Regulation & Governance

Regulation & Governance (2017) 11, 109–123 doi:10.1111/rego.12107

RegData: A numerical database on industry-specific regulations for all United States industries and federal regulations, 1997–2012

Omar Al-Ubaydli

International and Geo-Political Studies, Bahrain Center for Strategic, International and Energy Studies and George Mason University

Patrick A. McLaughlin

Mercatus Center, George Mason University

Abstract

We introduce RegData, formerly known as the Industry-specific Regulatory Constraint Database. RegData annually quantifies federal regulations by industry and regulatory agency for all federal regulations from 1997–2012. The quantification of regulations at the industry level for all industries is without precedent. RegData measures regulation for industries at the two, three, and four-digit levels of the North American Industry Classification System. We created this database using text analysis to count binding constraints in the wording of regulations, as codified in the *Code of Federal Regulations*, and to measure the applicability of regulatory text to different industries. We validate our measures of regulation by examining known episodes of regulatory growth and deregulation, as well as by comparing our measures to an existing, cross-sectional measure of regulation. Researchers can use this database to study the determinants of industry regulations and to study regulations' effects on a massive array of dependent variables, both across industries and time.

JEL codes: K2, L5, N4, Y1

Keywords: industry, RegData, regulation, regulatory accumulation.

1. Introduction

Scholars have been analyzing government regulation for decades because of its importance as a means of addressing market failure (Pigou 1938). However, regulatory policies may not always be virtuously conceived (Stigler 1971; McChesney 1987), and they may have adverse unintended consequences (see Peltzman 1975; for a more thorough discussion of the different theories of regulation, see Djankov *et al.* 2002).

Studies typically examine the causal effect of a unique regulation or a small collection of related regulations, such as air quality standards (e.g. Greenstone 2002), meaning that the intervention being studied is relatively limited in scope, even if its effects can be far-reaching. With a few notable exceptions, there has been no attempt to create aggregate time-series measures of regulation based on legal regulatory documents. Previous efforts have used proxies for the quantity of government regulations created or in effect each year.[1] Mulligan and Shleifer (2005) use the sizes of digitized versions of state-level statutes as a proxy for real state-level regulation. Coffey *et al.* (2012) use the

Correspondence: Patrick A. McLaughlin, Mercatus Center, George Mason University. Email: pmclaughlin@mercatus.gmu.edu

Accepted for publication 18 September 2015.

total number of pages published annually and quarterly in the *Federal Register*, the United States (US) government's daily journal of bureaucratic activity, including proposed and final regulations. Dawson and Seater (2013) use pages published annually in the *Code of Federal Regulations* (CFR), which contains the stock of final regulations. Crews (2011) counts both the annual number of final regulations published in the *Federal Register* and the annual number of *Federal Register* pages devoted to final regulations.

We advance these researchers' efforts in two ways. First, we provide a novel measure that quantifies regulatory demands by analyzing CFR text.[2] Second, we devise a measure, based on the analysis of regulatory text, for assessing the applicability of different divisions of text within the regulatory code to each of the industries that comprise the US economy, classified according to the two, three, and four-digit levels of the North American Industry Classification System (NAICS).[3] The result is RegData.[4] RegData is the first panel of federal regulation for the US annually for the years 1997–2012 that permits within-industry and between-industry econometric analyses of the causes and effects of federal regulations.

The Great Recession of 2008 has led to much controversy over the role of regulation in avoiding future crises. Some demand liberalization, viewing regulation through the lens of public choice theory (Stigler 1971). Others call for expanding regulation, especially in the financial sector, underlain by a Pigouvian trust in policymakers' ability to rectify rampant market failures (Pigou 1938). RegData can play an important role in helping to resolve this debate. The approach used to assemble RegData can facilitate empirical evaluation of the effectiveness of long-term regulatory trends. For example, how has the growth of environmental regulatory restrictions – which, incidentally, have exhibited tremendous growth relative to other types of regulatory restrictions – affected environmental outcomes, such as air pollution, and the human health outcomes that could be associated with them? At the same time, researchers can use RegData to assess the effects of these same regulatory trends on industry-specific performance metrics, such as total factor productivity or value added to gross domestic product (GDP). Furthermore, RegData permits the investigation of previously unanswerable questions related to the accumulation of regulation: as the stock of regulation has accumulated over the years, has the sheer quantity of regulation led to unanticipated and unintended consequences, driven by the interaction, duplication, or complexity of regulations?[5]

This paper proceeds as follows. In section 2, we explain the methods used to construct the database and provide some simple descriptive statistics. Section 3 offers closing remarks. All original data referred to in this paper are available to the public at http://www.regdata.org/, where several online appendices referred to in this paper can also be found. Appendix A contains supplemental tables and figures. Appendix B contains more details about the methods, and Appendix C explains how to use the data files made available at the website. Appendix D contains validation exercises. Appendix E delves into some of the database's more interesting implications.

2. Data and methods

The CFR is published annually and contains all regulations issued at the federal level in the US. The CFR is divided into 50 titles, each of which corresponds to a broad subject area. Each title is nominally divided into parts that cover specific regulatory areas within the broad subject area given by the title. Each title is also physically divided into volumes to permit publication in conveniently sized bindings. The relationship between parts and volumes is somewhat arbitrary and is subject to revision each year. RegData offers data at various levels of granularity, ranging from very granular (paragraph-level analysis) to very broad (title-level analysis) for the years 1997–2012. Table 1 describes the division scheme generally used in the CFR and reflected in RegData and gives some basic statistics on

Table 1 *Code of Federal Regulations* organization

CFR division	Typical contents	Mean annual observations	Mean word count
Title	Broad subject area of regulations	48	1,200,000
Chapter	Rules of an individual agency	410	150,000
Subchapter	Rules of a sub-agency	n/a	n/a
Part	Rules on a single program or function	8,100	7,500
Subpart	Rules on a particular aspect of a single program or function	n/a	n/a
Section	One provision of a program or function	190,000	310
Paragraph	Detailed requirement(s) related to the provision	1,600,000	39

Note: All numbers are rounded to two significant figures. CFR, code of federal regulations.

observations and word counts at each level. In Appendix A, Table A1 describes all titles used in the CFR in these years alongside more summary statistics on observations, word counts, and bytes.

