MERCATUS CENTER
George Mason University

POLICY BRIEF

RegData Canada: A Data-Driven Approach to Regulatory Reform

Patrick A. McLaughlin

March 2019

RegData Canada, a data project from the Mercatus Center at George Mason University that was launched in 2018, allows regulators and policymakers to better identify and prioritize regulations that may need reform. The RegData Canada project involves applying customized text-analysis software and machine learning algorithms to regulatory text issued by federal and provincial regulators, resulting in 14 unique datasets: a Canadian federal dataset and 13 provincial datasets. All of these datasets are freely available online at QuantGov.org.

RegData Canada provides a variety of quantitative data and indicators, including

- regulatory restriction counts,
- 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- the location of outdated language, and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 prescriptive design standards.

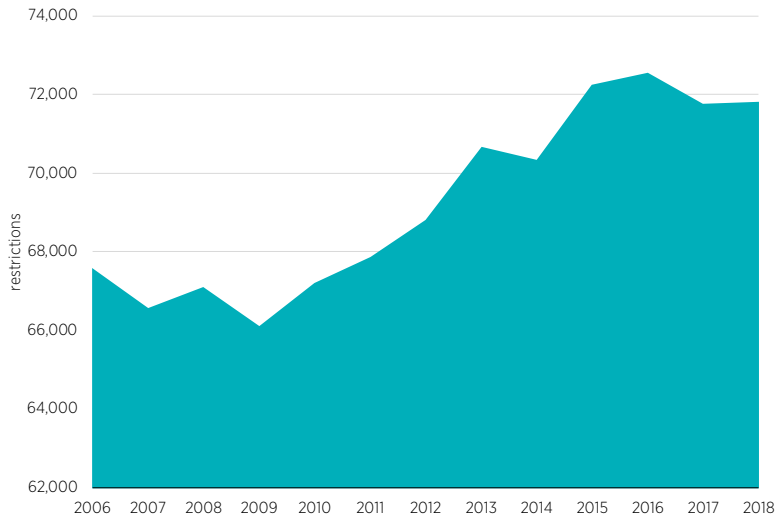
We describe some of these below. Additionally, these datasets have been used to create an interactive Canada Regulation Tracker, available online alongside the aforementioned datasets at QuantGov.org.

REGULATORY RESTRICTIONS

Regulatory restrictions are words and phrases in regulatory text that indicate specific obligations or prohibitions created by a regulation.

RegData Canada datasets include counts of the following words and phrases: *shall*, *may not*, *must*, *required*, and *prohibited*. Figure 1 shows the results for all Canadian federal regulations.

Figure 1. Federal Regulatory Restrictions, 2006–2018

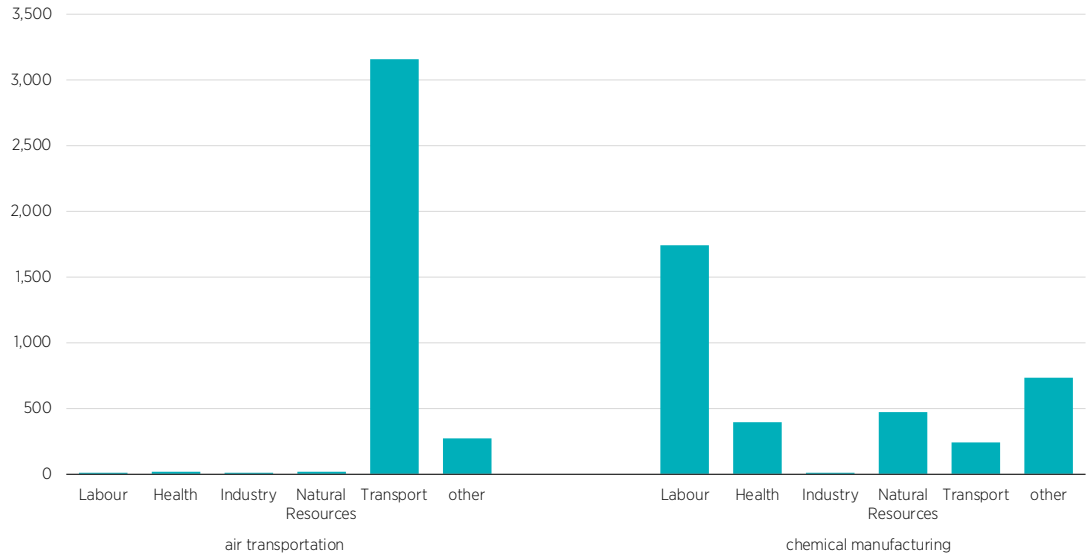


Source: Patrick A. McLaughlin, Scott Atherley, and Stephen Strosko, RegData Canada (dataset), QuantGov, Mercatus Center at George Mason University, Arlington, VA, 2018, https://quantgov.org/regdata-canada/.

