MERCATUS CENTER
George Mason University

POLICY BRIEF

RegData Canada: An Overview

Patrick A. McLaughlin, Scott Atherley, and Stephen Strosko

February 2019

The RegData Canada project from the Mercatus Center at George Mason University was launched in summer of 2018. The methodology for RegData Canada was initially established in a paper by Omar Al-Ubaydli and Patrick McLaughlin in 2015 and extended in papers by McLaughlin and coauthors in 2017 and McLaughlin and Oliver Sherouse in 2018.[1] To date, the RegData Canada project has released 14 unique datasets: a Canadian federal dataset and 13 provincial datasets. Following the methodology established in the original, US-focused RegData project (hereafter, RegData US), the datasets for the RegData Canada project (hereafter, RegData Canada) were created by applying text analysis and machine-learning algorithms to regulatory text issued by federal and provincial regulators. The datasets provide a variety of quantitative data and indicators, including regulatory restriction counts (by ministry or department), relevance of regulations to economic sectors and industries, the prevalence of incorporation by reference, linguistic complexity, the location of outdated language, and the likelihood that a regulation includes design standards. Additionally, these datasets have been used to create an interactive Canada Regulation Tracker, available online alongside the aforementioned datasets at QuantGov.org. RegData Canada allows policymakers to better identify and prioritize regulations that may need reform. RegData Canada follows the same methodology as other RegData projects, such as RegData US, facilitating comparative analysis of the economic effects of regulation and regulatory reform across Canadian provinces,[2] US states, and countries.

RegData Canada uses the algorithms developed for the most recent version of RegData US—RegData 3.1—to quantify regulation at both the federal and provincial levels through natural language processing and machine learning. This process yields the same two core components as all other RegData projects: quantification of regulatory restrictions and classification of regulatory text into the sectors and industries that are the likely subjects of regulations.

The views presented in this document do not represent official positions of the Mercatus Center or George Mason University.

Regulatory restrictions are words and phrases in regulatory text that indicate specific obligations or prohibitions created by the regulation. Following the RegData US methodology, RegData Canada datasets include quantifications of the following strings: *shall*, *may not*, *must*, *required*, and *prohibited*.

Like its American corollary dataset, the second core component of RegData Canada involves the estimation of the applicability of regulations. We use a set of customized machine-learning algorithms developed over the course of the RegData project that maps segments of regulatory text to the sectors and industries to which they are most relevant, based on the text of the regulation itself. This approach allows us to develop estimates of the number of restrictions and words that apply to specific sectors of the economy. We use the North American Industry Classification System (NAICS) to define industries. The NAICS standard is widely used across academia and in government, and it has the distinct advantage of being conceptually identical across the United States and Canada.[3]

SOURCES OF REGULATORY TEXT AND BASIC UNIT OF ANALYSIS

Regulatory text for this project is derived from three primary sources. First, for the federal regulation dataset, we use the XML point-in-time regulatory files provided by the Treasury Board of Canada Secretariat on the FTP server cited in the endnotes.[4] Second, for the 13 provincial datasets, we use regulatory text scraped or downloaded from individual provincial websites. Finally, several of our extended metrics leverage data from the current consolidated regulatory code available via the Canadian Department of Justice's laws and regulations portal. This section describes these data sources and the processing methods used and briefly discusses the relevant unit of analysis for the project.

Federal Regulatory Text: XML Consolidation of Point-in-Time Regulations

XML regulatory text accessed via FTP server serves as the source of text for the core metrics in the RegData Canada federal dataset. We leverage the point-in-time files—specifically, the PITXML/regulations folder—to construct a time series of changes in the consolidated regulatory code. Statutory instruments (flagged with "SI" prefixes) are excluded from analysis.

Regulations are mapped to departments using the Table of Public Statutes and Responsible Ministers and the regulation's enabling authority,[5] recorded in the XML. Mapping regulations to departments makes it possible to summarize the regulatory code by department. Out of all the regulations represented in the RegData Canada federal dataset, only 0.6 percent were unable to be matched to a designated department.