No divisions of the CFR correspond to individual industries in a self-contained way. Thus, for example, despite the existence of a title called "Shipping" (Title 46), the owner of a ship may need to pay attention to regulations in Title 33 (Navigation and Navigable Waters) and in Title 49 (Transportation). There is no definitive mapping between industries and titles, parts, sections, or other divisions of the CFR based purely on division name.

The CFR is based on a complementary publication called the *Federal Register*. The *Federal Register* is the government's official daily publication of rules, proposed rules, and notices of federal agencies and organizations, as well as executive orders and other presidential documents. Loosely speaking, the *Federal Register* corresponds to the flow of regulations and the CFR corresponds to the stock. We focus on the CFR principally because it contains the legally binding rules, and only the rules. By contrast, the *Federal Register* may measure bureaucratic activity more than regulatory growth. For each final regulation published in the *Federal Register*, pages of preamble text explaining the regulation, economic analyses of the regulation, a Paperwork Reduction Act analysis, and a multitude of other obligatory pages may also exist that, while related to the regulation, do not directly affect economic agents. Furthermore, the *Federal Register* contains notices of proposed rulemaking and advanced notices of proposed rulemaking – documents that explain regulatory agencies' plans but are not binding regulations.

Beyond this, the *Federal Register* contains a large number of non-regulatory pages, including notices of public meetings, announcements of legal settlements, administrative notices and waivers, corrections, presidential statements, and, on occasion, hundreds of blank pages. In short, the *Federal Register* is, at best, a noisy measure of regulation and, at worst, a biased measure because the number of pages associated with individual rulemaking has increased over time as acts of Congress or executive orders have required more analyses.[6]

Another significant advantage of the CFR over the *Federal Register* is that it allows for decreases in regulatory burden. Various titles decrease in length at various points in time, perhaps reflecting some degree of deregulation. Using simple measures based on the *Federal Register* restricts measures of the flow of regulations to always equal zero or greater (as it is not possible to have negative numbers of pages or rulemaking), even when the precise content of the *Federal Register* might reflect deregulation.

2.1. Simple methods for quantifying aggregate regulations

A number of researchers have introduced simple methods for quantifying regulations (Coglianese 2002; Mulligan & Shleifer 2005; Dawson & Seater 2013; Crews 2011; Coffey *et al.* 2012). The first

method is to collect page-count data from either the *Federal Register* or the CFR. These page counts provide an excellent departure point and have furnished several insightful regulation studies.

Page-count data are subject to the criticism that not all pages are equal. A page could be of enormous or trivial consequence to the economy. Also, page-formatting guidelines may change over time. Further, some CFR titles (e.g. Title 50: Wildlife and Fisheries) use maps, schematic diagrams, or a disproportionate number of tables rather than dense text. Thus, the complexity and impact of the associated regulations are potentially not well captured or comparable across titles by using raw page counts of the CFR. A similar critique is applicable to counting the number of final rules published on an annual basis.

Mulligan and Shleifer (2005) use file-size data from the statutes of 37 US states, allowing them to overcome the possibility of differences in formatting. We gather file-size data but omit it from this paper for reasons of parsimony; those interested should contact the authors. However, word count data at every CFR division is available. Words are also unaffected by large graphics that affect file sizes. Moreover, we devise and gather two additional novel measures.

Regardless of the method used, a major limitation of previous approaches is that the data show only longitudinal (time-series) variation in total regulation. Casual observation suggests that some industries are more heavily regulated than others. If this is indeed the case, then quantifying the cross-sectional variation will surely enhance our understanding of regulation. We attempt this quantification below.

2.2. Quantifying regulations using text analysis

Regulations affect economic agents primarily through constraining or expanding their legal choice sets. Regulatory texts typically use a relatively standard suite of verbs and adjectives to indicate a binding constraint, such as "shall," "must," and "prohibited." This observation motivated us to search the CFR for keywords that are likely to indicate binding constraints. As a departure point, we search for five strings that are likely to limit choice sets: "shall," "must," "may not," "prohibited," and "required." We refer to this set of five strings as "restrictions."

We use computer programs to count the occurrences of each of these five strings in each division of the CFR 1997–2012, with the exception of Title 35.[7] Title 2 addresses government grants and procurement procedures and starts in 2005. Title 6 covers domestic security and starts in 2004 (Department of Homeland Security).

One of our new measures of regulations – restrictions – is the total number of restrictions in a division of the CFR. To be clear, measurements of restrictions are subject to the same criticism as measurements of pages – just as one page may not be equal to another page, one restriction may carry more consequence than another. However, one advantage of using restrictions, and, more generally, text analysis, over page counts, is that the user can select different levels of granularity. RegData 2.0 offers this measure at four levels of divisions given in table 1 (title, chapter, part, and paragraph), with "title" being the broadest and "paragraph" the narrowest. Restrictions are measured by the total number of (case insensitive) occurrences in a CFR division of the five aforementioned restricting strings. Table 2 provides summary statistics of the variable restrictions for each CFR title over the 16-year period. Figure 1 depicts restrictions over this time period for the four CFR titles with the greatest number of restrictions, on average, of any of the 50 titles.

