No. 12-20
October 2012

WORKING PAPER

THE INDUSTRY-SPECIFIC REGULATORY CONSTRAINT DATABASE (IRCD): A Numerical Database on Industry-specific Regulations for All U.S. Industries and Federal Regulations, 1997-2010

By Omar Al-Ubaydli and Patrick A. McLaughlin

MERCATUS CENTER
George Mason University

The Industry-specific Regulatory Constraint Database (IRCD): A Numerical Database on Industry-Specific Regulations for All U.S. Industries and Federal Regulations, 1997–2010

July 2012

Omar Al-Ubaydli and Patrick A. McLaughlin[1]

Abstract

We introduce the Industry-specific Regulatory Constraint Database (IRCD). IRCD annually quantifies federal regulations by industry for all U.S. industries and regulations from 1997 to 2010. The quantification of federal regulations at the industry level for all industries is without precedent. 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 across time. The database parses industries at the two-digit and three-digit North American Industry Classification System (NAICS) levels. We created this database by 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 groups of regulations to different industries.

Key words: regulation; industry

JEL codes: K2; L5; N4; Y1

[1] The authors thank Steve Balla, Bentley Coffey, Susan Dudley, Jerry Ellig, Patrick Fuchs, Don King, Paul Large, Carlos Ramirez, Richard Williams, participants at the George Washington University's Regulatory Studies Center research seminar, and participants at the U.S. Department of Transportation FRA brown bag lunch for useful comments. The authors also thank Tim McLaughlin for providing immense help in creating the text analysis programs. Al-Ubaydli: Bahrain Center for Strategic, International and Energy Studies, and Department of Economics at Mercatus Center, George Mason University; email: omar@omar.ec. McLaughlin (corresponding author): Senior Research Fellow, Mercatus Center at George Mason University; email: pmclaug3@gmu.edu.

1. Introduction

Scholars have been analyzing the causes and consequences of government regulation for decades, leading to a vast and still-growing literature. A principle reason for the popularity of such inquiries is that regulations are an invaluable policy tool for addressing market failure.[2] However, the complexity of the political process means that regulations may not always be virtuously conceived,[3] and the intricacy of the modern economy means that regulations may have adverse unintended consequences.[4]

Studies typically examine (theoretically or empirically) the causal effect of a unique regulation or a small collection of related regulations, such as air quality standards.[5] Compared to the thousands of actual regulations that govern a large economy, the intervention typically 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 the voluminous legal documents that literally specify the regulations. Previous efforts to measure the extent of regulation in the United States have used proxy variables designed to measure the quantity of federal or state regulation created or in effect each year.[6] Mulligan and Shleifer used the sizes, measured in kilobytes, of the digitized versions of state-level statutes as a proxy for real state-level regulation.[7] Coffey et al. used the total number of pages published annually and quarterly in the *Federal Register* (*FR*),[8] the government's daily journal of newly proposed and final regulations. Dawson and Seater used pages published annually in the *Code of Federal Regulations* (*CFR*),[9] which contains the stock of final regulations. Coglianese also used annual *CFR* page counts to informally test whether judicial review has led to a decline in rulemaking.[10] Crews counted both the annual number of final regulations published in the *FR* and the annual number of *FR* pages devoted to final regulations.[11]

[2] A. Pigou, *The Economics of Welfare* (London: Macmillan, 1938).
[3] G. Stigler, "The Theory of Economic Regulation," *Bell Journal of Economics and Management Science* 2, (1971): 3–21; F. S. McChesney, "Rent Extraction and Rent Creation in the Economic Theory of Regulation," *Journal of Legal Studies* 16 (1987): 101–118.
[4] S. Peltzman, "The Effects of Automobile Safety Regulation," *Journal of Political Economy* 83 (1975): 677–725. For a more thorough discussion of the different theories of regulation, see S. Djankov, R. La Porta, F. Lopez-de-Silanes, and A. Shleifer, "The Regulation of Entry," *Quarterly Journal of Economics* 117 (2002): 1–37, which is also a typical study of the consequences of a small group of regulations (namely regulations that constitute entry barriers).
[5] See, e.g., M. Greenstone, "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 (2002): 1175–1219.
[6] We focus on those studies that have attempted to quantify broad swathes 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, including Djankov et al., "The Regulation of Entry," which employs a business entry regulation index, and J. Botero, S. Djankov, R. La Porta, F. Lopez-de-Silanes, and A. Shleifer, "The Regulation of Labor," *Quarterly Journal of Economics* 119 (2004): 1339–1382, which creates indices that measure the extent of worker protection laws and regulations. Some other papers that apply these measures include P. Aghion, Y. Algan, P. Cahuc, and A. Shleifer, "Regulation and Distrust," *Quarterly Journal of Economics* 125 (2010): 1015–1049, and E. Glaeser and A. Shleifer, "The Rise of the Regulatory State," *Journal of Economic Literature* 41 (2003): 401–425.
[7] C. Mulligan and A. Shleifer, "The Extent of the Market and the Supply of Regulation," *Quarterly Journal of Economics* 120, (2005): 1445–1473.
[8] B. Coffey, P. A. McLaughlin, and R. D. Tollison, "Regulators and Redskins," *Public Choice* (published online in March 2011).
[9] J. Dawson and J. Seater, "Federal Regulation and Aggregate Economic Growth" (working paper, 2008).
[10] Cary Coglianese, "Empirical Analysis and Administrative Law," *University of Illinois Law Review* 4 (2002): 1111–1138.
[11] C. W. Crews, *Ten Thousand Commandments: An Annual Snapshot of the Federal Regulatory State* (Washington, DC: Competitive Enterprise Institute, 2011).

We advance these researchers' efforts in two principal ways. First, in addition to providing page-count and file-size data for the *CFR*, we provide a complementary and novel measure that quantifies regulations by analyzing *CFR* text.[12] Second, we devise a measure, based on the analysis of regulatory text, for assessing the applicability of each regulation to each of the industries that comprise the U.S. economy, classified according to the two- and three-digit levels of the North American Industry Classification System (NAICS). The result is the Industry-specific Regulatory Constraint Database (IRCD). IRCD is the first panel of federal regulation for the U.S. annually for the years 1997–2010 that permits within-industry and between-industry econometric analyses of the causes and effects of federal regulations.

A particularly worrying consequence of the Great Recession of 2008 has been the polarization of views on how best to avoid future crises. Nowhere is the extremity of contrary diagnoses more apparent than in the realm of regulation. On the one hand, some demand liberalization, viewing regulation through the lens of public choice theory.[13] On the other hand, others call for expanding regulation, especially in the financial sector, underlain by a Pigouvian trust in policymakers' ability to rectify rampant market failures. In the analysis of regulation,[14] the stakes have risen to an all-time high. We believe our new database could play an important role in resolving the controversy.

This paper explains the methods used in constructing the database and provides some simple descriptive statistics. Appendices A and C contain more details about the methods. All data referred to in this paper are available to the public at www.regulationdata.org, and appendix B explains how to use the data files made available at the website.

2. Data and Methods

The *CFR* is published annually and contains all regulations issued at the federal level.[15] It is divided into 50 titles, each of which corresponds to a broad subject area covered by federal regulation. 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; some volumes contain dozens of parts, while some parts span across multiple volumes. We report data at the title level for the years 1997–2010. Table 1 describes all titles used in the *CFR* in these years.

[12] See M. Gentzkow and J. Shapiro, "What Drives Media Slant? Evidence from U.S. Daily Newspapers," *Econometrica* 78 (2010): 35–71, for other examples of the use of text analysis in economics.
[13] Stigler, "The Theory of Economic Regulation"; J. Buchanan and G. Tullock, *The Calculus of Consent* (Ann Arbor, MI: University of Michigan Press, 1962).
[14] Pigou, *The Economics of Welfare*.
[15] A regulation may be in effect for up to one year prior to actual publication in the *CFR*, but ultimately, all regulations are published in the *CFR*.

