RegData

A Numerical Database on Industry-Specific Regulations for All US Industries and Federal Regulations, 1997–2012

Omar Al-Ubaydli and
Patrick A. McLaughlin

November 2014

MERCATUS WORKING PAPER

MERCATUS CENTER
George Mason University

3434 Washington Blvd., 4th Floor, Arlington, Virginia 22201
www.mercatus.org

All studies in the Mercatus Working Paper series have followed a rigorous process of academic evaluation, including (except where otherwise noted) at least one double-blind peer review. Working Papers present an author's provisional findings, which, upon further consideration and revision, are likely to be republished in an academic journal. The opinions expressed in Mercatus Working Papers are the authors' and do not represent official positions of the Mercatus Center or George Mason University.

Omar Al-Ubaydli and Patrick A. McLaughlin. “RegData: A Numerical Database on Industry-Specific Regulations for All US Industries and Federal Regulations, 1997–2012.” Mercatus Working Paper, Mercatus Center at George Mason University, Arlington, VA, November 2014. http://mercatus.org/publication/regdata-numerical-database-industry-specific-regulations.

Please note that this particular working paper, which relates to an interactive, proprietary tool, was not subject to the standard review process for Mercatus Working Papers. Its methodology is open to further refinement and revision as the authors release subsequent versions of the tool, the second of which is available in beta at the time of this paper’s publication.

This paper updates an earlier version, published in October 2012 under the title “The Industry-Specific Regulatory Constraint Database (IRCD): A Numerical Database on Industry-Specific Regulations for All U.S. Industries and Federal Regulations, 1997–2010.”

Abstract

We introduce RegData, formerly known as the Industry-Specific Regulatory Constraint Database. RegData annually quantifies federal regulations by industry and by regulatory agency for all federal regulations from 1997 to 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. We then demonstrate several plausible relationships between industry regulation and variables of economic interest. 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.

JEL codes: K2, L5, N4, Y1

Keywords: regulation, industry

Author Affiliation and Contact Information

Omar Al-Ubaydli
Program Director, International and Geo-Political Studies
Bahrain Center for Strategic, International and Energy Studies
Affiliated Senior Research Fellow
Mercatus Center at George Mason University
omar@omar.ec

Patrick A. McLaughlin
Senior Research Fellow
Mercatus Center at George Mason University
pmclaughlin@mercatus.gmu.edu

RegData: A Numerical Database on Industry-Specific Regulations for All US Industries and Federal Regulations, 1997–2012

Omar Al-Ubaydli and Patrick A. McLaughlin

1. Introduction

Scholars have been analyzing the causes and consequences of government regulation for decades, leading to a vast and still-growing literature. A principal reason for the popularity of such inquiries is that regulations are an invaluable policy tool for addressing market failure (Pigou, 1938). However, the complexity of the political process means that regulations may not always be virtuously conceived (Stigler, 1971; McChesney, 1987), and the intricacy of the modern economy means that regulations 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 (theoretically or empirically) the causal effect of a unique regulation or a small collection of related regulations, such as air quality standards (Greenstone, 2002). 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 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 regulations created or in effect each year.[1] Mulligan and

[1] 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; 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).

Shleifer (2005) use the sizes, measured in kilobytes, of the digitized versions of state-level statutes as a proxy for real state-level regulation. Coffey et al. (2012) use the total number of pages published annually and quarterly in the *Federal Register*, the government's daily journal of bureaucratic activity including proposed and final regulations. Dawson and Seater (2008) 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 principal ways. First, we provide a novel measure that quantifies regulations by analyzing CFR text.[2] 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 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 United States annually for the years 1997–2012 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, including in the realm of regulation. Some demand liberalization, viewing regulation through the lens of public choice theory (Stigler,

[2] See Gentzkow and Shapiro (2010) and Baker, Blook, and Davis (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 subagency-specific search results.

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). We believe our new database could play an important role in resolving this controversy and in finding areas of common ground.