Provincial Regulatory Text: Independent Scraping and Downloading Algorithms

Data collection for the provincial regulatory codes was considerably more involved than the federal collection process. The 13 individual RegData Canada provincial datasets take an almost identical form to the RegData Canada federal dataset. Unlike the federal dataset, however, the provincial datasets are derived from 13 distinct locations online, each with its own quirks. The sources for the text used to form each dataset and any important notes for that dataset are listed in appendix A.

In addition to differences in sources, the RegData Canada provincial datasets differ in two major ways from the federal dataset. First, the provincial datasets are cross sections rather than time series. Many provinces do not make point-in-time consolidations available, and bulk collection tends to be more difficult for those that do. This is not unusual; many US states also do not provide point-in-time examples of their regulatory codes (in fact, one state—Arkansas—does not even provide a current version of its code). Second, we have not yet collected department mappings for any of the Canadian provinces; this is mainly owing to the variability of provincial websites, which made collecting this information more challenging than it was in the federal case. This is an opportunity for future work.

Finally, the precision of the tools that we use to collect, read, and analyze regulatory text rely on the consistency and presentation of the input documents. Across many of the provinces, the input documents were older PDF files, many of which were bilingual. We developed several algorithms to deal with bilingual documents (essentially a combination of splitting columns in PDF files and detecting lines with large numbers of French words), but this is not a perfect process. Consequently, word count variables (discussed in the following section) should be considered noisier across the provincial datasets compared with the RegData Canada federal dataset. However, this point does not apply to other variables that we present in the following section, such as restriction counts.

Federal Regulatory Text: Department of Justice Portal

The final source of regulatory text used in this project is the current form of the Canadian federal regulatory code, obtained via the Department of Justice's laws and regulations portal.[6] These data were collected via web scraping. The main advantage of this source over the provided XML files is its HTML markup, which provides some additional information relevant to several metrics we constructed. These additional metrics are discussed later in this brief.

Unit of Analysis for Regulatory Code

We analyze regulatory text at the individual regulation level (as opposed to splitting into more granular units, such as sections or paragraphs). Splitting text into smaller sections would involve a series of judgments about what constitutes a "subregulation," which introduces subjectivity

into the process of classifying the text. For example, consider SOR/83-190, Gas Pipeline Uniform Accounting Regulations.[7] Like many regulations, this document contains a series of sections and subsections, some of which are standardized (section 1 denotes Short Title, section 2 denotes Interpretation [i.e., definitions], etc.). In theory, it might have been feasible to split regulations into these sections and subsections. In practice, however, there are complications involved in attempting this. While sections are easily identified, at least in that particular regulation, it is not so easy to automatically identify "clear" differentiations between topically distinct subsections (or even paragraphs).

Finally, based on our experience with US federal regulations as well as subnational regulations (i.e., state regulations), specific sections of regulatory text tend to focus on a specific industry or a closely related set of industries. It is generally not the case that a given section of the *Code of Federal Regulations* (CFR) discusses disparate industries in the same group of sentences. Our investigation of the Canadian code so far seems to bear this out. A given regulation may cover a variety of policy issues yet remain applicable to the same major industry throughout. For our purposes—and especially for the purpose of using machine-learning algorithms to classify regulatory text by industry—the fact that individual regulations tend to be industry specific means the appropriate unit of analysis is likely to be the individual regulation.

REGDATA CANADA METHODOLOGY AND CORE METRICS

The information in this section will review the core metrics found in RegData Canada. These metrics map to the variable names found in the corresponding RegData Canada datasets. For additional information on specific metrics, refer to the RegData 3.0 User's Guide.[8]

Word Counts

"Words," by the strict definition of the typical parser (including the one we use), are not always literally English words. We allow the algorithm to parse other discrete units of information, such as numbers and citation references, as if they are words. We take this approach in order to capture a very broad definition of a given regulation's size alongside a much more focused one (restrictions). And in a sense, the number of "tokens" (a common term in natural-language processing referring to independent units of language) present in a given text is a measure of size. Certain types of regulations will receive relatively inflated word counts using this approach. Most obviously, any regulation consisting primarily of tables of numbers will be recorded as having many "words," even though it may merely consist of many pages of numbers. In any case, the logic of the metric remains the same: in order to fully comprehend a regulation, the regulated entity must read and understand these tables (and citations, abbreviations, etc.). One might also conceptualize word count as a high-level measurement of document size.