Table 1. Titles Used in the *Code of Federal Regulations*, 1997–2010

CFR title	Subject
1	General Provisions
2	Grants and Agreements
3	The President
4	Accounts
5	Administrative Personnel
6	Domestic Security
7	Agriculture
8	Aliens and Nationality
9	Animals and Animal Products
10	Energy
11	Federal Elections
12	Banks and Banking
13	Business Credit and Assistance
14	Aeronautics and Space
15	Commerce and Foreign Trade
16	Commercial Practices
17	Commodity and Securities Exchanges
18	Conservation of Power and Water Resources
19	Customs Duties
20	Employees' Benefits
21	Food and Drugs
22	Foreign Relations
23	Highways
24	Housing and Urban Development
25	Indians
26	Internal Revenue
27	Alcohol, Tobacco Products and Firearms
28	Judicial Administration
29	Labor
30	Mineral Resources
31	Money and Finance: Treasury
32	National Defense
33	Navigation and Navigable Waters
34	Education
35	Panama Canal
36	Parks, Forests, and Public Property
37	Patents, Trademarks, and Copyrights
38	Pensions, Bonuses, and Veterans' Relief
39	Postal Service
40	Protection of Environment
41	Public Contracts and Property Management
42	Public Health
43	Public Lands: Interior
44	Emergency Management and Assistance
45	Public Welfare
46	Shipping
47	Telecommunication
48	Federal Acquisition Regulations System
49	Transportation
50	Wildlife and Fisheries

Source: Legal Information Institute, http://www.law.cornell.edu/cfr/text, accessed June 25, 2012.

Titles do not 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), as well as

many other regulations in many other titles. There is no convincing mapping between titles and industries based purely on the name of the title.

The *CFR* itself is based on a complementary publication called the *Federal Register*. The *FR* 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 *FR* corresponds to the flow of regulations and the *CFR* corresponds to the stock. We focus our attention on the *CFR* principally because the *FR* may measure bureaucratic activity more than regulatory growth. For each final regulation published in the *FR*, there are also pages of preamble text explaining the regulation, economic analyses of the regulation, a Paperwork Reduction Act analysis, and a multitude of other obligatory pages that, while related to the regulation, do not directly affect economic agents. Furthermore, the *FR* contains notices of proposed rulemakings and advanced notices of proposed rulemakings—documents that explain regulatory agencies' plans but that are not binding regulations.

Even worse (from the econometrician's perspective), the *FR* also contains a large number of nonregulatory 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 *FR* is at best a noisy measure of regulation and at worst a biased measure because the number of pages associated with individual rulemakings has increased over time as acts of Congress or executive orders have required more analyses.[16]

Perhaps the most significant advantage of the *CFR* over the *FR* is that it allows for decreases in regulations. Various titles decrease in length at various points in time, perhaps reflecting some degree of deregulation. Using simple measures based on the *FR* restricts measures of the flow of regulations to always equal zero or greater (since you cannot have negative numbers of pages or rulemakings), even when the precise content of the *FR* might reflect deregulation.

A. Simple Methods for Quantifying Aggregate Regulations

A number of researchers have introduced simple methods for quantifying regulations.[17] The first method is to collect page-count data from either the *FR* or the *CFR*. These page counts provide an excellent departure point and have furnished several insightful inquiries into the causes and consequences of regulations.

Page-count data are subject to the criticism that not all pages are equal. A page, or an entire set of pages in a final rulemaking, could be of enormous consequence to the economy or could go virtually unnoticed. Also, page-formatting guidelines may change over time. Further, some 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

[16] Crews, *Ten Thousand Commandments*, somewhat mitigates this drawback by focusing only on pages devoted to final rules. P. A. McLaughlin, "The Consequences of Midnight Regulations and Other Surges in Regulatory Activity," *Public Choice* 147, (2011): 395–412.

[17] Coglianese, "Empirical Analysis and Administrative Law"; Mulligan and Shleifer, "The Extent of the Market"; Dawson and Seater, "Federal Regulation"; Coffey, McLaughlin, and Tollison, "Regulators and Redskins"; and Crews, *Ten Thousand Commandments*.

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: not all rules are of equal consequence.

Mulligan and Shleifer use file-size data from the statutes of 37 U.S. states.[18] The use of file-size data permits the researcher to overcome the possibility of differences in formatting, such as font sizes, that would distort the comparison of page-count data across states. Following their lead, we have gathered a second measure: file-size data for the electronically published versions of the *CFR*.

Our database provides both *CFR* page-count data and file-size data in addition to a third measure described below in section 2B.

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 our understanding of the causes and consequences of regulation will surely be enhanced by quantifying the cross-sectional variation. We attempt this quantification below in section 2C.

B. 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 the modal verbs "shall" and "must" and the adjective "prohibited." This observation motivated us to search the *CFR* for key words 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 used custom computer programs to count the occurrences of each of these five strings in each title of the *CFR* published from 1997 through 2010, with the exception of title 35. Title 35 contained regulations relevant to the Panama Canal and has not been amended since 2000.[19] Titles 2 and 6 do not exist at the start of our dataset, but they are included in our dataset after their respective inceptions in 2005 and 2004.[20] Table 2 provides summary statistics of our constraint count by *CFR* title.

[18] Mulligan and Shleifer, "The Extent of the Market."

[19] The Panama Canal was ceded to Panama on December 31, 1999, though an unchanged Title 35 was published for several additional years before being terminated in 2004.

[20] Title 2 addresses government grants and procurement procedures. These procedures previously existed in the form of memorandums and other guidance documents, but they were formally added to the *CFR* beginning in 2005. Title 6, which covers domestic security, was first published in 2004 when the newly created Department of Homeland Security began rule promulgation.

Table 2. Summary Statistics of Constraint Count, 1997–2010, by *CFR* Title

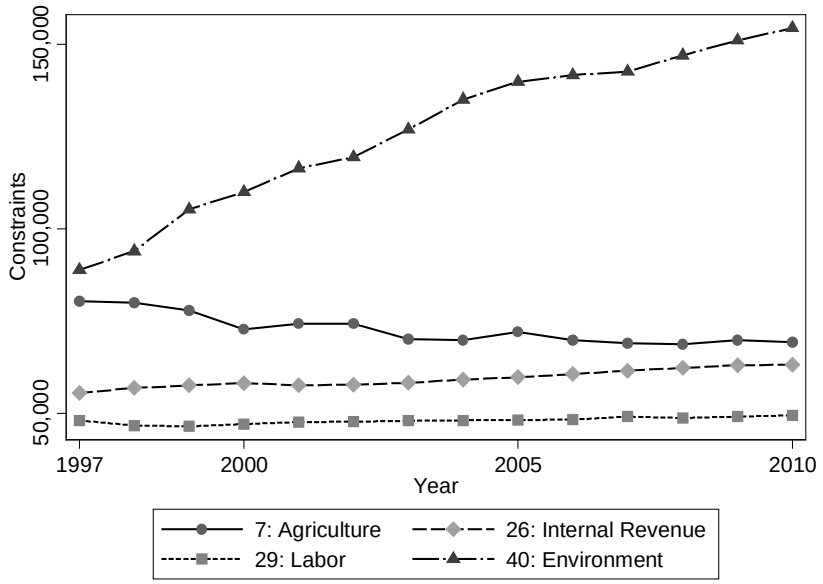
CFR Title	Subject	# Obs	Mean	SD	Min.	Max.
1	General Provisions	14	397	1	396	399
2	Grants and Agreements	6	1,232	465	338	1,642
3	The President	14	750	276	416	1,410
4	Accounts	14	764	110	665	971
5	Administrative Personnel	14	11,862	748	10,837	13,062
6	Domestic Security	7	1,047	246	754	1,386
7	Agriculture	14	72,786	4,065	68,844	80,398
8	Aliens and Nationality	14	8,663	1,875	6,016	10,733
9	Animals and Animal Products	14	17,628	530	16,729	18,504
10	Energy	14	23,824	1,782	21,491	27,269
11	Federal Elections	14	3,164	387	2,700	3,658
12	Banks and Banking	14	25,280	2,770	21,676	31,534
13	Business Credit Assistance	14	3,895	596	2,897	4,710
14	Aeronautics and Space	14	29,305	3,460	26,002	35,312
15	Commerce and Foreign Trade	14	9,276	398	8,477	9,709
16	Commercial Practices	14	9,756	503	9,017	10,679
17	Commodity and Securities Exchange	14	18,085	1,116	16,504	19,582
18	Conservation of Power and Water Resources	14	10,877	1,050	9,818	12,415
19	Custom Duties	14	12,270	537	11,138	13,047
20	Employees' Benefits	14	16,628	3,293	5,823	18,788
21	Food and Drugs	14	20,239	1,099	18,761	22,193
22	Foreign Relations	14	11,337	176	10,998	11,702
23	Highways	14	3,820	140	3,632	4,054
24	Housing and Urban Development	14	23,193	820	21,617	24,637
25	Indians	14	10,092	947	8,244	11,361
26	Internal Revenue	14	59,442	2,409	55,596	63,243
27	Alcohol, Tobacco Products and Firearms	14	10,781	131	10,609	10,980
28	Judicial Administration	14	9,837	578	8,778	10,423
29	Labor	14	48,108	888	46,528	49,509
30	Mineral Resources	14	21,415	459	21,067	22,603
31	Money and Finance: Treasury	14	8,238	965	6,593	9,421
32	National Defense	14	22,618	969	21,425	24,199
33	Navigation and Navigable Waters	14	14,675	1,175	13,248	16,454
34	Education	14	9,882	430	9,283	10,799
35	Panama Canal	3	1,348	798	426	1,809
36	Parks, Forests, and Public Property	14	11,474	459	10,425	12,120
37	Patents, Trademarks, and Copyrights	14	4,679	778	3,640	5,898
38	Pensions, Bonuses, and Veterans' Relief	14	8,540	814	7,514	10,102
39	Postal Service	14	3,375	121	3,165	3,545
40	Protection of Environment	14	126,594	21,275	88,852	154,350