Figure 2 shows the total restrictions published each year in the CFR. The persistent growth seems to confirm the popular notion that federal regulation has grown regardless of the political party in charge of the executive branch. Total restrictions increased from 830,000 in 1997 to 1 million in 2012.

Table 2 Summary statistics for restrictions in *Code of Federal Regulations* titles, 1997–2012

Title	subject	No. of years	Mean	SD	Min.	Max.
1	General Provisions	16	400	13	400	450
2	Grants and Agreements	8	1,400	490	340	1,900
3	The President	16	770	270	420	1,400
4	Accounts	16	790	120	670	980
5	Administrative Personnel	16	12,000	830	11,000	13,000
6	Domestic Security	9	1,100	270	750	1,400
7	Agriculture	16	71,000	4,600	61,000	80,000
8	Aliens and Nationality	16	8,800	1,800	6,000	11,000
9	Animals and Animal Products	16	18,000	590	17,000	19,000
10	Energy	16	24,000	2,200	21,000	28,000
11	Federal Elections	16	3,200	390	2,700	3,700
12	Banks and Banking	16	27,000	6,300	19,000	47,000
13	Business Credit and Assistance	16	4,000	670	2,900	5,000
14	Aeronautics and Space	16	30,000	4,000	24,000	35,000
15	Commerce and Foreign Trade	16	9,300	410	8,500	9,800
16	Commercial Practices	16	9,900	610	9,000	11,000
17	Commodity and Securities Exchanges	16	18,000	2,600	9,300	21,000
18	Conservation of Power and Water Resources	16	11,000	1,100	9,800	12,000
19	Customs Duties	16	12,000	570	11,000	13,000
20	Employees' Benefits	16	17,000	3,200	5,800	19,000
21	Food and Drugs	16	21,000	1,300	19,000	23,000
22	Foreign Relations	16	11,000	1,100	7,100	12,000
23	Highways	16	3,900	180	3,600	4,200
24	Housing and Urban Development	16	23,000	920	22,000	25,000
25	Indians	16	10,000	980	8,200	11,000
26	Internal Revenue	16	60,000	3,100	54,000	64,000
27	Alcohol, Tobacco Products, and Firearms	16	11,000	130	11,000	11,000
28	Judicial Administration	16	10,000	910	8,800	12,000
29	Labor	16	48,000	1,900	43,000	51,000
30	Mineral Resources	16	22,000	700	21,000	23,000
31	Money and Finance: Treasury	16	8,200	1,000	6,600	9,400
32	National Defense	16	22,000	1,500	18,000	24,000
33	Navigation and Navigable Waters	16	15,000	1,600	11,000	17,000
34	Education	16	10,000	560	9,300	11,000
35	Panama Canal	3	1,300	800	430	1,800
36	Parks, Forests, and Public Property	16	12,000	480	10,000	12,000
37	Patents, Trademarks, and Copyrights	16	4,800	840	3,600	6,100
38	Pensions, Bonuses, and Veterans' Relief	16	8,600	820	7,500	10,000
39	Postal Service	16	3,400	110	3,200	3,500
40	Protection of Environment	16	130,000	25,000	74,000	160,000
41	Public Contracts and Property Management	16	9,300	280	8,900	9,900
42	Public Health	16	15,000	2,600	11,000	20,000
43	Public Lands: Interior	16	14,000	810	13,000	17,000
44	Emergency Management and Assistance	16	4,000	210	3,800	4,500

(*Continues*)

Table 2. (Continued)

Title	subject	No. of years	Mean	SD	Min.	Max.
45	Public Welfare	16	17,000	1,300	13,000	19,000
46	Shipping	16	35,000	250	34,000	35,000
47	Telecommunication	16	25,000	1,200	22,000	27,000
48	Federal Acquisition Regulations System	16	29,000	1,600	25,000	31,000
49	Transportation	16	42,000	5,600	34,000	51,000
50	Wildlife and Fisheries	16	16,000	4,900	10,000	24,000

Note: All numbers and are rounded to two significant figures. SD, standard deviation.

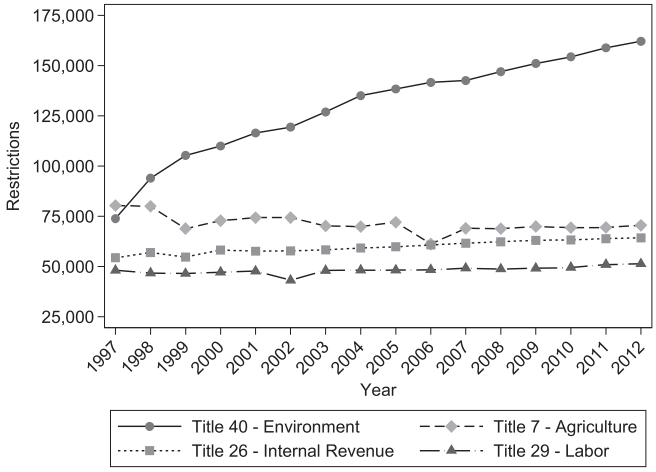

Figure 1 *Code of Federal Regulations* restrictions, 1997–2012, for four titles.
Note: these are the titles with the greatest number of restrictions on average of any of the 50 titles.

Both figures 1 and 2 show total restrictions, rather than adjusting restrictions for total word counts (an adjustment we do perform for other metrics later in the paper). There are two reasons we do this. First, total restrictions are an interesting measure of regulatory accumulation, regardless of how many words accompany the restrictions. Second, dividing restrictions by word counts does not change the relative placement of the series shown in figure 1, because word counts and restriction counts are highly correlated, averaging about 0.93 at the part level. This correlation is depicted in figure A2 in Appendix A.