Regulatory Restrictions

"Restrictions," as measured in data from the RegData project, are designed as a cardinal proxy for the number of regulatory restrictions contained in regulatory text. This variable is devised by counting select words and phrases that are typically used in legal language to create binding obligations or prohibitions. The words used to proxy the number of restrictions are *shall*, *must*, *may not*, *required*, and *prohibited*. This relatively straightforward approach to summarizing regulatory text has a rich history in the United States and Canada. Where point-in-time compilations of regulatory text are available, shifting restriction counts serve as an effective proxy for the aggregate regulatory burden in a given jurisdiction. Counting restrictions is also a considerable improvement on earlier research, which typically emphasized page counts.

Industry Relevance

The final component of RegData Canada's core metrics is industry relevance. A regulation's industry relevance is the probability that the specific regulation is associated with the specific industry. RegData Canada maps Canadian regulations to the sectors and industries most likely to be associated with them through the North American Industry Classification System (NAICS).[9] NAICS codes are used in a wide variety of economic research when categorizing by industry. RegData's use of NAICS codes permits users to merge RegData with other datasets that may reflect the results of regulatory policies that interact at the industry level. 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. Similarly, Statistics Canada reports many industry metrics that are built around NAICS, such as the Business Dynamics series.[10] RegData Canada features data at the 3-digit NAICS level.

Industry Restrictions

The design of the restriction counts and industry relevance variables facilitates the construction of a single variable, "industry restrictions," which approximates the total number of regulatory restrictions that are relevant to a particular industry or set of industries. Note that the process of aggregation used to construct this metric results in overlap by design. Some regulations, such as those involving labor and workplace safety, may apply to nearly the entire productive economy.

The industry restrictions variable allows for verifiable estimates of regulatory accumulation across time and across sectors of the economy. It facilitates a great deal of economic research that would otherwise be impossible. Prior studies of regulation were limited to annual time series, such as pages in the *Federal Register* or the CFR. RegData allows for complete time-series and cross-sectional research designs and integration with standard national economic aggregations; for example, many regulatory and statistical agencies in both the US and Canada publish a number

of statistical data series using NAICS codes as the basis of comparison and aggregation. A number of studies have exploited these features to investigate various features of the US economy. Several were cited earlier in this document at endnote 5.

HOW MACHINE LEARNING IS USED TO DETERMINE INDUSTRY RELEVANCE

In order to assign industry classifications to each body of text, we employ a set of machine-learning algorithms called supervised classification to assess the probability that a unit of regulatory text targets a specific NAICS industry. The first step in supervised classification is to collect training documents that can be used by the program to learn what words, phrases, and other features can best identify when a unit of text is relevant to a specific industry.

Training document selection. Our training documents are derived from the XML version of the *Federal Register*.[11] 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 nonoverlapping, 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 (Chemical Manufacturing). While most industry names are only meaningful in the context of the industry, a few (such as NAICS 51: Information) have more generally used names and are therefore blacklisted as matches, although child and parent industries of these are still used.

Training document exceptions. Two US agencies, the Small Business Administration and the Office of Personnel Management, both frequently publish informational rules about, for example, the definition of a small business within each of the NAICS industries—these are not actually restrictions on those industries, so rules from these two agencies are excluded. Finally, any rules that match more than 2.5 standard deviations above the mean number of matched industries are assumed not to be industry specific and are therefore excluded from the industry classification process.

Training document processing. 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 at least five individual documents that are positively labeled for that industry. The training documents were vectorized using bigram counts, which carry more semantic specificity than unigrams. Words for vectorization were defined as sequences of two or more alphabetic characters, with English stop words excluded, and bigrams were defined as two consecutive words.[12] The vocabulary was limited to bigrams occurring in at least 0.5 percent but not more than 50 percent of trainers to keep the number of features computationally manageable and to exclude very common phrases.