This paper proceeds as follows. In section 2, we explain the methods used in constructing the database and provide some simple descriptive statistics. Section 3 describes our validation exercises. Section 4 demonstrates plausible relationships between our metrics of regulation and variables of interest, and it delves into some of the database's more interesting implications. Section 5 offers closing remarks. Appendix A contains more details about the methods. All original data referred to in this paper are available to the public at http://www.regdata.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. A regulation may be in effect for up to one year before publication in the CFR, but ultimately, all regulations are published in the CFR. The CFR 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 multiple volumes. 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 observations and word counts at each level. Table 2 describes all titles used in the CFR in these years alongside more summary statistics on observations, word counts, and bytes.

Table 1. ***Code of Federal Regulations*** **(CFR) Organization**

CFR division	Granularity	Typical contents	Mean annual observations	Mean word count
Title	Least	Broad subject area of regulations	48	1,200,000
Chapter		Rules of an individual agency	410	150,000
Subchapter		Rules of a subagency	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	Most	Detailed requirement(s) related to the provision	1,600,000	39

Note: All numbers are rounded to two significant figures.

Table 2. ***Code of Federal Regulations*** **(CFR) Titles with Summary Statistics**

CFR title	Subject	No. of years	Mean (SD) words (thousands)	Mean (SD) bytes (thousands)
1	General Provisions	16	44 (3.2)	300 (21)
2	Grants and Agreements	8	150 (58)	1,000 (380)
3	The President	16	160 (28)	1,100 (190)
4	Accounts	16	72 (11)	460 (72)
5	Administrative Personnel	16	1,400 (130)	9,200 (860)
6	Domestic Security	9	120 (43)	780 (290)
7	Agriculture	16	6,100 (300)	40,000 (1900)
8	Aliens and Nationality	16	670 (150)	4,300 (930)
9	Animals and Animal Products	16	1,100 (64)	7,100 (400)
10	Energy	16	2,100 (230)	14,000 (1600)
11	Federal Elections	16	260 (39)	1,800 (260)
12	Banks and Banking	16	2,900 (710)	19,000 (4600)
13	Business Credit and Assistance	16	420 (74)	2,800 (500)
14	Aeronautics and Space	16	2,200 (260)	15,000 (1800)
15	Commerce and Foreign Trade	16	1,100 (91)	7,600 (580)
16	Commercial Practices	16	850 (63)	5,500 (400)

continued on next page

CFR title	Subject	No. of years	Mean (SD) words (thousands)	Mean (SD) bytes (thousands)
17	Commodity and Securities Exchanges	16	1,600 (260)	11,000 (1,700)
18	Conservation of Power and Water Resources	16	860 (86)	5,700 (590)
19	Customs Duties	16	1,200 (110)	8,300 (750)
20	Employees' Benefits	16	2,000 (300)	13,000 (2,000)
21	Food and Drugs	16	2,500 (120)	17,000 (760)
22	Foreign Relations	16	960 (98)	6,300 (640)
23	Highways	16	340 (28)	2,300 (180)
24	Housing and Urban Development	16	1,900 (76)	12,000 (460)
25	Indians	16	720 (99)	4,700 (650)
26	Internal Revenue	16	9,700 (920)	64,000 (6,200)
27	Alcohol, Tobacco Products and Firearms	16	970 (55)	6,200 (370)
28	Judicial Administration	16	1,000 (130)	6,600 (850)
29	Labor	16	3,600 (190)	24,000 (1200)
30	Mineral Resources	16	1,300 (77)	8,500 (490)
31	Money and Finance: Treasury	16	1,100 (210)	7,200 (1,400)
32	National Defense	16	2,500 (150)	17,000 (1,000)
33	Navigation and Navigable Waters	16	1,400 (170)	9,100 (1,100)
34	Education	16	1,300 (86)	8,400 (560)
35	Panama Canal	3	120 (40)	870 (250)
36	Parks, Forests, and Public Property	16	1,000 (71)	6,700 (470)
37	Patents, Trademarks, and Copyrights	16	450 (77)	3,100 (500)
38	Pensions, Bonuses, and Veterans' Relief	16	1,100 (110)	7,300 (730)
39	Postal Service	16	310 (13)	2,000 (83)
40	Protection of Environment	16	12,000 (2,400)	88,000 (16,000)
41	Public Contracts and Property Management	16	860 (23)	5,800 (190)
42	Public Health	16	1,900 (340)	12,000 (2,200)
43	Public Lands: Interior	16	1,100 (59)	7,200 (370)
44	Emergency Management and Assistance	16	410 (15)	2,700 (96)
45	Public Welfare	16	1,600 (160)	11,000 (1,100)
46	Shipping	16	2,100 (34)	15,000 (390)
47	Telecommunication	16	2,200 (91)	15,000 (530)
48	Federal Acquisition Regulations System	16	2,600 (160)	18,000 (1,200)
49	Transportation	16	3,300 (480)	22,000 (3,100)
50	Wildlife and Fisheries	16	2,400 (1,100)	17,000 (8,000)

Note: All means and standard deviations are in thousands and are rounded to two significant figures.