41	Public Contracts and Property Management	14	9,251	303	8,941	9,928
42	Public Health	14	14,955	2,178	11,450	18,486
43	Public Lands: Interior	14	14,095	873	13,430	16,658
44	Emergency Management and Assistance	14	3,985	218	3,768	4,480
45	Public Welfare	14	16,881	664	15,509	17,596
46	Shipping	14	34,790	239	34,384	35,279
47	Telecommunication	14	24,350	1,038	22,484	25,754
48	Federal Acquisition Regulations System	14	29,371	1,578	27,064	32,335
49	Transportation	14	41,378	4,690	34,220	48,380
50	Wildlife and Fisheries	14	15,285	4,713	10,010	26,135

Source: Authors' calculations.

Our new measure of regulations, denoted *constraints*, is the total number of restrictions in a title. Restrictions are measured by the total number of occurrences in a title of the five constraining strings that we searched for. All searches used to create this database are case insensitive. Table 2 gives summary statistics of the variable constraints for each *CFR* title over the 14-year period. Figure 1 depicts the constraints over this time period for the four *CFR* titles with the greatest number of constraints, on average, of any of the 50 titles.

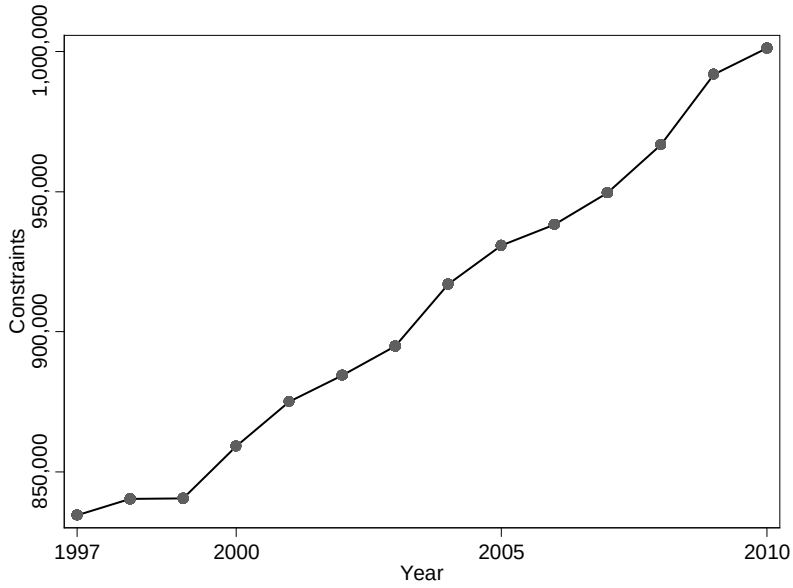
Figure 1. Constraints of the Four *CFR* Titles with Highest Overall Means, 1997–2010



Source: Authors' calculations.

Figure 2 shows the total constraints published each year in the *CFR*—that is, the summation of all of the occurrences of the five constraint strings annually in all the titles. The persistent growth of the total number of constraints in the *CFR* seems to confirm the popular notion that federal regulation has grown regardless of the political party in charge of the executive branch. Total constraints increased from 834,649 in 1997 to 1,001,153 in 2010.

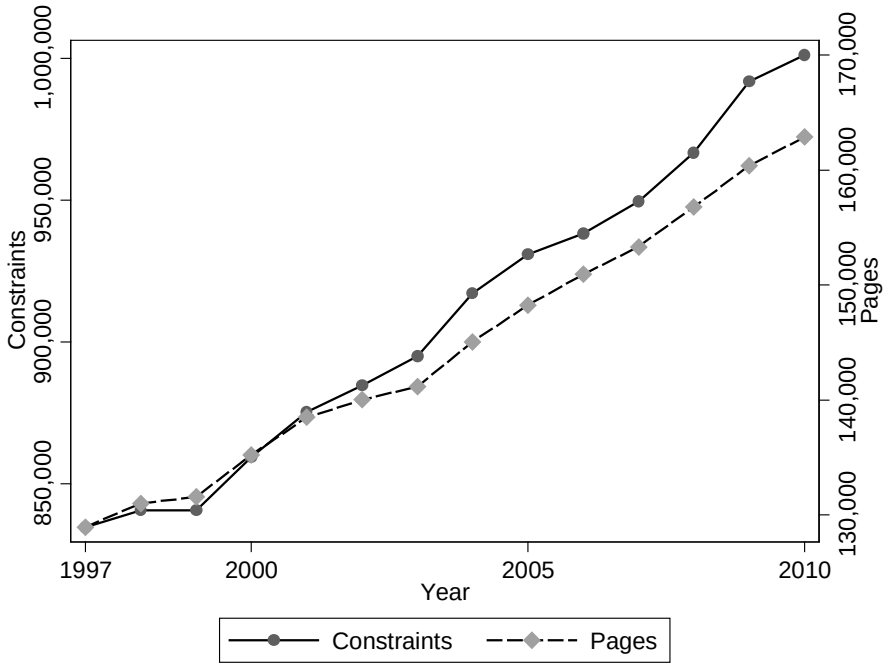
Figure 2. Total Annual Regulatory Constraints, 1997–2010



Source: Authors' calculations.

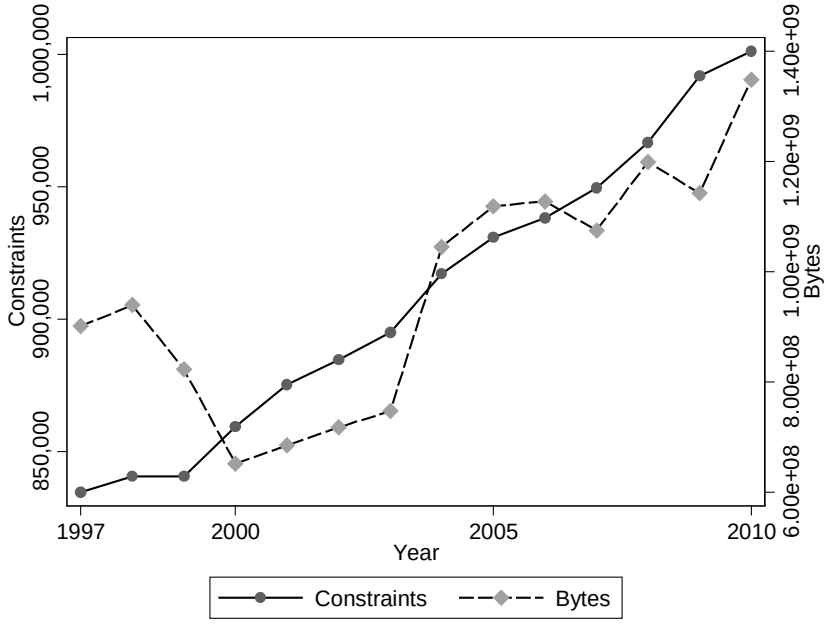
Figure 3 juxtaposes constraints with page count data, while figure 4 shows constraints alongside file-size data. The superiority of constraints compared to page counts and file sizes is an open empirical question. Figure 4 suggests that page-count data and file-size data are less correlated over time than page-count data and constraints. At the title level, the correlation between constraints and page counts is 0.96, the correlation between constraints and file size is 0.58, and the correlation between page counts and file size is 0.66 (671 observations, all significant at the $p < 1\%$ level). We do not offer an explanation.