Because a regulation can be relevant to more than one industry, we use a multilabel approach for classification. We tested one parametric and one nonparametric model: a logistic regression (Logit) implemented with the LASSO (L1 penalty and regularization) and a random forest.[13] The logit model was designed for multilabeling using a one-versus-rest strategy. Simply put, the final algorithm runs a sequence of classification models for each industry and produces probabilities for each document for all industries at the relevant level of the NAICS hierarchy. The probabilities reflect the algorithm's estimate of the likelihood that the document is about a given industry.

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 parameters selected using cross-validation, we estimated a variety of performance metrics for each class individually (F1, precision, recall, accuracy, and ROC curves). Each of these metrics is a means of measuring the performance of a classifier. Recall and accuracy are straightforward success metrics, while F1 and AUC (area under the ROC curve) attempt to balance competing classification priorities. Accuracy is the simplest possible classification metric: it measures the percentage of correct classifications as a percentage of total classification attempts. Recall denotes the percentage of correct classifications among items classified as correct. For example, if we are predicting whether or not a document is relevant to Air Transportation, and we score 10 documents as relevant, but 19 in the data should have been scored as relevant, we would score a recall of 10/19. This measure is an improvement over accuracy, particularly in unbalanced classification problems, as classes with large numbers of 0-values (nonmembership) will dominate the accuracy statistic. This is illustrated in endnote 10. AUC and F1 scores are somewhat more involved. The F1 score balances recall and precision—the number of correctly classified values as a percentage of values *identified as correct by the model*. The AUC calculation measures the probability that a given model will rank a randomly chosen positive instance of a class higher than a randomly chosen negative instance.

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.[14] 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 training documents.

The primary RegData 3.1 datasets (including RegData Canada) 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, clearly marked, unfiltered dataset for researchers wishing to use their own threshold. Median normalized score results for each NAICS level are presented in table 1, while median nonnormalized score results are presented in table 2.

Table 1. Normalized Metric Scores

NAICS LEVEL	TOTAL 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 2. Nonnormalized Metric Scores

NAICS LEVEL	TOTAL 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

APPENDIX A. SOURCES FOR PROVINCIAL REGULATORY TEXT

Alberta: "Alberta Queen's Printer" home page, accessed August 6, 2018, http://www.qp.alberta.ca/index.cfm.

British Columbia: Queen's Printer for British Columbia, "Consolidated Regulations of British Columbia," accessed August 6, 2018, http://www.bclaws.ca/civix/content/crbc/crbc/?xsl=/templates/browse.xsl.

Manitoba: Manitoba, "Consolidated Regulations of Manitoba," accessed August 6, 2018, http://web2.gov.mb.ca/laws/regs/index.php.

New Brunswick: Attorney General of New Brunswick, "Browse Regulations by Regulation Number," accessed August 6, 2018, http://laws.gnb.ca/en/BrowseRegChapter?letter=all.

Newfoundland and Labrador: Newfoundland and Labrador Office of the Legislative Counsel, "Alphabetical List of Regulations," accessed July 9, 2018, https://www.assembly.nl.ca/legislation/sr/regulations/titleindex2.htm.

Northwest Territories: Justice Department of the Northwest Territories, "Legislation of the Northwest Territories," accessed July 12, 2018, https://www.justice.gov.nt.ca/en/legislation /#gn-filebrowse-0:/.

Nova Scotia: Nova Scotia Department of Justice, Registry of Regulations, "Regulations Listed by Act," accessed July 19, 2018, https://www.novascotia.ca/just/regulations/regsxact.htm.

Nunavut: Nunavut Legislation, "Current Consolidated Statutes and Regulations," accessed July 10, 2018, https://www.nunavutlegislation.ca/en/consolidated-law/current?title=A.

Ontario: Government of Ontario, "Consolidated Laws," accessed July 2, 2018, https://www.ontario.ca/laws.

Prince Edward Island: Government of Prince Edward Island, "Statutes and Regulations," accessed July 19, 2018, https://www.princeedwardisland.ca/en/legislation/all/all/all.

Quebec: Publications Quebec, *LégisQuébec*, accessed September 12, 2018, http://legisquebec.gouv.qc.ca/en/BrowseChapter?corpus=regs.

Saskatchewan: Government of Saskatchewan, "Regulations," accessed September 12, 2018, http://www.publications.gov.sk.ca/deplist.cfm?d=1&c=43.