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), as well as many other regulations in many other titles. 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 our attention on the CFR principally because the *Federal Register* may measure bureaucratic activity more than regulatory growth. For each final regulation published in the *Federal Register*, there may also exist 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 *Federal Register* contains notices of proposed rulemaking and advanced notices of proposed rulemaking—documents that explain regulatory agencies' plans but that are not binding regulations.

Furthermore, the *Federal Register* 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 *Federal Register* 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.[5]

Perhaps the most significant advantage of the CFR over the *Federal Register* is that it allows for decreases in regulations. Various titles decrease in length at various points in time,

[5] Crews (2011) somewhat mitigates this drawback by focusing only on pages devoted to final rules (McLaughlin, 2011).

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 (since it is not possible to have negative numbers of pages or rulemakings), 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 and Shleifer, 2005; Dawson and Seater, 2008; Coffey, McLaughlin, and Tollison, 2012; Crews, 2011). 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 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 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: not all rules are of equal consequence.

Mulligan and Shleifer (2005) use file-size data from the statutes of 37 US states. 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. We gather file-size data but omit it from this paper for reasons of parsimony; those interested should contact the authors. However, we make available word count data at every CFR division. Word counts also overcome formatting and font size issues, and, simultaneously, are not affected by large graphics that often affect file sizes. Moreover, we devise and gather two additional, novel measures, which we describe below.

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.

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 the modal verbs "shall" and "must" and the adjective "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 "regulatory restrictions," or simply "restrictions" because they restrict legal choice sets.

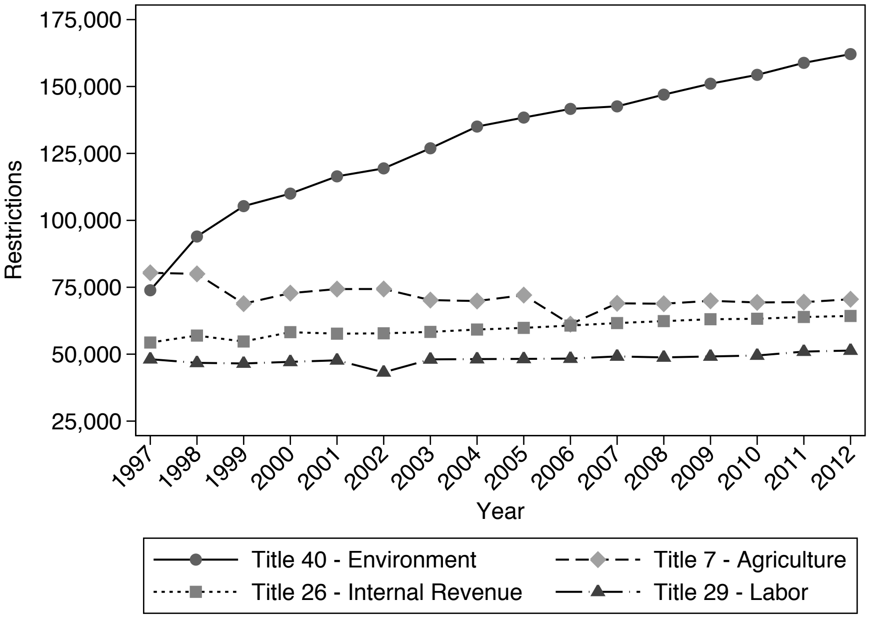
We use custom computer programs to count the occurrences of each of these five strings in each division of the CFR published from 1997 through 2012, with the exception of Title 35.[6] Titles 2 and 6 do not exist at the start of our dataset, but they are included in our dataset after

[6] Title 35 contained regulations relevant to the Panama Canal and has not been amended since 2000. 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.

their respective inceptions in 2005 and 2004. 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.