Figure 3. Total Annual Regulatory Constraints Compared to Total Annual *CFR* Pages, 1997–2010



Source: Authors' calculations.

Figure 4. Total Annual Regulatory Constraints Compared to Bytes of Digitized *CFR*, 1997–2010



Source: Authors' calculations.

C. Quantifying the Applicability of Regulations to Specific Industries Using Text Analysis

The North American Industry Classification System (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. The coarsest is two-digit, followed by three-digit, four-digit, five-digit, and finally the finest, six-digit.[21] Table 3 illustrates the gradation with an example. Table 4 shows the two-digit classification, and Table 5 shows the three-digit classification.

[21] For more on the NAICS codes and descriptions, see U.S. Census Bureau, "North American Industry Classification System: Introduction," http://www.census.gov/eos/www/naics/.

Table 3. An Example of NAICS Descriptions from Two- to Six-Digit Specificity

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

Source: U.S. Census Bureau, http://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart_code=31&search=2007%20NAICS%20Search, accessed June 25, 2012.

Table 4. Two-Digit NAICS Classifications

Code	Description
11	Agriculture, Forestry, Fishing and Hunting
21	Mining, Quarrying, and Oil and Gas Extraction
22	Utilities
23	Construction
31	Manufacturing
42	Wholesale Trade
44	Retail Trade
48	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate and Rental and Leasing
54	Professional, Scientific, and Technical Services
55	Management of Companies and Enterprises
56	Administrative and Support and Waste Management and Remediation Services
61	Educational Services
62	Health Care and Social Assistance
71	Arts, Entertainment, and Recreation
72	Accomodation and food services
81	Other Services (except Public Administration)
92	Public Administration

Source: U.S. Census Bureau, http://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2007, accessed June 25, 2012.

Table 5. Three-Digit NAICS Classifications

Code	Description	Code	Description
111	Crop Production	482	Rail Transportation
112	Animal Production	483	Water Transportation
113	Forestry and Logging	484	Truck Transportation
114	Fishing, Hunting and Trapping	485	Transit and Ground Passenger Transportation
115	Support Activities for Agriculture and Forestry	486	Pipeline Transportation
211	Oil and Gas Extraction	487	Scenic and Sightseeing Transportation
212	Mining (except Oil and Gas)	488	Support Activities for Transportation
213	Support Activities for Mining	491	Postal Service
221	Utilities	492	Couriers and Messengers
236	Construction of Buildings	493	Warehousing and Storage
237	Heavy and Civil Engineering Construction	511	Publishing Industries (except Internet)
238	Specialty Trade Contractors	512	Motion Picture and Sound Recording Industries
311	Food Manufacturing	515	Broadcasting (except Internet)
312	Beverage and Tobacco Product Manufacturing	517	Telecommunications
313	Textile Mills	518	Data Processing, Hosting, and Related Services
314	Textile Product Mills	519	Other Information Services
315	Apparel Manufacturing	521	Monetary Authorities-Central Bank
316	Leather and Allied Product Manufacturing	522	Credit Intermediation and Related Activities
321	Wood Product Manufacturing	523	Securities, Comm. Contracts, and Other Fin. Invest.
322	Paper Manufacturing	524	Insurance Carriers and Related Activities
323	Printing and Related Support Activities	525	Funds, Trusts, and Other Financial Vehicles
324	Petroleum and Coal Products Manufacturing	531	Real Estate
325	Chemical Manufacturing	532	Rental and Leasing Services
326	Plastics and Rubber Products Manufacturing	533	Lessors of Nonfinancial Intangible Assets
327	Nonmetallic Mineral Product Manufacturing	541	Professional, Scientific, and Technical Services
331	Primary Metal Manufacturing	551	Management of Companies and Enterprises
332	Fabricated Metal Product Manufacturing	561	Administrative and Support Services
333	Machinery Manufacturing	562	Waste Management and Remediation Services
334	Computer and Electronic Product Manufacturing	611	Educational Services
335	Electrical Equipment, Appliance, and Component Manufacturing	621	Ambulatory Health Care Services
336	Transportation Equipment Manufacturing	622	Hospitals
337	Furniture and Related Product Manufacturing	623	Nursing and Residential Care Facilities
339	Miscellaneous Manufacturing	624	Social Assistance
423	Merchant Wholesalers, Durable Goods	711	Performing Arts, Spectator Sports, and Related Industries
424	Merchant Wholesalers, Nondurable Goods	712	Museums, Historical Sites, and Similar Institutions
425	Wholesale Electronic Markets and Agents and Brokers	713	Amusement, Gambling, and Recreation Industries
441	Motor Vehicle and Parts Dealers	811	Repair and Maintenance
442	Furniture and Home Furnishings Stores	812	Personal and Laundry Services
443	Electronics and Appliance Stores	813	Religious, Grantmaking, Civic, Prof., and Similar Organiz.
444	Building Material and Garden Equipment and Supplies Dealers	814	Private Households
445	Food and Beverage Stores	921	Executive, Legislative, and Other General Government Support
446	Health and Personal Care Stores	922	Justice, Public Order, and Safety Activities
447	Gasoline Stations	923	Administration of Human Resource Programs
448	Clothing and Clothing Accessories Stores	924	Administration of Environmental Quality Programs
451	Sporting Goods, Hobby, Book, and Music Stores	925	Admin. of Housing Programs, Urban Planning, and Comm. Develop.
452	General Merchandise Stores	926	Administration of Economic Programs
453	Miscellaneous Store Retailers	927	Space Research and Technology
454	Nonstore Retailers	928	National Security and International Affairs
481	Air Transportation		

Source: U.S. Census Bureau, http://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2007, accessed June 25, 2012.

Regulation data for all three measures—constraints, page counts, and file sizes—were gathered or created at the title level. Each title corresponds to a broad area subject to federal regulation. As argued earlier, there is no convincing mapping from title to NAICS codes based purely on title name. Our goal was to use text analysis to measure the applicability of the regulations contained in a specific title to a specific industry.

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 so the search strings were “finance and insurance,” “finance,” and “insurance.”

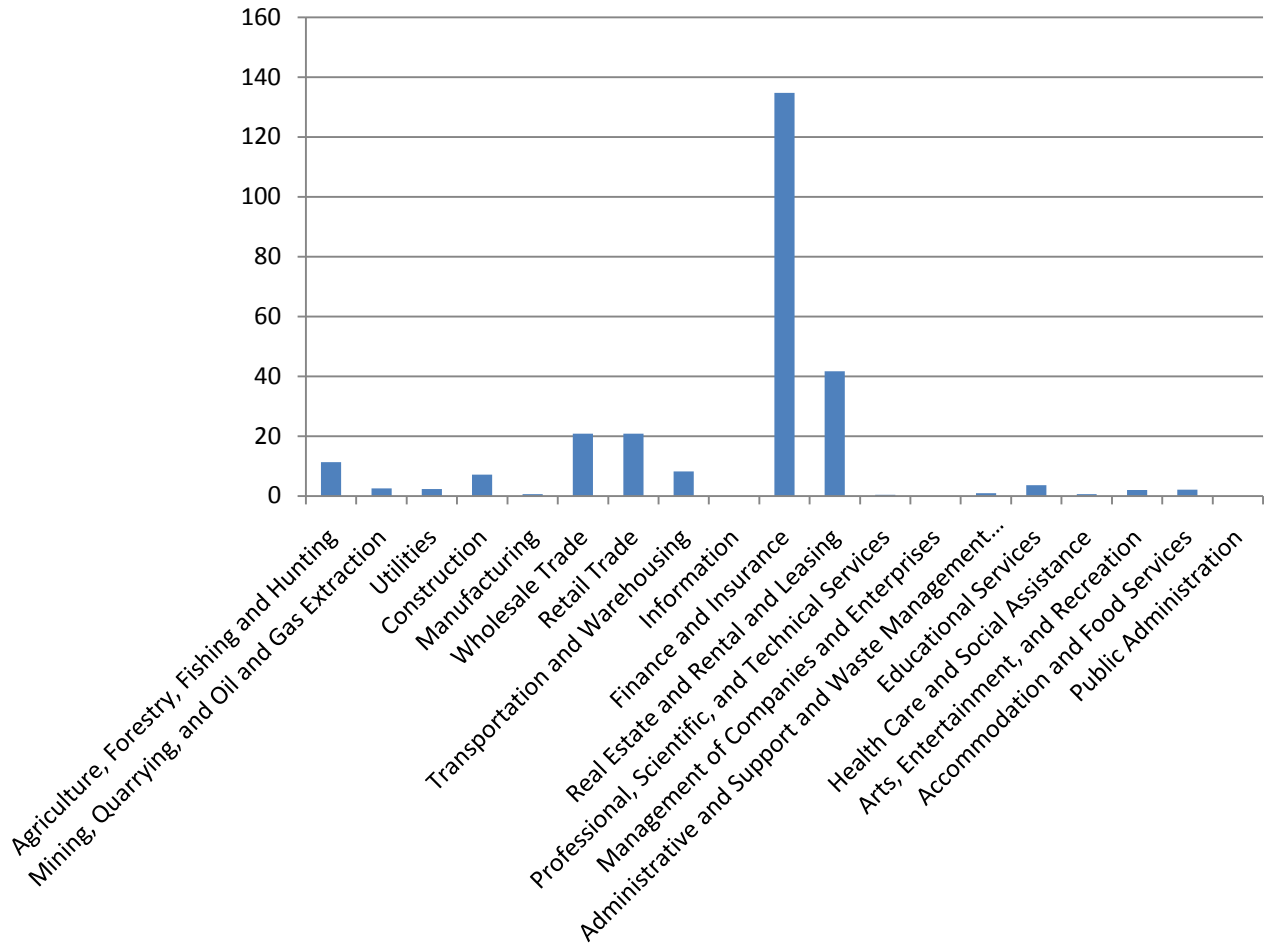
We created these search strings through rules we devised to transform NAICS descriptions into multiple search strings. The decision of what rules to create and to follow necessarily required some subjective judgment. In the interest of transparency, these rules are fully explained in appendix A. All search strings created with this approach are given on the website (www.regulationdata.org), along with the rule used to create each string. Thus, if another researcher disagrees with any particular rule, she can elect to remove all strings based on that rule.