INDUSTRY AND SECTOR DATA

The second core component of RegData Canada involves the estimation of the applicability of regulations. We use a set of machine learning algorithms developed over the course of the RegData project—which initially launched in 2012—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.[1] 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.[2] Figure 2 shows the results for two sample industries, broken down by ministry.

Figure 2. Federal Industry-Relevant Regulatory Restrictions for Select Industries

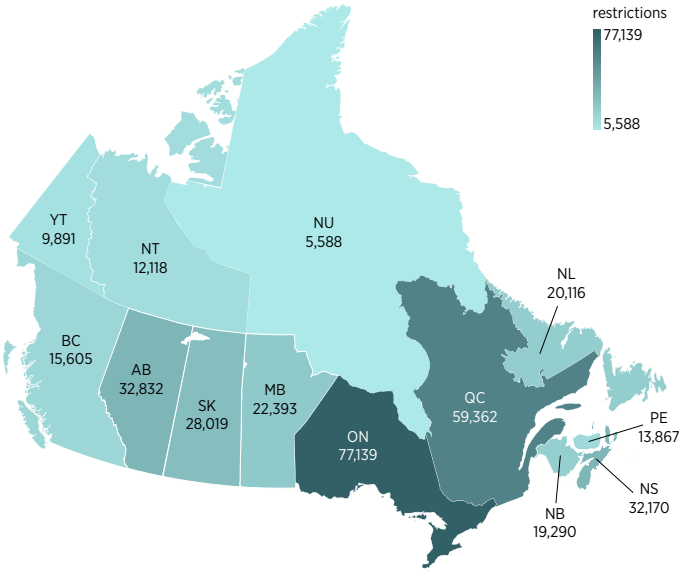


Source: McLaughlin, Atherley, and Strosko, RegData Canada (dataset).

PROVINCES AND TERRITORIES

In addition to processing Canadian federal regulations, we also apply a comparable methodology to regulations published by individual provinces and territories. The 13 Canadian province and territory datasets offer a comprehensive cross section of provincial and territorial regulations as of July 2018. Figure 3 presents total restriction counts across the provinces and territories.

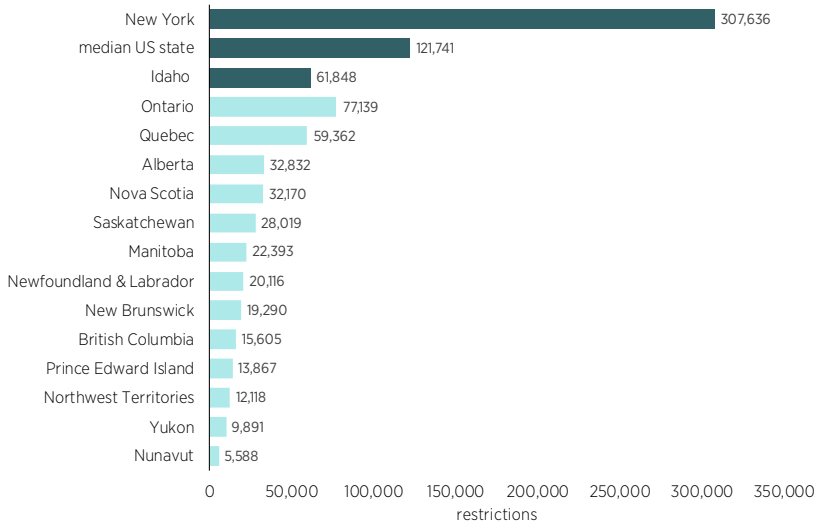
Figure 3. Map of Provincial Regulatory Restrictions



Source: McLaughlin, Atherley, and Strosko, RegData Canada (dataset).

Figure 4 also presents provincial and territorial regulatory restriction counts alongside some results from a sample of US states. The US states selected for figure 4—New York and Idaho—represent the maximum and minimum (respectively) restriction counts of the states for which we have data at this point in the project. The median US state is Maryland at the time of this writing.

Figure 4. Provincial Regulatory Restrictions and Sample US State Regulatory Restrictions



Source: McLaughlin, Atherley, and Strosko, RegData Canada (dataset); Patrick A. McLaughlin, Oliver Sherouse, Daniel Francis, and Jonathan Nelson, State RegData (dataset), QuantGov, Mercatus Center at George Mason University, Arlington, VA, 2018, https://quantgov.org/state-regdata/.