2.3. Quantifying the applicability of regulations to specific industries using text analysis

The NAICS classifies industries into mutually exclusive and exhaustive bins that are assigned numbers. There are five versions of the NAICS, depending on the granularity of the classification: two-digit (coarsest) to six-digit (finest).[8] Table 3 illustrates the gradation with an example. In Appendix A, table A2 shows the two-digit classification, and table A3 shows the three-digit classification.[9]

We developed a method that uses these NAICS industry descriptions to measure the applicability of the regulatory text contained in a unit of the CFR to a specific industry. As with

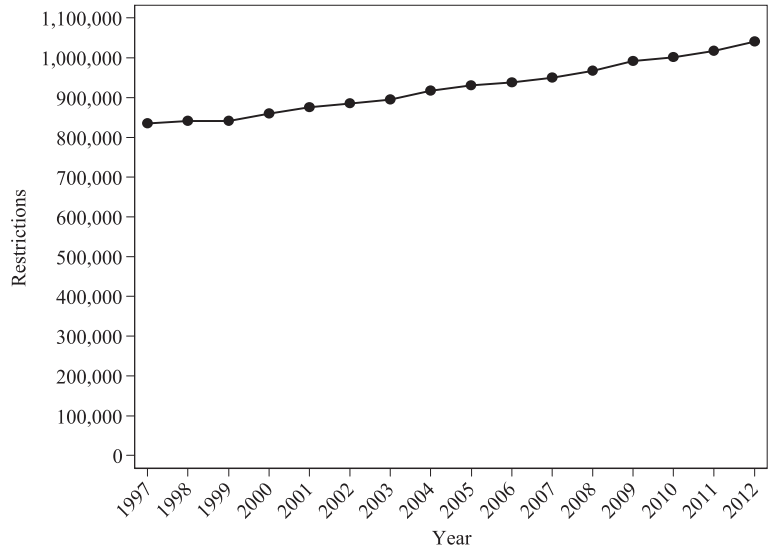

Figure 2 Total *Code of Federal Regulations* restrictions, 1997–2012.
Note: This graph covers all five restrictions across all titles.

Table 3 An example of North American industry classification system gradation

Digits	Industry number and description
2	31 Manufacturing
3	311 Food Manufacturing
4	3112 Grain and Oilseed Milling
5	31121 Flour Milling and Malt Manufacturing
6	311211 Flour Milling
6	311212 Rice Milling
6	311213 Malt Manufacturing
5	31122 Starch and Vegetable Fats and Oils Manufacturing
6	311221 Wet Corn Milling
6	311222 Soybean Processing
6	311223 Other Oilseed Processing
6	311225 Fats and Oils Refining and Blending

restrictions, we measured applicability of regulatory text to industries at four CFR levels of granularity: title, chapter, part, and paragraph.

2.3.1. Main method

For each NAICS code, we created a collection of strings based on combinations and transformations of words in the code's description. We denote this collection the "search strings." Thus, for example, code 52 is "Finance and Insurance," and the search strings included strings such as "finance," "insurance," and "insurer."

We created these search strings using rules we devised – according to our subjective judgment – to transform NAICS descriptions into multiple search strings. We fully explain these rules in

Appendix B. The website lists all search strings created with this approach, along with the rule used to create each string, thereby permitting researchers to modify our approach.

After forming each code's search strings, we counted the occurrences of each search string for each two, three, and four-digit industry in each division of the 1997–2012 CFR.[10] The resulting data give industry-specific measures of relevance – measures of the extent to which a CFR division in a given year relates to specific industries as defined in the corresponding NAICS classifications. Our measure of industry relevance is deflated by the number of words in the same CFR unit (see Appendix B). As with many aspects of this database, users are also able to modify or remove this deflation.

As an illustration, figure 3 shows the relevance of one particular CFR title, Title 12: Banks and Banking, to all the two-digit NAICS industries. The bars show the number of occurrences of the industry-specific search strings found in Title 12 in the year 2012, divided by Title 12's word count. Predictably, Title 12 appears most relevant to the "Finance and Insurance" industry (code 52) and "Real Estate and Rental and Leasing" industry (code 53).

Our data also permit the user to examine the relevance of other CFR divisions, such as parts or chapters, as well as the relevance of regulations issued by a department or agency. (In fact, chapters and agencies are nearly equivalent, because chapters can typically be easily mapped to a specific regulatory department or agency. In contrast, parts often correspond to particular regulatory programs of interest to researchers and policymakers alike.) Figure 4, for example, shows the relevance of agency regulations to the animal slaughtering and processing industry. The most relevant agency by far is the Food Safety and Inspection Service in the Department of Agriculture, followed by the Food and Drug Administration.

