Quarterly Data Release Report

July 2020

Introduction:

Welcome to the second quarterly data release report from the QuantGov team at the Mercatus Center at George Mason University. In this data release, our team is looking to continue the facilitation of effective and informative policy-relevant research through the quantification of policy. In this release, the QuantGov team focused primarily on producing the second iteration of one of our biggest RegData projects, State RegData. Additionally, we have produced a new dataset that counts the number of terms in the Federal Register that remove regulation from the Code of Federal Regulation (CFR).

While State RegData 1.0 is both unique and informative, the data was developed over the span of three years, leading to inconsistencies that arise during any innovative process. When comparing data from one jurisdiction to another, those data will ideally use a consistent methodology for collection and analysis, while also being collected during the same relative time period. While collection method, methodology, and time were all held relatively consistent during the creation of State RegData 1.0, the overall result was sub-optimal for a couple of reasons. To start, each state publishes their regulatory code in a different structure and on different websites, which requires at least one customized Python web-scraping script for each state. Naturally, as our team wrote each of these scripts, our method of data collection naturally improved over time. Eventually, we developed a playbook that included best practices, processes, pipelines, and quality assurance checkpoints informed by experience.

The second iteration of State RegData leverages the lessons learned in the development of State RegData 1.0 and provides a consistent snapshot of the 2020 regulatory landscape in all of the states. Rather than gathering document over three years, State RegData 2.0 saw the collection of each state regulatory code within a span of three months. Instead of having to write a new playbook, the QuantGov team utilized the previously built pipelines and processes and applied consistent quality assurance checks to the data collected. State RegData 2.0 provides a solid basis for genuine comparative analysis of regulations across different states and regions.

Broadly speaking, the efficiency gains from making an initial time and resource downpayment for developing people, processes, and tools is the bread and butter of what the team here at QuantGov strives for. Our hope is that every 2.0, 3.0, 4.0 etc. version of any data project will take less time, be more accurate, and most importantly, lead to more effective research. Without further ado, here are the high level details about our Q2 data additions!

Stephen Strosko

Data Engineer
State RegData 2.0

Full Documentation for State RegData 2.0

State RegData 2.0 is the second iteration of the State RegData data project. Version 2.0 contains data from 44 states and the District of Columbia. Alaska, Arkansas, Connecticut, Hawaii, New Jersey, and Vermont were not included in the 2.0 version due to not having a regulatory code, a usable website, or paywalls (or a combination of the three). If we are able to acquire the regulatory code for any of these states in the near future, we will release an update (e.g., State RegData 2.1). All 44 states’ regulatory codes were collected between the dates 2020-03-09 and 2020-06-12, providing users with comparable point-in-time data for the year 2020. As version 2.0 is a continuation of State RegData 1.0, the following metrics will continue to be included:

- Regulatory restrictions and word counts for each state regulatory document.
- Probabilistic association for each state regulatory document to the 3-digit NAICS industries.
- Metadata describing each state regulatory document.

In addition to these metrics, the 2.0 version provides the updated and new metrics for every state in the 2.0 version:

- Complexity metrics for each state regulatory document.
- Agency/department oversight for each state regulatory document.
- Improved unit of analysis aggregation.

The QuantGov team is really excited about the new metrics in State RegData 2.0. Each state regulatory document is now associated with three different measures of complexity: Shannon Entropy, Sentence Length, and Conditional Counts. Shannon entropy measures the likelihood of encountering new words and concepts in a given body of text. Conditional counts, counts the number of branching words such as “if”, “but”, and “provided” found in any given piece of text. Sentence length measures the mean sentence length in any given body of text.

In the 2.0 release, each state document is now associated with an issuing agency or department. This is one of the more commonly referenced features in our Federal RegData datasets, and is now available at the state level. Finally, the QuantGov team improved unit of analysis aggregation with the 2.0 release. Each state regulatory code is written in a different structure. Some states simply have titles filled with regulations, while other states have more complex structures. An example of this is Texas, which has titles, parts, chapters, subchapters, divisions, and rules. To be able to compare regulations across states, the QuantGov team had to determine the appropriate level of unit of analysis aggregation for each state – so that researchers can better compare apples to apples.

Sample API Calls for State RegData:

1) Retrieve the number of restrictions in both version 1.0 and version 2.0 for the state of Nevada:
2) Retrieve the number of restrictions in Pennsylvania the Department of Environmental Protection is responsible for:

3) Retrieve the complexity metrics for Florida’s regulatory documents:

Deregulation Data

[Full Documentation for Deregulation](#)

Another dataset that is being released in the second quarter data release is the deregulation dataset. The deregulation dataset attempts to capture active attempts at deregulation of the [Code of Federal Regulations](https://www.gpo.gov/fdsys/browse/collection.action?size=full&sys=GPO) through the publication of final rules in the [Federal Register](https://www.federalregister.gov). While RegData projects successfully track changes to the total stock of regulation RegData projects don’t offer much insight into changes that are in the flow of regulations, which is described in the [Federal Register](https://www.federalregister.gov) in the form of notices of proposed rulemakings and final rules. Furthermore, some changes could include additions of text to the stock of regulation that are actually **deregulatory** in nature, despite adding more text. This dataset attempts to capture just that, deregulatory text, as indicated by its presence in the [Federal Register](https://www.federalregister.gov).

Specifically, the deregulation dataset provides counts of the following terms: **13771, delay, deregulate, eliminates, eliminate, unnecessary, obsolete, repeal, withdraw, withdrawing, remove, and replace**. We have identified these terms as terms that strongly proxy deregulatory activity. This release will provide data ranging back to 2012, with future iterations pushing that data back even further.

General Data Additions and Fixes

There are a few general additions and fixes to our existing data library this quarter. To start, a few states from State RegData 1.0 contained errors that were found during the 2.0 collection process. For the most part, all of these were extremely small, though Texas was found to have significantly more regulation than previously thought. All of these changes have been implemented and both 1.0 and 2.0 data points now contain accurate numbers.

The API has been updated to include citation and documentation information when a call for data is made. The team encourages users to read the documentation that is associated with any downloaded data.
Questions or Data Inquiries?
info@quantgov.org

Media Inquiries or Interview Requests?
media@mercatus.gmu.edu