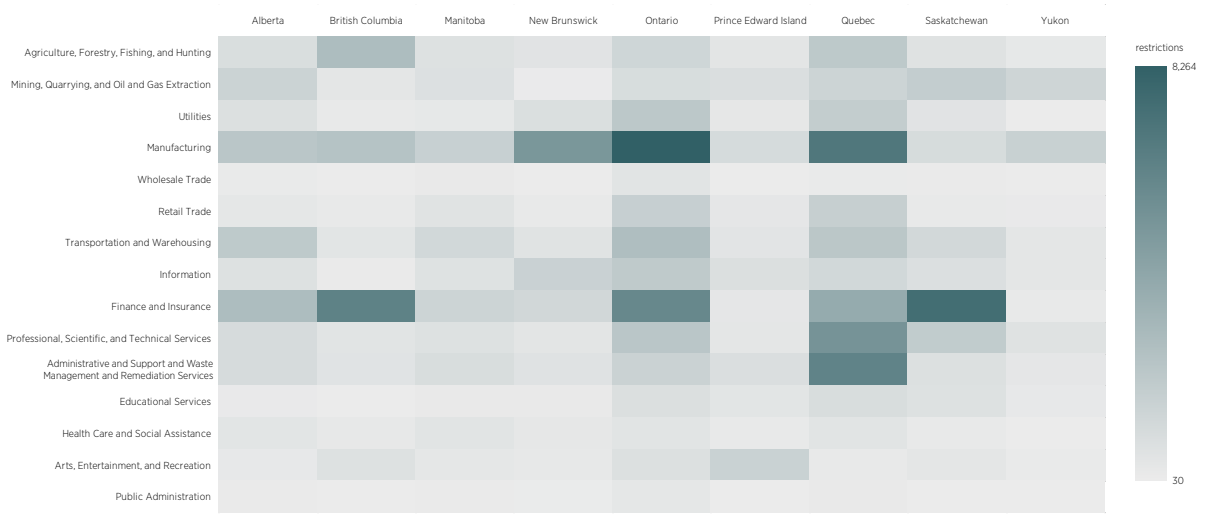
Restrictions by industry vary considerably across the provinces. Oil and gas extraction and related industries show particularly high levels of restrictions in high-production areas, including Alberta and the Northwest Territories, where such activities make up a substantial portion of the provincial GDP. By contrast, Ontario and British Columbia, whose economies are comparatively more service oriented, show more service-related industry restrictions, particularly on financial and insurance-related industries and professional services.

Figure 5 summarizes total restriction counts by substantive NAICS two-digit industries and Canadian provinces. This graphic allows for comparisons of regulatory distribution. For example, finance and insurance regulations are abundant in British Columbia and Saskatchewan, while manufacturing is more highly regulated in Alberta and Ontario.

LINGUISTIC COMPLEXITY

The complexity of a regulation is significant for a number of reasons. Prominently, complexity is expected to raise compliance costs, as regulated entities need to spend more time to under-

Figure 5. Regulatory Restrictions by Industry, Province



Source: McLaughlin, Atherley, and Strosko, RegData Canada (dataset).

stand complex regulations. This may force regulated entities to employ more lawyers, which is an additional cost.

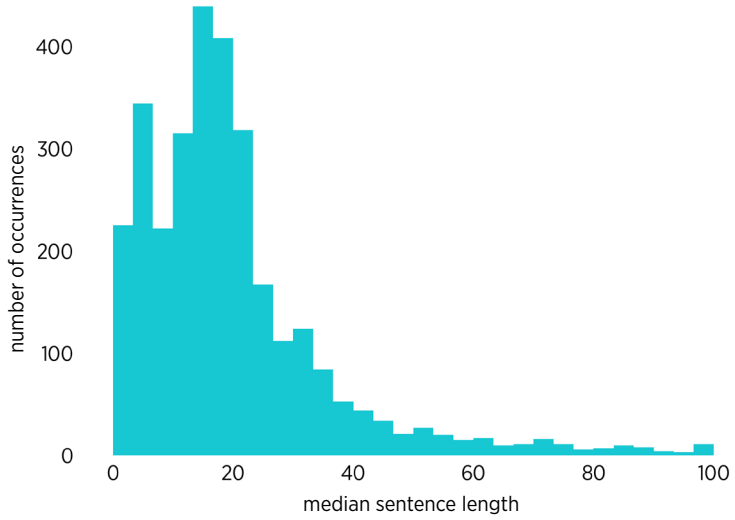
RegData uses two different metrics to compare the complexity of regulations. The first of these is a simple, commonly used measure: sentence length. This measure uses the median-length sentence in a document, which avoids outliers caused by limitations in parsing sentences, such as large tables or other nonstandard bodies of text.