There are a variety of uses for the data. For example, if one wants to compare Title 40's relevance to "Chemical Manufacturing" (code 325) with Title 40's relevance to "Motor Vehicle and

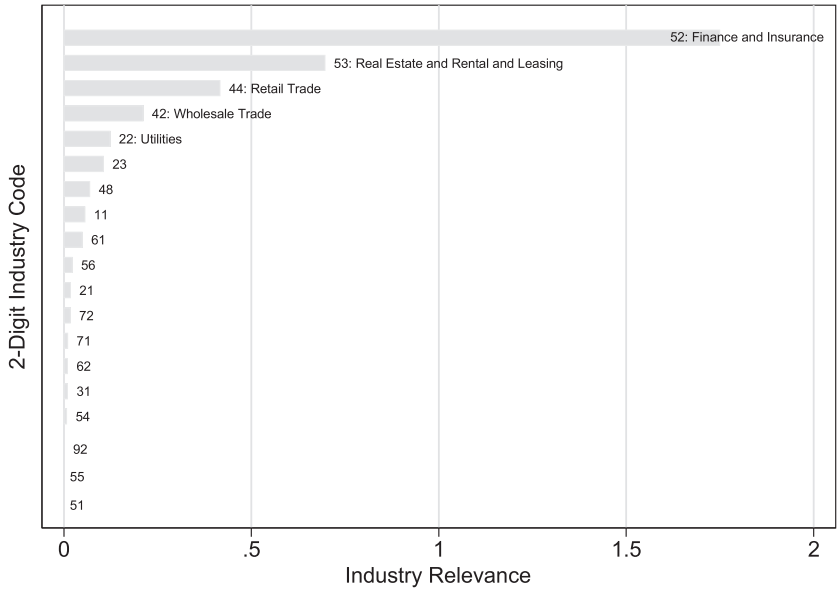

Figure 3 Relevance of *Code of Federal Regulations* Title 12, "Banks and Banking," to all two-digit North American Industry Classification System Industries.
Note: data are from 2012. The top five industries are labeled; see table A2 for a list of all two-digit North American Industry Classification System industries and codes.

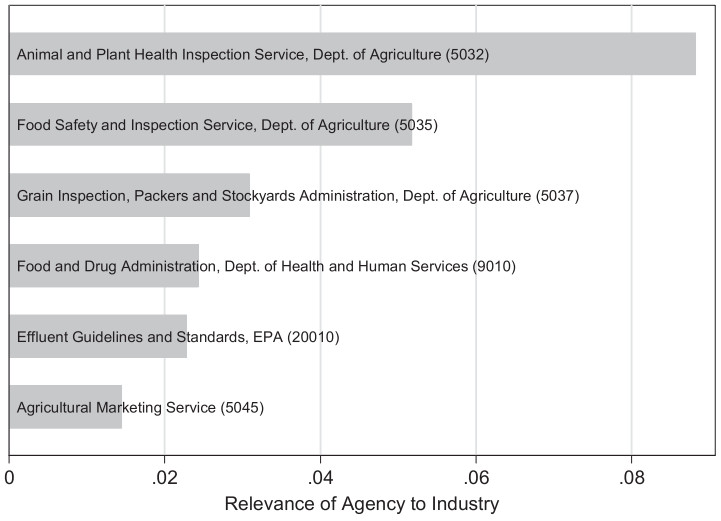

Figure 4 Relevance of US Federal Agencies to North American Industry Classification System Industry 3116, "Animal Slaughtering and Processing".
Note: 13 other agencies are relevant but have industry relevance values of less than 0.01. Data are from 2012. EPA, Environmental Protection Agency.

Parts Dealers" (code 441) for the year 2000, one can compare the hits on the strings from code 325 to those from code 441. Another method is to include parent codes additively – that is, to compare the hits on the strings from code 32 plus the hits on the strings from code 325 against the hits on the strings from code 44 plus the hits on the strings from code 441. We explain some different methods in Appendix B.

This method of mapping regulatory text to specific industries has some limitations. We discuss these limitations and some solutions to them in Appendix B. We show that these limitations do not invalidate our method of quantifying regulation through validation exercises described in Appendix D.

2.4. Combining the two databases to create a panel

We provide an example of how a user could construct an industry regulation index that combines measures of restrictiveness of specific units of the CFR with the relevance of those units to specific industries. We use the term "industry regulation index" in a fairly loose sense. Depending on the specific index constructed, it may be more accurately described as an industry restriction index (which would be appropriate in the example below).

Let i denote industry and y denote year; let I denote the set of industries and Y the set of years; let $S = \{I \times Y\}$ (Eqn. 1). (In our case, I depends on which granularity of the NAICS is used, while Y covers the period 1997–2012.) Title, part, and agency-specific measures of regulation – for example, restrictions – can be combined with our data on the relevance of CFR units to specific industries to create a panel dataset indicating industry-specific regulation from 1997 through 2012. For an example of a part-specific panel, let Rpy be the number of regulations in part p in year y, based on one of our two measures of regulation (word count or restrictions). Assuming that the weight a restriction receives in total restrictions does not depend on the part, $R_y = \sum_p R_{py}$ (Eqn. 2) is a measure of the total number of restrictions in year y.

Let *apyi* be the applicability of the regulations in part p in year y to industry i taken from the industry relevance data described above. We want to construct a new index *rpyi* measuring the restrictions for industry i in part p in year y. The relationship will be of the form:

$$r_{pyi} = f\left(a_{pyi}, R_{py}\right), \tag{3}$$

where f is increasing in both elements and the cross-partial is also positive. The simplest possibility is:

$$f\left(a_{pyi}, R_{py}\right) = a_{pyi}R_{py}; \tag{4}$$

alternatively, one could use a function of the form:

$$f\left(a_{pyi}, R_{py}\right) = D\left(a_{pyi}\right)R_{py}, \tag{5}$$

where D is a dummy variable that takes the value 1 when *apyi* is above a threshold. Finally, assuming equal part weighting as above:

$$r_{yi} = \sum_{p} r_{pyi} \tag{6}$$

will be a measure of the restrictions on industry i in year y. We provide:

$$r_{pyi} = a_{pyi}R_{py} \tag{7}$$

as the default industry regulation index. However, as above, to promote fruitful experimentation, we make the entire dataset available, permitting anyone to construct different industry-specific regulatory indexes using different weightings or combinations of *apyi* and *Rpy*.