One of our new measures of regulations, denoted restrictions, is the total number of restrictions in a division of the CFR. RegData offers this measure at all levels of divisions given in table 1, with "title" being the broadest and "paragraph" the narrowest. Restrictions are measured by the total number of occurrences in a CFR division of the five restricting strings that we searched for. All searches used to create this database are case insensitive. Table 3 gives 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 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.

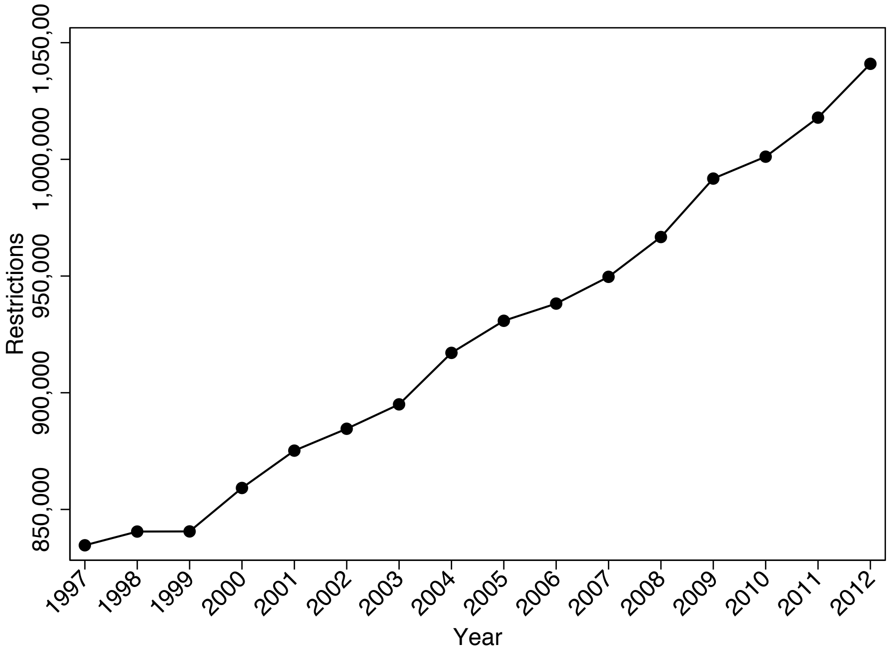
Table 3. Summary Statistics for Restrictions in *Code of Federal Regulations* Titles, 1997–2012

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
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 are rounded to two significant figures.

Figure 2 shows the total restrictions published each year in the CFR—that is, the summation of all occurrences of the five restriction strings annually in all the titles. The persistent growth of the total number of restrictions 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 restrictions increased from 830,000 in 1997 to 1 million in 2012. Over the same period, the total number of words in the CFR increased from 73 million to 100 million.

Figure 2. Total *Code of Federal Regulations* Restrictions, 1997–2012



Note: This graph covers all five restrictions across all titles.

Figure 3 juxtaposes total annual restrictions with total annual word counts, another measure of regulation contained in RegData. Figure 4 shows the yearly correlation between restrictions and word counts at the title, chapter, and part text levels. The superiority of restrictions compared to word counts is an open empirical question. The correlation between restrictions and word counts is 0.94 at the title level, 0.96 at the chapter level, and 0.93 at the part level, all of which are significant at the $p < 1\%$ level.

Figure 3. ***Code of Federal Regulations*** **Total Restrictions vs. Total Word Counts, 1997–2012**