Figure 6 shows the distribution of median sentence length for all of the regulations. While the analysis shows a mean of 25 words per sentence across all regulations, the median for all regulations is 17, indicating that many regulations are in line with the Canadian Treasury Board recommendation of 20 words per sentence.[3] There are, however, a number of regulations with greater than the recommended 20-word-per-sentence average, which may have room for improvement in terms of readability by rewriting the regulations to simplify or break up sentences.

The second metric examined was Shannon entropy, which is slightly more complex.[4] Shannon entropy measures, in broad terms, the frequency of new ideas introduced in documents, with simpler and more focused documents having a lower entropy score. The average federal regulation has an entropy score of 6.79. For the sake of comparison, compositions by Shakespeare tend to have an entropy score between 9.0 and 9.7.

Figure 7 shows the distribution of Shannon entropy for federal regulations. The red area shows the range of Shannon entropy for Shakespeare's major plays. The figure readily suggests that a

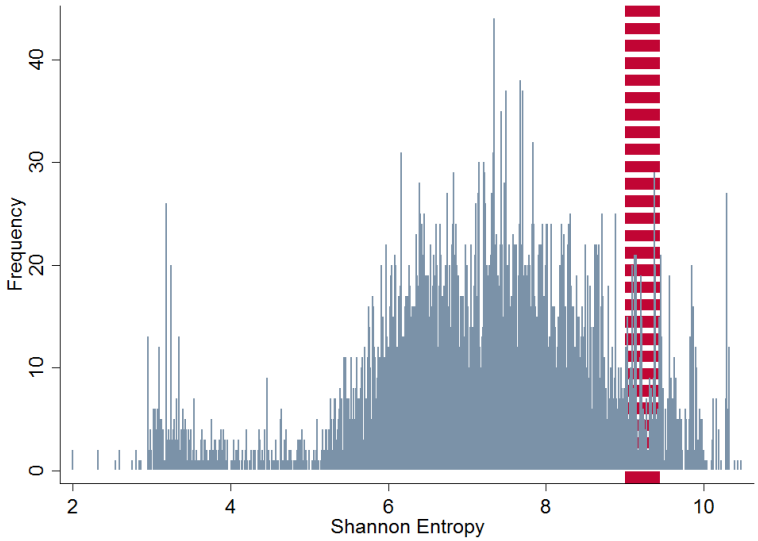
Figure 6. Sentence Length of Federal Regulations



Source: McLaughlin, Atherley, and Strosko, RegData Canada (dataset).

"Shakespeare Test," in which any regulation with a Shannon entropy score greater than the Shakespeare range would qualify for a closer read to determine if it needs simplification or updating, may be useful.

Figure 7. Shannon Entropy of Federal Regulations and of Shakespeare's Plays



Source: Author's calculations; Marcin Lawnik, "Shannon's Entropy in Literary Works and Their Translations," *Journal of Computer Science* 1, no. 3 (2012): 1–3.

ABOUT THE AUTHOR

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.

This policy brief was produced in part using **QuantGov**, a policy analytics platform that facilitates analysis of the causes and effects of various government actions. The QuantGov project treats policy text as data, allowing researchers to quickly and effectively examine broad policies (as articulated in bodies of text) by using some of the latest advances from data science, such as machine learning and other artificial intelligence technology. The Mercatus Center's team of data engineers, analysts, and developers created this platform and continually utilize and update it to produce data that support a variety of research products and to provide policymakers with data that inform positive policy change. More information is available at **quantgov.org**.

Patrick A. McLaughlin
Policy Analytics Director

Stephen Strosko
Data Engineer

Jonathan Nelson
Software Developer

Thurston Powers
Data Analyst

NOTES

1. For details on our machine learning algorithms, see Patrick A. McLaughlin, Oliver Sherouse, Daniel Francis, Michael Gasvoda, Jonathan Nelson, Stephen Strosko, and Tyler Richards, "RegData 3.0 User's Guide," accessed January 28, 2019, https://quantgov.org/regdata/users-guide/.
2. NAICS classifications range from extremely broad economic sectors (two-digit codes such as NAICS 52: Finance and Insurance) to very specific industries (six-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, "Introduction to NAICS," December 3, 2018, https://www.census.gov/eos/www/naics/.
3. Government of Canada, "Canada.ca Content Style Guide," November 13, 2018, https://www.canada.ca/en/treasury-board-secretariat/services/government-communications/canada-content-style-guide.html#toc5.
4. C. E. Shannon, "A Mathematical Theory of Communication," *Bell Systems Technical Journal* 27, no. 3 (1948): 379–423.