As an example, using our default, part-level method, figure 5 shows the growth path of an industry regulation index (where the base year is 1997) for a selection of four-digit industries. According to the part-level industry regulation index, the average four-digit industry has experienced a 28 percent increase in restrictions since 1997. To give a broader look at the data,

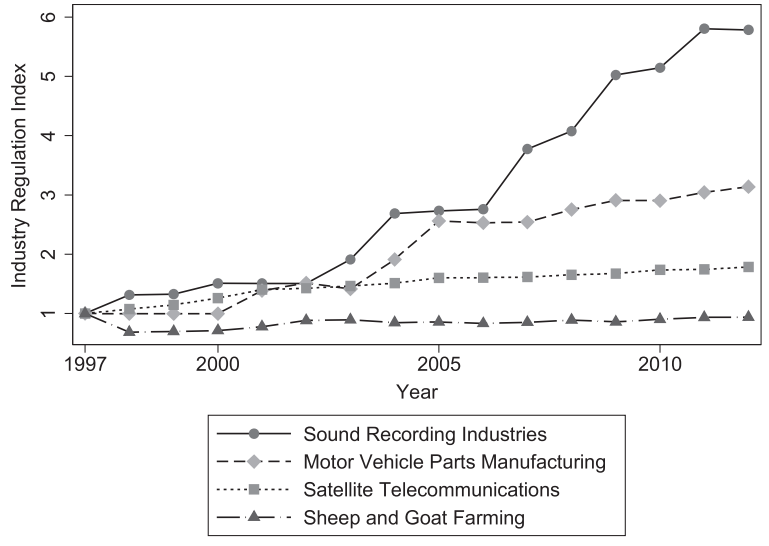

Figure 5 Industry Regulation Index for a selection of North American Industry Classification System four-digit industries.
Note: The base year is 1997; the index is calculated at the part level.

table 4 reports summary statistics for our default industry regulation level for a selection of two, three, and four-digit level industries.

RegData also permits the user to calculate how regulated an industry is by a specific department, agency, or set of departments and agencies. Figure 6 shows the agency-level industry regulation index series for three different industries. The agencies included are classified as creating "workplace" regulations in the Regulators' Budget (Dudley & Warren 2012; see the note below the figure for the list). All other metrics – word counts, restriction counts, industry search terms, and industry relevance – are also available at the department and agency levels.

The default version of RegData calculates industry regulation by linking relevance to restrictions at the part level and aggregating. As remarked above, one can produce alternative measures of regulation by linking at the agency, chapter, or title level. In fact, the correlation between these alternatives and our default part-level measure is so high (above 0.95) that choosing one over another has a barely discernible effect.

North American Industry Classification System categories are extensively applied to a wide variety of economic data. For example, the Bureau of Economic Analysis provides GDP value-added data by industry according to two and three-digit NAICS codes. Many opportunities exist to merge our database with other data to explore the causes and outcomes of regulations. We conduct an exploratory analysis in Appendix E.

3. Closing remarks

This paper introduces RegData. The version of RegData described here is the second iteration of an ongoing research effort that will later include further refinements of this approach to measuring regulation quantity, as well as the development of other metrics of law and regulation.

RegData allows users to combine two datasets to create a panel database that annually quantifies federal regulations by industry for all US industries and regulations from 1997 –2012. The

Table 4 Summary statistics for the regulation level for a selection of North American industry classification system two, three, and four-digit industries

NAICS	Description	Mean	SD	Min.	Max.
22	Utilities	180,000	16,000	140,000	210,000
53	Real Estate and Rental and Leasing	290,000	22,000	230,000	330,000
56	Admin. and Supp. and Waste Mgmt.	650,000	52,000	570,000	740,000
62	Health Care and Social Assistance	35,000	7,100	2,000	43,000
42	Wholesale Trade	18,000	1,200	15,000	19,000
112	Animal Production	97,000	9,300	85,000	110,000
313	Textile Mills	15,000	1,400	11,000	17,000
611	Educational Services	92,000	10,000	73,000	110,000
324	Petroleum and Coal Products Mfg.	141,000	12,000	110,000	160,000
312	Beverage and Tobacco Product Mfg.	88,000	19,000	65,000	110,000
3358	Tobacco Manufacturing	79,000	20,000	54,000	100,000
1123	Poultry and Egg Production	43,000	2,300	40,000	48,000
4862	Pipeline Transportation of Natural Gas	40,000	7,000	24,000	50,000
6222	Psychiatric and Substance Abuse Hospitals	29,000	1,700	27,000	33,000
4511	Sporting Goods, Hobby, and Musical Stores	710	95	580	830

Note: The regulation index was constructed using the default method (that is, at the part level). Numbers are reported to two significant figures. SD, standard deviation.

