970 to 2016
title	Title containing the CFR part, from 1 to 50
part	Part number within the title, range varies by title
Quantitative variables	
Name	Description
naics_code	Industry code in NAICS level, where X (2, 3, 4, 5, 6) corresponds to levels in the 2007 NAICS hierarchy
industry_description	NAICS industry description
relevance	Probability that the CFR part is relevant to the industry

The "RegData Unfiltered" dataset contains the same files, but the relevance files are renamed to "naics_Xdigit_relevance_unfiltered.csv." The columns retain the same meaning.

RegData users often prefer to work with aggregated or summary data instead of with granular data at the level of the CFR part. We have created summary extracts in which data are aggregated to the agency level or the industry level—levels of aggregation that are fairly common in observed usage of RegData. The tables below summarize these downloadable RegData extracts. Table 6 describes the agency-level extract, and table 7 describes the industry-level extract.

Table 6. *agency_extract.csv*, Containing Annual Summary Statistics by Agency

Name	Description
year	Year of observation
agency	Name of the agency as given in the CFR Index
agency_id	Budget code for agency
department	Name of the department associated with the agency
department_id	Budget code for the department
words	Total number of words in agency text
shall	Occurrences of the word *shall* in agency text
must	Occurrences of the word *must* in agency text
may not	Occurrences of the phrase *may not* in agency text
required	Occurrences of the word *required* in agency text
prohibited	Occurrences of the word *prohibited* in agency text
restrictions	Occurrences of any regulatory restriction in agency text

Table 7. *naics_Xdigit_extract.csv*, Annual Industry-Level Summary Statistics, Where X (X =2, 3, 4, 5, 6) Corresponds to the Level of Granularity of NAICS Classifications

Name	Description
year	Year of observation
industry_code	Name of the agency as given in the CFR Index
industry	NAICS industry description
industry-relevant restrictions	Industry relevance of a CFR part multiplied by restrictions in that CFR part, summed across all parts, as introduced in Al-Ubaydli and McLaughlin (2015). NB: Previous versions of RegData have referred to this as the *Industry Regulation Index*.
industry-relevant words	Industry relevance of a CFR part multiplied by total word count of that CFR part, summed across all parts.

ABOUT THE MERCATUS CENTER

The Mercatus Center at George Mason University is the world's premier university source for market-oriented ideas—bridging the gap between academic ideas and real-world problems.

As a university-based research center, the Mercatus Center trains students, conducts research of consequence, and persuasively communicates economic ideas to solve society's most pressing problems and advance knowledge about how markets work to improve people's lives.

Our mission is to generate knowledge and understanding of the institutions that affect the freedom to prosper and to find sustainable solutions that overcome the barriers preventing individuals from living free, prosperous, and peaceful lives.

Since 1980, the Mercatus Center has been located at George Mason University.

Mercatus Center at George Mason University
3434 Washington Blvd., 4th Floor
Arlington, Virginia 22201
http://www.mercatus.org • http://quantgov.org/data