After forming each code’s search strings, we counted the occurrences of each search string for each two- and three-digit industry in each title of the 1997–2010 *CFR*. The resulting data give industry-specific measures of relevance—that is, a measure of the extent to which a title in a given year relates to a specific industry as defined in the two- and three-digit NAICS classifications.[22]

We offer a few ways to visualize the results of our measurements of industry relevance. Figure 5 shows the relevance of one particular *CFR* title—Title 12, Banks and Banking—to all the two-digit NAICS industries, which are shown along the horizontal axis. The bars show the number of occurrences of the industry-specific search strings that were found in Title 12 in the year 2010. As we would expect, Title 12 appears most relevant to the “Finance and Insurance” industry (code 52), followed by the “Real Estate and Rental and Leasing” industry (code 53).

[22] Our suggested measure of industry relevance is deflated by the number of pages in a title; we explain this measure more fully in appendix A. As with many aspects of this database, users are also able to modify or remove this deflation.

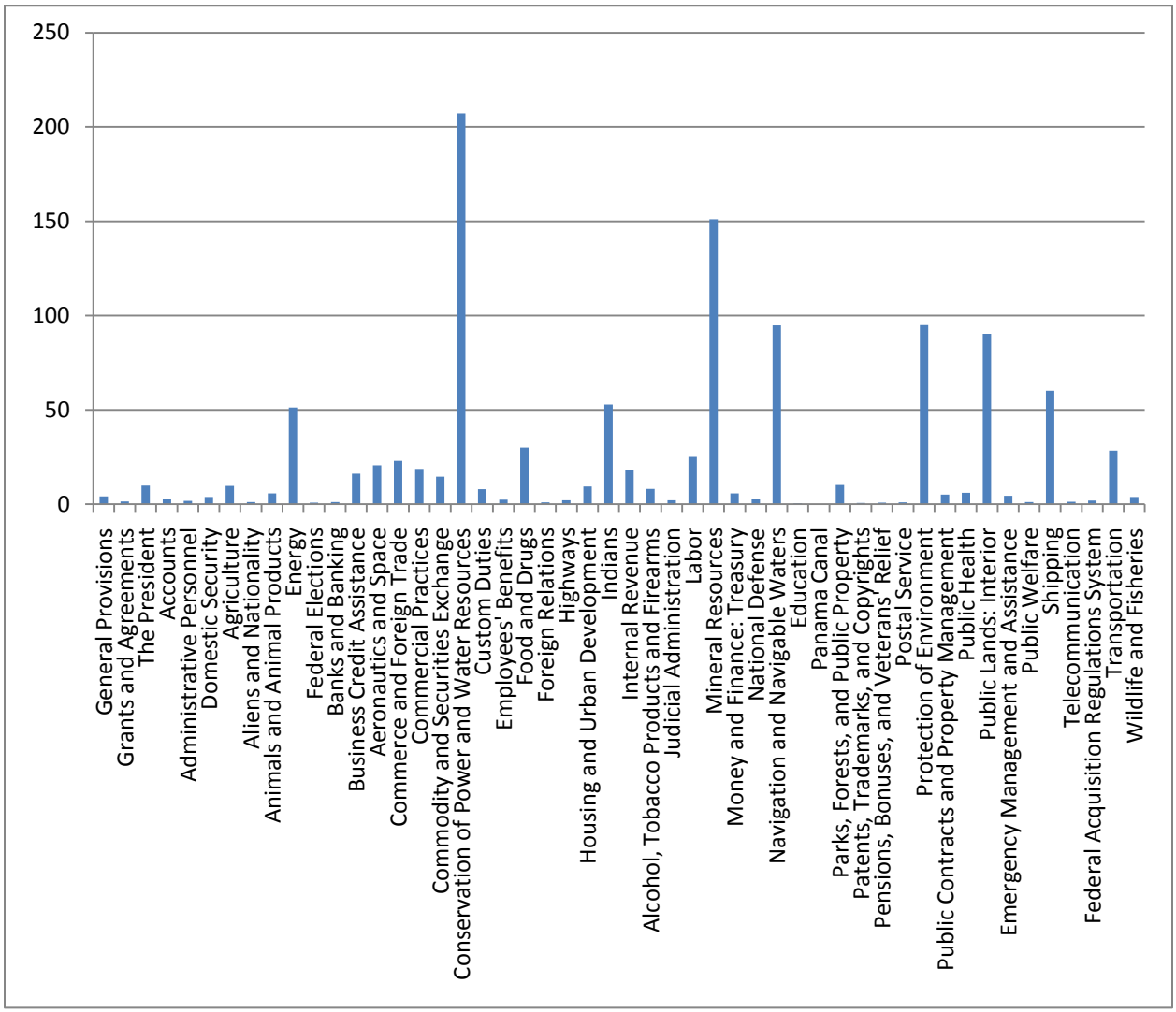
Figure 5. Relevance (Citations per 100 Pages) of Title 12 to Two-Digit Industries in 2010



Source: Authors' calculations.

As an alternative visualization, figure 6 shows an example of relevance of each *CFR* title to three-digit NAICS industry code 211, "Oil and Gas Extraction," for the year 2010. Figure 6 shows that the search strings for the oil and gas extraction industry show up most often (after deflating for the number of pages in a title) in Title 18 (Conservation of Power and Water Resources), Title 30 (Mineral Resources), Title 33 (Navigation and Navigable Waters), Title 40 (Protection of Environment), and Title 43 (Public Lands: Interior). These are exactly the titles that we would expect to most intensively regulate this industry.