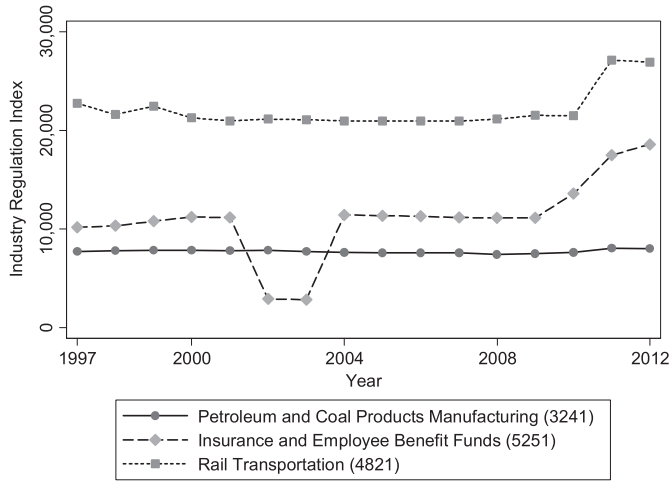

Figure 6 Industry Regulation Index for a selection of North American Industry Classification System four-digit industries by workplace regulatory agencies.
Note: these agencies include six agencies within the Department of Labor and five independent boards or commissions. Department of Labor: Employment Standards Administration, Office of Workers Compensation Programs, Office of Federal Contract Compliance Programs, Employee Benefits Security Administration, Mine Safety and Health Administration, and Occupational Safety and Health Administration. Independent boards and commissions: Architectural and Transportation Barriers Compliance Board, Equal Employment Opportunity Commission, National Labor Relations Board, and Occupational Safety and Health Review Commission.

first database contains three metrics of regulation quantity: CFR page-count data, digitized CFR file-size data, and a novel measure (restrictions) that counts the number of legally binding words contained in regulatory text.

In the second dataset, we offer the first measure of the relevance of units of the CFR to industries in the US. This measure was created by searching each unit of the CFR for text strings that describe each industry in the US, as defined by the two through four-digit codes of the NAICS, and summing the number of hits in each unit and each year. We based the descriptions of industries on two through four-digit NAICS industry descriptions, in part to allow RegData to be combined with data on specific outcomes that may be affected by or determinants of regulation, such as industrial performance, safety data, or environmental outcomes. Many publicly available datasets are also based on the NAICS, such as employment levels or value added to GDP by industry, thus, lending compatibility with RegData.

RegData offers users numerous choices that we hope will permit maximum experimentation and minimize subjectivity. Users can decide how to combine the databases, which measure the quantity of regulation to use, and whether to omit or include specific strings from the constraints database or from the industry search strings.

In the Appendix, we have employed two validation strategies: a bottom-up approach that appeals to the steps in the database's construction; and two top-down approaches: one that demonstrates the consistency of (a modified version of) the database with stylized facts about regulation, namely notable episodes of deregulation, and one that demonstrates some consistency with a three-value, cross-sectional measure of regulation developed by Coates (2012). Our efforts to validate RegData are hampered by its novelty (no other panel measures of regulation exist) and by economists' lack of certainty regarding the relationship between regulation and other economic variables. To demonstrate its value, we have also conducted a rudimentary econometric

exploration of the relationship between RegData regulation and various aggregate economic time series thought to be causally linked to regulation. Our rudimentary analysis shows, for example, that the relationships between RegData regulation and employment and regulation and total labor compensation warrant more complete research, which we hope to facilitate.

This iteration of RegData is freely available to the public with the goal of facilitating regulatory research, and we hope to refine RegData in several ways and release those refined versions in the future. First, our novel measure of regulation – restrictions – treats all occurrences of a binding constraint equally. We plan to develop more nuanced measures of constraints that take into account the context of the word. For example, some binding constraints may be followed or prefaced by an exception, or may only apply in special circumstances.

Second, we plan to develop other measures of regulatory text that will serve as proxies for regulatory quality. These measures will serve as companion databases that supplement RegData. We intend to start this process by creating rules based on the plain language guidance that federal regulators are directed to use when writing regulatory text. Despite this guidance, some parts of the CFR do not hew to the precepts of plain language. As a starting point, we will develop a plain language score, which can then be combined with industry-specific outcomes to test whether the quality of regulatory writing affects economic outcomes.

Acknowledgments

The authors thank Steve Balla, Bill Beach, John Coates IV, Bentley Coffey, Tyler Cowen, Susan Dudley, Jerry Ellig, Peter French, Patrick Fuchs, Don King, Paul Large, Randy Lutter, Carlos Ramirez, Jo Strang, Thomas Stratmann, Richard Williams, Dima Yazigi, and participants at workshops and seminars at the George Washington University's Regulatory Studies Center, the American Law and Economics Association Annual Meeting, the Society for Benefit-Cost Analysis Annual Meeting, Resources for the Future, the Southern Economics Association Annual Meeting, the Western Economics Association Annual Meeting, the Congressional Budget Office Microeconomics Group, and the US Department of Transportation for useful comments. The authors also thank Tim McLaughlin for providing immense help in creating the text analysis programs. Dima Yazji Shamoun, Schinria Islam, Stuart Sanders, and Sam Toolan provided excellent research assistance. Brian Deignan, Jacob Feldman, Andrew McKeown, and Kellen Rosenfelder all assisted in quality assurance.

NOTES

1 We focus on those studies that have attempted to quantify broad swaths of regulation rather than regulation focused on a particular industry or issue. Other studies have used measures of specific types of regulations or proxies of regulation across countries; these studies include Djankov *et al.* (2002), which employs a business entry regulation index, and Botero *et al.* (2004), which creates indexes that measure the extent of worker protection laws and regulations. Some other papers that apply these measures include Aghion *et al.* (2010) and Glaeser and Shleifer (2003).

2 See Gentzkow and Shapiro (2010) and Baker *et al.* (2013) for other examples of the use of text analysis in economics.

3 We anticipate completing and releasing five- and six-digit industry data in the near future and will release those data on our website: http://www.regulationdata.org

4 RegData was first introduced in a working paper published in July 2012; see http://ssrn.com/abstract=2099814. The version of RegData we introduce in this paper contains several improvements over the July 2012 version. Improvements include data for years 2011 and 2012; data for NAICS four-digit industries; search-term weightings derived from Google's Ngram database; scalable granularity for CFR search results, ranging from CFR title-level results to CFR paragraph-level results; and regulatory agency and sub-agency-specific search results.