Yukon: Yukon Government, "Acts and Regulations," last updated March 31, 2018, http://www.gov.yk.ca/legislation/legislation/page_a.html.

ABOUT THE AUTHORS

Patrick A. McLaughlin is the director of Policy Analytics and a senior research fellow at the Mercatus Center at George Mason University. His research focuses primarily on regulations and the regulatory process. He created and leads the RegData and QuantGov projects, deploying machine-learning and other tools of data science to quantify governance indicators found in federal and state regulations and other policy documents. McLaughlin has authored more than a dozen peer-reviewed studies in diverse areas, including regulatory economics, administrative law, industrial organization, and international trade.

Scott Atherley is a former research fellow at the Mercatus Center at George Mason University.

Stephen Strosko is a data engineer for Policy Analytics at the Mercatus Center at George Mason University. Stephen specializes in regulatory research and has notably worked on the RegData, Quantgov, FRASE, and RegData Canada projects.

NOTES

1. Omar Al-Ubaydli and Patrick A. McLaughlin, "RegData: A Numerical Database on Industry-Specific Regulations for All U.S. Industries and Federal Regulations, 1997-2012" (Mercatus Working Paper, Mercatus Center at George Mason University, Arlington, VA, 2014); Patrick A. McLaughlin et al., "RegData 3.0 User's Guide," accessed February 15, 2018, https://quantgov.org/regdata/users-guide/; Patrick A. McLaughlin and Oliver Sherouse, "RegData 2.2: A Panel Dataset on US Federal Regulations," *Public Choice*, online (2018): 1–13.
2. Provinces in this sense will refer to both Canadian provinces and Canadian territories.
3. NAICS classifications range from extremely broad economic sectors (2-digit codes such as NAICS 52: Finance and Insurance) to very specific industries (6-digit codes such as NAICS 315225: Men's and Boys' Cut and Sew Work Clothing Manufacturing). For a detailed description of the NAICS classification standard, see US Census Bureau, "North American Industry Classification System: Introduction to NAICS," accessed August 31, 2018, https://www.census.gov/eos/www/naics/.
4. Treasury Board of Canada Secretariat, Canada Federal Point in Time Regulations (regulatory database), accessed November 2, 2018, ftp://205.193.86.89/.
5. Government of Canada, Justice Laws Website, "Table of Public Statutes and Responsible Ministers," last modified August 22, 2018, http://laws-lois.justice.gc.ca/eng/TablePublicStatutes/index.html.
6. Government of Canada, "Consolidated Regulations of Canada," January 8, 2019, http://laws-lois.justice.gc.ca/eng/regulations/.
7. Government of Canada, "Gas Pipeline Uniform Accounting Regulations," January 8, 2019, http://www.laws.justice.gc.ca/eng/Regulations/SOR-83-190/FullText.html.
8. McLaughlin et al., "RegData 3.0 User's Guide."
9. US Census Bureau, "North American Industry Classification System: Introduction to NAICS."
10. Statistics Canada, "Business Dynamics Measures, by Industry," January 10, 2019, https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3310016401.
11. The *Federal Register* is the US equivalent of the *Canada Gazette*.

12. Stop words are generally defined as words that do not contain any useful information in search queries or text segmentation tasks. At this point in the process, we use a common dictionary of English stop words (containing words such as *the*, *as*, *a*, etc.). Stop words may also be identified empirically (i.e., by testing for words that appear in an extremely high percentage of documents and provide no useful ability to assist segmentation). In our case, words meeting this definition are dealt with later in the modeling process, via the TF-IDF weighting algorithm.

13. The LASSO augments regression and classification algorithms to incorporate both variable-selection and regularization (in the form of a penalty function designed to prevent overfitting) components.

14. Think of this problem like detecting credit card fraud. Almost all transactions are legitimate, just as almost all documents are irrelevant to a given industry (the classes are much less unbalanced in our case, but the concept is similar). Therefore, one can obtain a very high accuracy score by predicting that fraud never occurs, but it will not be a helpful prediction. In our case, it is straightforward to achieve an illusory sense of high accuracy by never classifying a document as relevant to a given industry. This is the baseline model. Improvements over this baseline should represent legitimate gains in useful predictive accuracy.