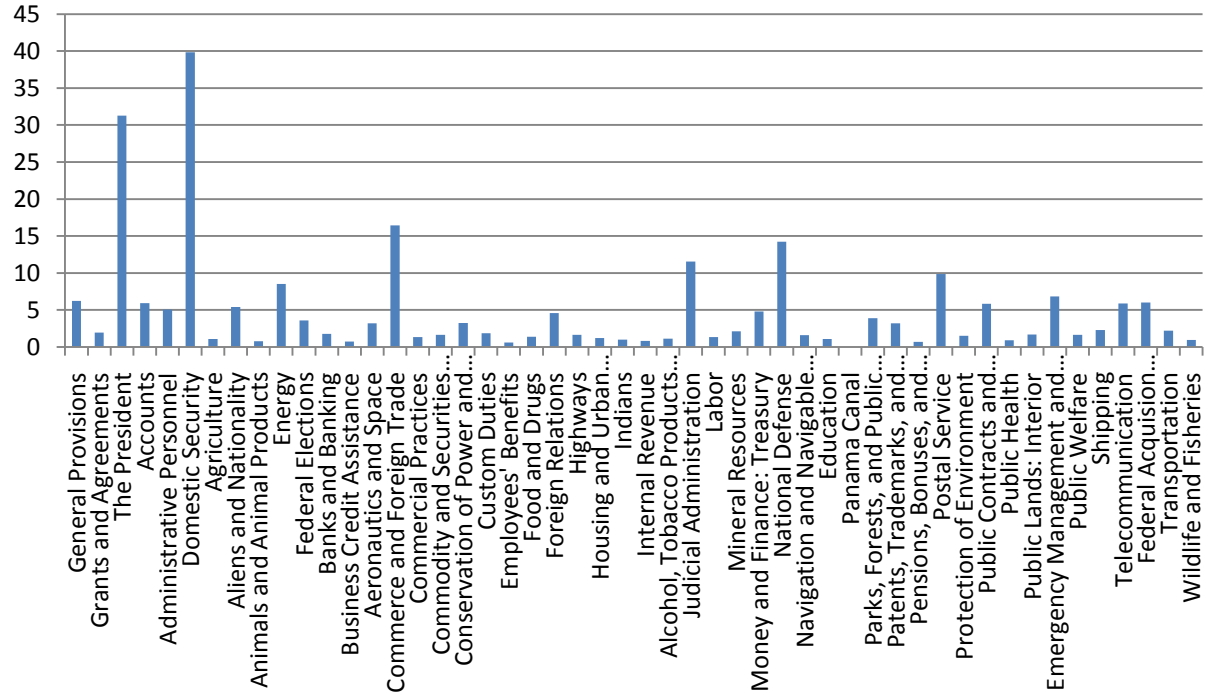
Figure 6. Industry Relevance (Citations per 100 Pages) for "Oil and Gas Extraction" (NAICS Code 211) by Title in 2010



Source: Authors' calculations.

As another example, figure 7 shows the cross-section of industry relevance by *CFR* title in year 2010 for the "National Defense and International Affairs" industry (code 928). This industry is largely comprised of the armed forces and government services related to immigration and international programs, so it is perhaps not surprising to see that Title 6 (Domestic Security) is the most relevant, followed closely by Title 3 (The President), which covers the commander-in-chief of the armed forces. Title 15 (Commerce and Foreign Trade), Title 32 (National Defense), and Title 28 (Judicial Administration) are the next three most relevant titles.

Figure 7. Industry Relevance (Citations per 100 Pages) for "National Security and International Affairs" (NAICS Code 928) by Title in 2010



Source: Authors' calculations.

There are a variety of ways to interpret and use the data. For example, if one wants to compare Title 40's relevance to "Chemical Manufacturing" (code 325) to Title 40's relevance to "Motor Vehicle and Parts Dealers" (code 441) for the year 2000, one method is to directly compare the hits on the strings from 325 to those from 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 the different methods in appendix A.

One of the database's drawbacks is that some search strings associated with a code are likely to occur more frequently for linguistic reasons unrelated to the relevance of a title to the industry in question. For example, gauging the relevance of a title to the "Information" industry (code 51) based on occurrences of the word "information" will likely lead to an exaggeration compared to, say, the "Construction" industry (code 23), because the word "information" may be used without any reference to the activities of the information sector. To address this shortcoming, we have flagged those search strings that we deem likely to occur in irrelevant text, and we make this information available in the data on the website (www.regulationdata.org).

We speculate that a better correction can be made by combining our data with data on the probability of certain words or phrases occurring in natural language. Alternatively, humans can be employed to assess applicability for random subsets of occurrences of the words. We hope to do both in future versions of this database. In the meantime, in the interests of transparency and to promote fruitful experimentation, we make the entire database available.

D. Combining the Two Databases to Create a Panel

A title-specific measure of regulation—for example, constraints, page counts, or file sizes—can be combined with our data on the relevance of *CFR* titles to specific industries to create a panel data set indicating industry-specific regulation from 1997 through 2010. Let R_{ty} be the number of regulations in title t in year y, based on a measure of regulation from section 2A (*CFR* pages or file sizes) or 2B (constraints). Assuming that the weight a regulation receives in *total* regulations does not depend upon the title, $R_y = \sum_t R_{ty}$ is a measure of the total number of regulations in year y.

Let a_{tyi} be the applicability of the regulations in title t in year y to industry i taken from the data in section 2C. We want to construct a new index r_{tyi} measuring the regulations for industry i in title t in year y. The relationship will be of the form $r_{tyi} = f(a_{tyi}, R_{ty})$, where f is increasing in both elements and the cross-partial is positive, too. The simplest possibility is $f(a_{tyi}, R_{ty}) = a_{tyi} R_{ty}$; alternatively, one could use a function of the form $f(a_{tyi}, R_{ty}) = D(a_{tyi}) R_{ty}$, where D is a dummy variable that takes the value 1 when $a_{ty\,i}$ is above a threshold. Finally, assuming equal title weighting as above, $r_{yi} = \sum_t r_{tyi}$ will be a measure of the regulations on industry i in year y. We provide $r_{tyi} = a_{tyi} R_{ty}$ as the default industry-specific regulation index. However, as above, to promote fruitful experimentation, we make the entire data set available, permitting anyone to construct different industry-specific regulatory indices using different weightings or combinations of a_{tyi} and R_{ty}.

NAICS classifications 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. There are, therefore, many opportunities to merge our database with other data to explore the causes and outcomes of regulations.

Closing Remarks

This paper introduces the Industry-specific Regulatory Constraint Database (IRCD). IRCD is the first iteration of the first product 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 regulation.

IRCD allows users to combine two databases to create a panel database that annually quantifies federal regulations by industry for all U.S. industries and regulations from 1997 to 2010. The first database contains three metrics of regulation quantity: *CFR* page-count data, digitized *CFR* file-size data, and a novel measure called *constraints* that counts the number of binding words (e.g., “shall” or “must”) contained in regulatory text.

In the second database, we offer the first measure of the relevance of *CFR* titles to industries in the United States. This measure was created by searching each title of the *CFR* for text strings that describe each industry in the United States, as defined by the two- and three-digit codes of the North American Industry Classification System (NAICS), and summing the number of hits in each title and each year. We based the descriptions of industries on two- and three-digit NAICS industry descriptions in part to allow IRCD to be combined with data on specific outcomes that may be affected by regulation, such as industrial

performance, safety data, or environmental outcomes. Many publicly available data sets are also based on the NAICS, such as value added to GDP by industry, thus lending compatibility with IRCD.[23]

IRCD offers users numerous choices that we hope will permit maximum experimentation and minimize any subjectivity inherent in the creation of the database. Users can decide how to combine the databases (e.g., whether and how to weight constraints in a given *CFR* title by industry relevance to that title), which measures of the quantity of regulation to use, and whether to omit or include specific strings from the constraints database or from the industry search strings.

While we offer this iteration of IRCD to the public with the goal of facilitating regulatory research, we hope to refine IRCD in several ways and release those refined versions in the future. First, because some industry descriptions contain words or strings that are common in natural language, we hope to develop a method of weighting search string results according to the probability of the string occurring.

Second, our novel measure of regulation—constraints—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.

Third, we currently report data at the title level; as described above, titles are divided into thousands of parts covering specific regulatory areas. In the next iteration, we plan to report data at the part level to deliver a much higher resolution of industry-specific regulation.

Finally, 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 IRCD. 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.

[23] For GDP-by-industry data, see Bureau of Economic Analysis, "Industry Economic Accounts," http://www.bea.gov/industry/.

Appendix A: Construction of the Industry Relevance Metric

Appendix A explains how we constructed the industry relevance metric. First, by decomposing typical NAICS industry descriptions, we describe the structure of industry descriptions. Second, we explain the rules we developed to turn the NAICS industry descriptions into a set of search strings. Third, we cover some shortcomings of our systems and offer possible solutions to the individual user of the database. Finally, we explain how we calculated the industry relevance metric and discuss alternative ways to calculate it.

A. Industry Name Structure

The NAICS industry description is a collection of words or phrases linked by conjunctions or commas, e.g., "Agriculture, Forestry, Fishing and Hunting," or "Finance and Insurance" (we discuss some important exceptions below). The full description can be divided into an exhaustive collection of phrases that may have some overlap in shared words. For example, "Oil and Gas Extraction" can be divided into "Oil Extraction" and "Gas Extraction."

Each individual phrase is a **noun phrase**. The noun phrase has up to three components.