5 Mandel and Carew (2013) eloquently discuss some potential problems associated with regulatory accumulation. McLaughlin and Williams (2014) and McLaughlin and Greene (2014) review several studies that indicate regulatory accumulation, in and of itself, may be problematic.
6 Crews (2011) somewhat mitigates this drawback by focusing only on pages devoted to final rules (McLaughlin 2011).
7 Title 35 contained regulations relevant to the Panama Canal, which was ceded to Panama in 1999; the title was terminated in 2004.
8 See http://www.census.gov/eos/www/naics/ for more information.
9 See the NAICS homepage for the larger tables corresponding to four, five, and six-digit classifications.
10 We anticipate completing and releasing five and six-digit search results in a future update of the database. However, initial indications suggest that the five and six-digit versions of RegData suffer from some linguistic drawbacks compared with the coarser granularities, and, thus, at this point we endorse the three and four-digit versions over the remainder. The problem with the finer granularities is an abundance of sporadically distributed technical vocabulary that requires a more sophisticated string-generation procedure. An implausibly large number of industries report "zero" regulation according to RegData techniques when measuring at the six-digit level, for example, "Noncurrent-carrying Wiring Device Manufacturing" (335932), and this result biases the dataset when there are other six-digit industries, such as "Cheese Manufacturing" (311513), that attract a reasonable number of hits.

References

Aghion P, Algan Y, Cahuc P, Shleifer A (2010) Regulation and Distrust. *Quarterly Journal of Economics* 125, 1015–1049.

Al-Ubaydli O, McLaughlin PA (2012) The Industry-Specific Regulatory Constraint Database (IRCD): A Numerical Database on Industry-Specific Regulations for All U.S. Industries and Federal Regulations, 1997–2010. Mercatus Center Working Paper. George Mason University, Arlington, VA.

Baker SR, Blook N, Davis SJ (2013) Measuring Economic Policy Uncertainty. Chicago Booth Research Paper, No. 13-02. Stanford University, Stanford CA.

Botero JC, Djankov S, La Porta R, Lopez-de-Silanes F, Shleifer A (2004) The Regulation of Labor. *Quarterly Journal of Economics* 119, 1339–1382.

Coates JCIV (2012) Corporate Politics, Governance, and Value before and after Citizens United. *Journal of Empirical Legal Studies* 9, 657–696.

Coffey B, McLaughlin PA, Tollison RD (2012) Regulators and Redskins. *Public Choice* 153, 191–204.

Coglianese C (2002) Empirical Analysis and Administrative Law. *University of Illinois Law Review* 4, 1111–1138.

Crews CW (2011) Ten Thousand Commandments: An Annual Snapshot of the Federal Regulatory State. Competitive Enterprise Institute, Washington, DC.

Dawson JW, Seater JJ (2013) Federal Regulation and Aggregate Economic Growth. *Journal of Economic Growth* 18, 137–177.

Djankov S, La Porta R, Lopez-de-Silanes F, Shleifer A (2002) The Regulation of Entry. *Quarterly Journal of Economics* 117, 1–37.

Dudley S, Warren M (2012) Growth in Regulators' Budget Slowed by Fiscal Stalemate: An Analysis of the U.S. Budget for Fiscal Years 2012 and 2013. Regulators' Budget Report 34. Available online: http://wc.wustl.edu/files/wc/imce/2013regreport.pdf

Gentzkow M, Shapiro J (2010) What Drives Media Slant? Evidence from US Daily Newspapers. *Econometrica* 78, 35–71.

Glaeser E, Shleifer A (2003) The Rise of the Regulatory State. *Journal of Economic Literature* 41, 401–425.

Greenstone M (2002) The Impacts of Environmental Regulations on Industrial Activity: Evidence from the 1970 and 1977 Clean Air Act Amendments and the Census of Manufactures. *Journal of Political Economy* 110, 1175–1219.

Mandel M, Carew DG (2013) Regulatory Improvement Commission: A Politically-viable Approach to U.S. Regulatory Reform. Progressive Policy Institute Policy Memo. [Last accessed 4 April 2014.] Available online: http://www.progressivepolicy.org/wp-content/uploads/2013/05/05.2013-Mandel-Carew_Regulatory-Improvement-Commission_A-Politically-Viable-Approach-to-US-Regulatory-Reform.pdf

McChesney FS (1987) Rent Extraction and Rent Creation in the Economic Theory of Regulation. *Journal of Legal Studies* 16, 101–118.

McLaughlin PA (2011) The Consequences of Midnight Regulations and Other Surges in Regulatory Activity. *Public Choice* 147, 395–412.

McLaughlin P, Greene R (2014) The Unintended Consequences of Federal Regulatory Accumulation. Discussion Paper, Mercatus Center Economic Perspectives Series, George Mason University, Arlington, VA.

McLaughlin P, Williams R (2014) The Consequences of Regulatory Accumulation and a Proposed Solution. Mercatus Center Working Paper, No. 14-03, George Mason University, Arlington, VA.

Mulligan C, Shleifer A (2005) The Extent of the Market and the Supply of Regulation. *Quarterly Journal of Economics* 120, 1445–1473.

Peltzman S (1975) The Effects of Automobile Safety Regulation. *Journal of Political Economy* 83, 677–725.

Pigou A (1938) *The Economics of Welfare*. Macmillan, London.

Stigler G (1971) The Theory of Economic Regulation. *Bell Journal of Economics and Management Science* 2, 3–21.

United States Census Bureau. North American Industry Classification System. [Last accessed 13 January 2012]. http://www.census.gov/eos/www/naics/

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website.