Head noun: The main word in the phrase. This can be in the form of a present participle [*Fishing*] or not [*Construction*].

Pre-modifiers: Words that precede the head noun and modify its meaning. They can be adjectives [*Educational* in "Educational Services"], nouns [*Waste Management* in "Waste Management Services"] or a mixture [*Electronic Product* in "Electronic Product Manufacturing"]. They can also be absent [*Construction*].

Post-modifiers: Words that follow the head noun and modify its meaning. They can be nouns [*Companies* in "Management of Companies"] or a mixture of adjectives and nouns [*Economic Programs* in "Administration of Economic Programs"]. They can also be absent [*Construction*]. We ignore prepositions.

B. Rules for Strings

Each of the following rules applies to each of the full phrases derived from the industry description. All searches are case insensitive.

Rule 1: The full phrase.

- Conditions: None.
- Examples: [*wholesale trade*].
- Exceptions: None.

Rule 2: The singular form of the full phrase.

- Conditions: The full phrase is naturally pluralized.
- Examples: [*utility* in “utilities”].
- Exceptions: None.

Rule 3: The person who does the full phrase (singular).

- Conditions: A commonly used version actually exists.
- Examples: [*retail trader* in “retail trade”].
- Exceptions: None.

Rule 4: The person who does the full phrase (plural).

- Conditions: A commonly used version actually exists.
- Examples: [*retail traders* in “retail trade”].
- Exceptions: None.

Rule 5: The head noun.

- Conditions: The full phrase is composed of more than one word.
- Examples: [*trade* in “wholesale trade”].
- Exceptions: The head noun is used extensively in the *CFR* to convey a meaning that is fundamentally different from the meaning in the phrase. Exclude [*assistance* in “social assistance”] and [*services* in “educational services”].

Rule 6: The base form of the head noun.

- Conditions: The full phrase is only one word AND the head noun is a present participle.
- Examples: [*hunt* in “hunting”].
- Exceptions: None.

Rule 7: The pre-modifiers together as a whole string.

- Conditions: The head noun has pre-modifiers.
- Examples: [*waste management* in “waste management services”].
- Exceptions: The pre-modifiers are used extensively in the *CFR* to convey a meaning that is fundamentally different to the meaning in the phrase. Exclude [*public* in “public administration”].

Rule 8: The post-modifiers together as a whole string.

- Conditions: The head noun has post-modifiers.
- Examples: [*human resource programs* in “Administration of Human Resource Programs”].
- Exceptions: The post-modifiers are used extensively in the *CFR* to convey a meaning that is fundamentally different from the meaning in the phrase. Exclude [*enterprises* in “management of enterprises”].

Rule 9: Individual words and phrases in pre-modifiers and post-modifiers.

- Conditions: The head noun has more than one pre- or post-modifier.
- Examples: [*coal* in "coal products manufacturing"].
- Exceptions: The pre- or post-modifiers are used extensively in the *CFR* to convey a meaning that is fundamentally different from the meaning in the phrase. Exclude [*products* in "plastics products manufacturing"].

In our database, we begin by dividing the industry description into the individual noun phrases described above; within the industry, each noun phrase is assigned a group number to distinguish its strings from those belonging to the other noun phrases. For example, in the industry "oil and gas extraction," *oil extraction* is assigned group 1 and *gas extraction* is assigned group 2.

C. Situations When the Rules Are Ineffective

The above rules are ineffective in three infrequent classes of NAICS industry descriptions. The first is when the industry description involves a parenthetical comment, typically an exception, such as "mining (except oil and gas)." Our solution is to simply ignore the parenthetical comment. The following industries suffer from this problem:

- 212 (Mining [except Oil and Gas])
- 511 (Publishing Industries [except Internet])
- 515 (Broadcasting [except Internet)]
- 533 (Lessors of Nonfinancial Intangible Assets [except Copyrighted Works])

The second is the case of "other," "support," or "related" activities, such as "support activities for mining" or "furniture and related product manufacturing." We apply the rules in the normal fashion; however, in some of these cases, the outcome is unlikely to fully reflect the spirit of the NAICS industry description. The following industries suffer from this problem:

- 56 (Administrative and Support and Waste Management and Remediation Services)
- 115 (Support Activities for Agriculture and Forestry)
- 323 (Printing and Related Support Activities)
- 337 (Furniture and Related Product Manufacturing)
- 488 (Support Activities for Transportation)
- 519 (Other Information Services)
- 525 (Funds, Trusts, and Other Financial Vehicles)
- 562 (Administrative and Support Services)
- 711 (Performing Arts, Spectator Sports, and Related Industries)
- 712 (Museums, Historical Sites, and Similar Institutions)
- 813 (Religious, Grantmaking, Civic, Professional, and Similar Organizations)
- 921 (Executive, Legislative, and Other General Government Support)

The final case is that of industry names that contain the word "general" or "miscellaneous," such as "general merchandise stores." We apply the rules in the normal fashion. However, in some of these cases, the outcome is unlikely to fully reflect the spirit of the NAICS industry description. The following industries suffer from this problem:

- 339 (Miscellaneous Manufacturing)
- 452 (General Merchandise Stores)
- 453 (Miscellaneous Store Retailers)

We have omitted industry description 81 (Other Services [Except Public Administration]) because any search for strings based on the words "other services" would return useless results. We have also omitted 423 (Merchant Wholesalers, Durable Goods) and 424 (Merchant Wholesalers, Nondurable Goods); they are the only three-digit industries that fall under 42 (Wholesale Trade), and we cannot think of a sensible way of distinguishing them since they do not follow the phrase structure of the other NAICS industry names. Therefore, we direct the reader to the data on 42 (Wholesale Trade) only.

In the next subsection, we discuss alternative techniques for calculating industry relevance. Some of these techniques can remedy the problems of rule-ineffectiveness encountered in the above situations.

D. Calculating Industry Relevance

Each industry description is associated with a collection of strings. The strings are classified according to group (see the end of section B in this appendix) and rule. For each group in each industry, each rule in the range 1–8 is associated with at most *one* string. Rule 9 can yield multiple strings associated with the same group or industry.

As an illustration, consider industry 316 (Leather and Allied Product Manufacturing). The industry name is composed of two phrases: *leather manufacturing* (group 1) and *allied product manufacturing* (group 2). The resulting strings are in table A1.

Table A1. Strings Associated with Industry 316 (Leather and Allied Product Manufacturing)

String	Group	Rule
leather manufacturing	1	1
allied product manufacturing	2	1
leather manufacturer	1	3
allied product manufacturer	2	3
leather manufacturers	1	4
allied product manufacturers	2	4
manufacturing	1	5
manufacture	1	6
leather	1	7
allied product	2	7
allied	2	9
product	2	9

Source: Authors' compilation using rules described in this appendix.

In this example, based on our discretionary interpretation of the rules, we exclude *manufacturing*, *manufacture*, *allied*, and *product*. In the final database, there is a variable denoting which strings we recommend including/excluding, though we still measure the occurrence of every string to allow readers to judge for themselves. Though we judge each rule-9 string on individual merit, in the default version of the final database (which we use for the figures and tables in the main text), we exclude all rule-9 strings. In appendix C, we detail strings where we struggled to decide on inclusion or exclusion.

As table A1 shows, some of the smaller strings are contained in the larger strings from the same group. More formally, each string derived from rules 1, 2, 3, or 4 can potentially contain the head noun (string from rule 5), the pre-modifier (string from rule 7) or post-modifier (string from rule 8) from the same group. (We ignore containment of the strings from rule 9 because we are excluding rule 9 strings.) We therefore create three additional dummy variables: *contains_head_noun*, *contains_pre_modifier*, and *contains_post_modifier*. These variables make it easy to use statistical software to eliminate double-counting. For example, every occurrence of the string "leather manufacturing" automatically implies an occurrence of the string "leather," but we would only want to count such an occurrence once. We provide programming code for Stata that prevents double-counting by using these variables.

In some cases, a string is shared by multiple groups in the same industry, e.g., *manufacturing* in the example in table A1. We assign such shared strings to the first group that shares them since we are ultimately aggregating at the industry level, and so assigning them to multiple groups within the same industry will result in double-counting.

Once we have eliminated the possibility of double-counting, for each industry or title, we sum the total occurrences of the included strings in that title. We then divide that sum by the number of pages in the title and multiply by 100 to obtain a measure of **industry relevance per hundred pages**. This measure prevents longer titles from appearing to be more relevant to an industry simply by virtue of their length. Users have the opportunity to undo this act of deflation should they so desire.

What we have described above is the **standard/direct** approach. To address the shortcomings described in section C of appendix A, one can employ a **bottom-up** approach. For example, consider industry 81 (Other Services [except Public Administration]). No meaningful search based on the strings in its name can be made. However, it houses the following three-digit industries: 811 (Repair and Maintenance), 812 (Personal and Laundry Services), 813 (Religious, Grantmaking, Civic, Professional, and Similar Organizations), and 814 (Private Households). Thus, an index of its relevance can be constructed by aggregating the relevance of its three-digit subindustries. Future iterations of this database will include industries at the 4-digit, 5-digit, and 6-digit levels, permitting a much richer bottom-up approach.

Users may also wish to employ a **top-down hybrid** approach, where the relevance of a three-digit industry is calculated by applying the standard approach to the industry itself and adding the relevance of its two-digit parent industry.

Appendix B: Using the Data Files

The database is composed of 20 comma separated value (.csv) files, available at www.regulationdata.org. There are also two annotated Stata program (.do) files that transparently clean the data and can be easily modified according to the user's preferences. We describe each file and the variables contained therein.

---

data_constraints.csv: This file contains the frequency of each command string by year/title/volume.

- *year*: year from {1997, 1998, …, 2010}
- *title*: *CFR* title from {1, 2, …, 50}
- *volume*: *CFR* volume (positive integer)
- *string*: binding constraint, from {required, must, prohibited, shall, may not}
- *count*: the number of times the string appears in the year/title/volume (positive integer)

The search is case insensitive, but the whole string much be matched (e.g., the word "muster" will not result in a hit for the string "must").

---

data_file_size.csv: This file contains the file size in bytes by year/title/volume.

- *year*: year from {1997, 1998, …, 2010}
- *title*: *CFR* title from {1, 2, …, 50}
- *volume*: *CFR* volume from {1, 2, …}
- *file_size*: *CFR* year/title/volume file size in bytes

For a small subset of volumes, digital copies were obtained from an alternative source with a different file format, preventing a direct file size comparison. We used the following imputation procedure. Let the baseline source be source A and the alternative be source B. Suppose year_X_title_Y_volume_Z was acquired from the alternative source. We compared the file size across the two sources for the following neighboring volumes (where these actually existed):

- year_X-1_title_Y_volume_Z
- year_X+1_title_Y_volume_Z
- year_X_title_Y_volume_Z-1
- year_X_title_Y_volume_Z+1

For each neighboring volume, we divided file_size_B by file_size_A to obtain a factor. We found the arithmetic means of the factors for the four neighboring volumes. We divided file_size_B by the mean factor to obtain the imputed file_size_A. We undertook this procedure for the following volumes:

- Year 2002 / Title 4 / Volume 1 (mean factor = 13.23)
- Year 2007 / Title 14 / Volume 3 (mean factor = 12.21)
- Year 1999 / Title 20 / Volume 3 (mean factor = 14.81)
- Year 1999 / Title 26 / Volume 7 (mean factor = 14.18)

- Year 1997 / Title 40 / Volume 19 (mean factor = 7.77)

Since our database spans 2,960 volumes, this imputation procedure was barely used in the proportionate sense.

---

data_page_count.csv: This file contains the number of pages by year/title/volume.

- *year*: year from {1997, 1998, …, 2010}
- *title*: *CFR* title from {1, 2, …, 50}
- *volume*: *CFR* volume from {1, 2, …}
- *page_count*: *CFR* year/title/volume number of pages (positive integer)

Approximately half the page counts can be considered too long by one page. The digital versions of the *CFR* volumes randomly included a blank page at the end of the document, and we included that page in our page counts. We also included the title page and the explanation pages (which are numbered with Roman numerals). Since the average volume length is almost 700 pages, the blank page is insignificant.

---

naics2_X.csv (where X is an element of {1, 2, 3}): This file contains the frequency of each two-digit industry-relevance string by year/title/volume.

- *year*: year from {1997, 1998, …, 2010}
- *title*: *CFR* title from {1, 2, …, 50}
- *volume*: *CFR* volume (positive integer)
- *string*: string derived from NAICS industry description according to rules specified above
- *count*: the number of times the string appears in the year/title/volume (positive integer)
- *code*: two-digit NAICS industry code
- *group*: when the industry code can be divided into multiple noun phrases, each noun phrase and its associated strings are assigned a group number (positive integer) that is unique at the industry level
- *rule*: the rule number generating the string
- *excluded*: a dummy variable taking the value 1 if the authors think that the string should be excluded according to the exclusion criteria in the rules
- *contains_head_noun*: dummy variable taking the value 1 if the string contains the string specified in the head noun; missing observation for strings associated with rules 5, 6, 7, 8, or 9
- *contains_pre_modifier*: dummy variable taking the value 1 if the string contains the string specified in the pre-modifier; missing observation for strings associated with rules 5, 6, 7, 8, or 9
- *contains_post_modifier*: dummy variable taking the value 1 if the string contains the string specified in the post-modifier; missing observation for strings associated with rules 5, 6, 7, 8, or 9

The search is case insensitive, but the whole string much be matched (e.g., the word “manufacturers” will not result in a hit for the string “manufacturer”).

---

naics3_X.csv (where X is an element of {1, 2, …, 11}): This file contains the frequency of each three-digit industry-relevance string by year/title/volume. Everything else is identical to **naics2_X.csv**.

---

names_naics2.csv: This file contains the full names of the two-digit NAICS industries.

- *code*: two-digit NAICS industry code
- *industry_name*: the industry description taken directly from the NAICS definitions

---

names_naics3.csv: This file contains the full descriptions of the three-digit NAICS industries. Everything else is identical to **names_naics2.csv**.

---

names_titles.csv: This file contains the full names of the *CFR* titles.

- *title*: *CFR* title from {1, 2, …, 50}
- *title_name*: the *CFR* title name

---

cleaning_naics2.do: this Stata .do file cleans and combines the above data files. It aggregates over volumes and presents two-digit industry data at the year/title level.

- *year*: year from {1997, 1998, …, 2010}
- *title*: *CFR* title from {1, 2, …, 50}
- *title_name*: the *CFR* title name
- *code_2*: two-digit NAICS industry code
- *industry_2_name*: the industry description taken directly from the NAICS definitions
- *industry_2_relevance*: the total number of times each individual string associated with the two-digit industry appears in that title/year per 100 pages
- *count_X*: the total number of times the string "X" appears in that title/year, where X is from {required, must, prohibited, shall, may not}
- *page_count*: *CFR* year/title number of pages (positive integer)
- *file_size*: *CFR* year/title file size in bytes

This Stata file has been tested with all versions of Stata including and beyond Stata 9.

---

cleaning_naics3.do: This Stata .do file cleans and combines the above data files. It aggregates over volumes and presents three-digit industry data at the year/title level. All variables are identical or

analogous to **cleaning_naics2.do**, except that we include two-digit industry codes and names in case the user wants to use a bottom-up approach (see appendix A, section D).

This Stata file has been tested with all versions of Stata including and beyond Stata 9.

Appendix C: Controversial Inclusion/Exclusion

The authors found the following strings particularly controversial. Table A2 shows our final decisions concerning inclusion/exclusion; users are free to make their own decisions.

Table A2. Controversial Strings (Exclusion Dummy = 1 Implies Exclusion)

Code	String	Exclusion dummy
236	buildings	0
311	food	0
313	mills	0
314	mills	0
321	wood	0
322	paper	0
327	mineral product	0
331	metal	0
332	metal products	0
332	fabricated	0
332	metal	0
333	machinery	0
334	computer	0
335	appliance	0
339	manufacturing	0
339	manufacturer	0
339	manufacturers	0
444	garden	0
446	stores	1
446	personal care	0
448	stores	1
448	clothing	0
451	stores	1
451	book	0
451	music	0
452	stores	1
453	store	1
453	retailers	0
454	retailers	0
481	air	1
483	water	1
485	passenger	0
486	transportation	1
487	scenic	0
487	sightseeing	0
488	transportation	1
518	hosting	1
525	funds	1
525	trust	0
533	intangible assets	0
533	nonfinancial assets	0
562	waste	0
621	health care	0
812	launderer	0
812	launderers	0
813	grantmaking	0
813	civic	0
921	general government	0
923	human resource	0
928	security	1

Source: Authors' compilation.