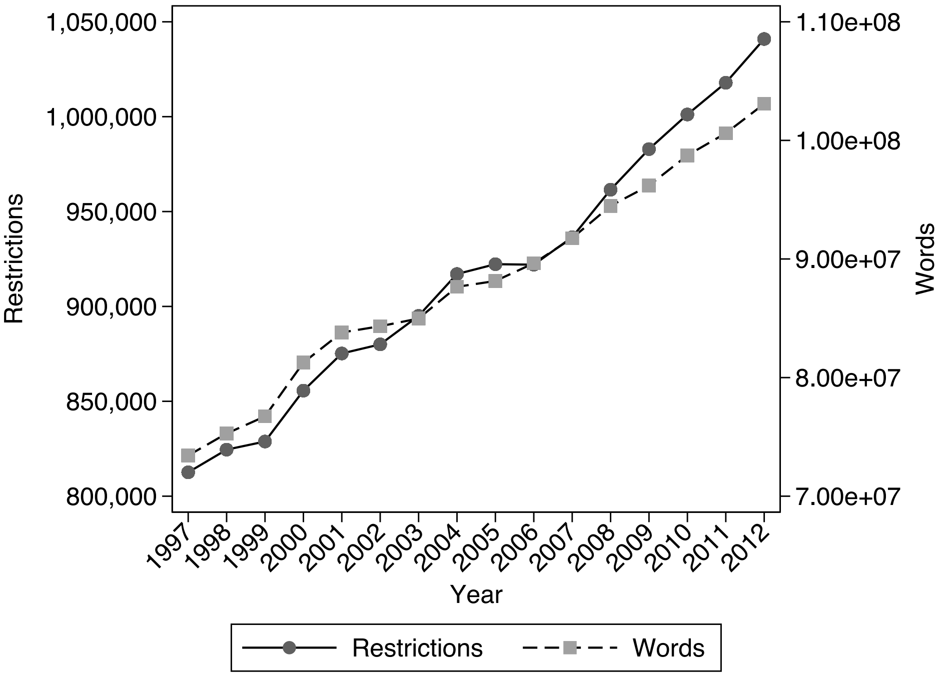
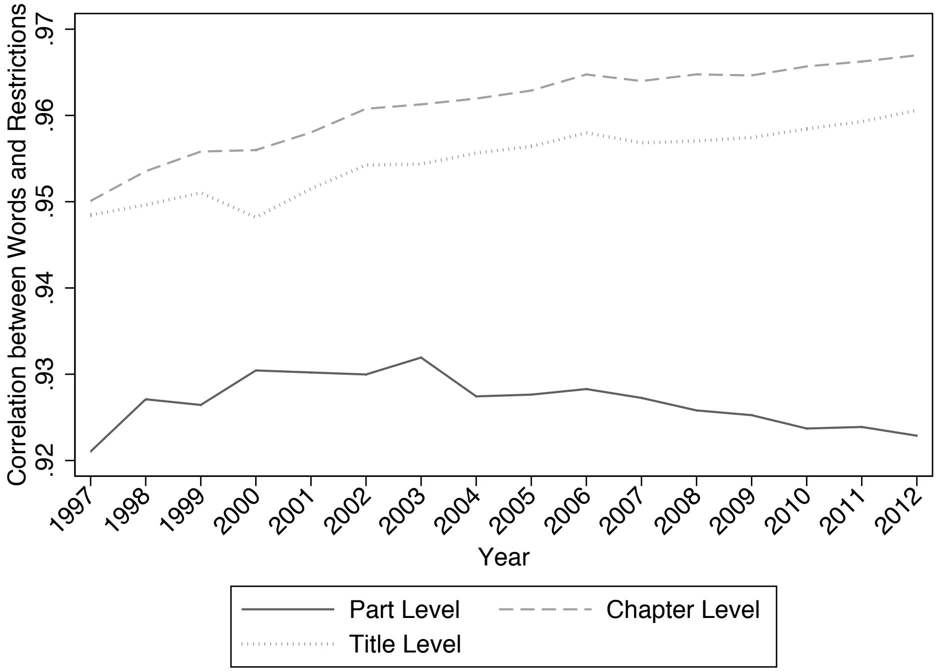
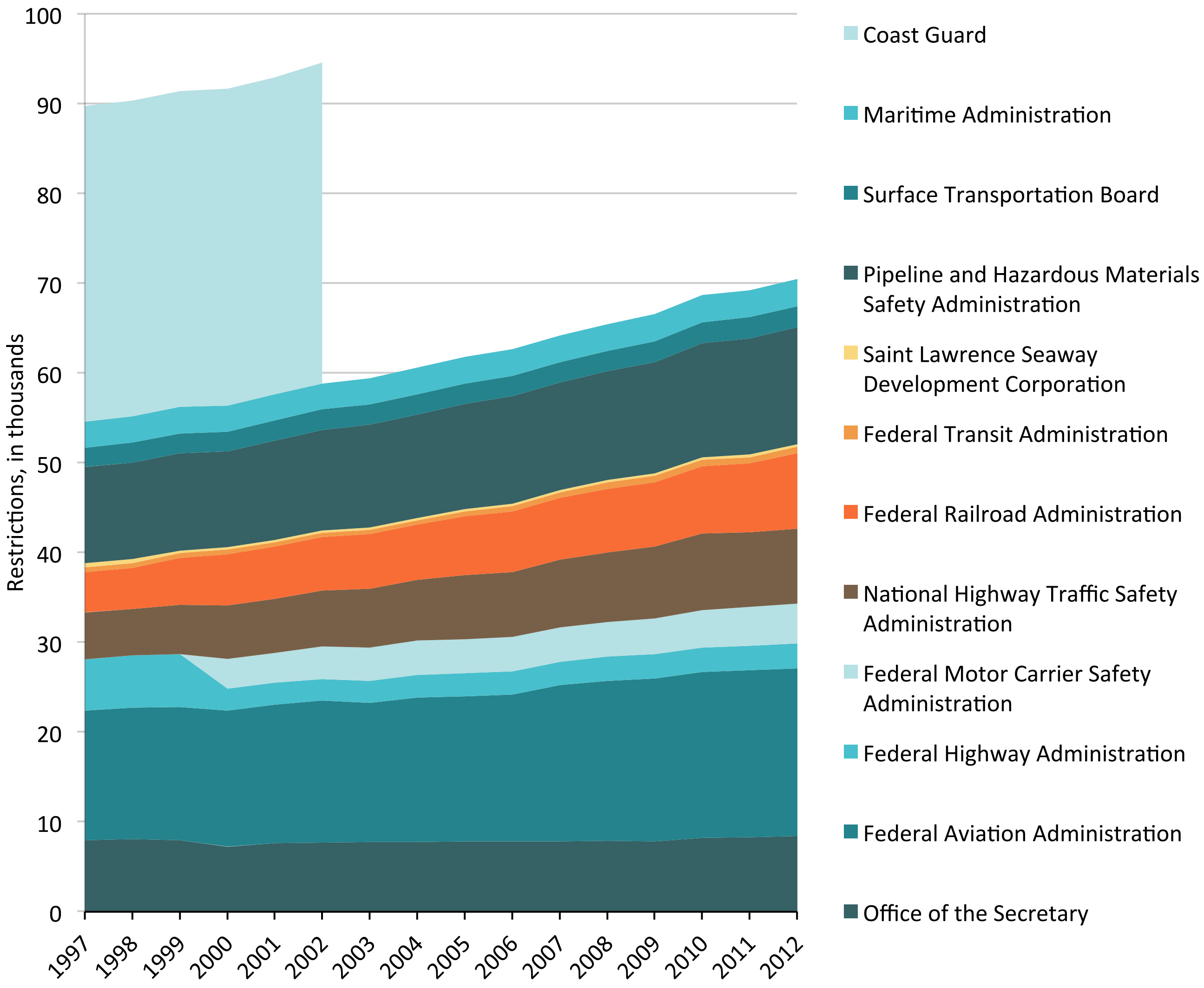


Figure 4. ***Code of Federal Regulations*** **Correlation between Word Counts and Restrictions, 1997–2012**



Our data also permit the user to examine regulation by department or agency. Figure 5 shows one of the larger and more complex departments, the Department of Transportation, and most of the agencies housed within that department. The most prominent feature of figure 5 is that the Coast Guard's restrictions fall to zero beginning in 2003. This is because the Homeland Security Act of 2002 moved the Coast Guard to the newly created Department of Homeland Security. The Coast Guard's restrictions did not disappear from the CFR or even from the titles they are printed in, but they stopped being included in the department-level time series for the Department of Transportation. Another shifting of regulations from one agency to another can be seen by tracing the Federal Motor Carrier Safety Administration, which was split off from the Federal Highway Administration in 2000.

Figure 5. Restrictions from the Department of Transportation by Agency, 1997–2012



Note: In the interest of space, we exclude three minor (in terms of restrictions) agencies: the Board of Contract Appeals, the Bureau of Transportation Statistics, and the Research and Innovative Technology Administration.

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. The coarsest is two-digit, followed by three-digit, four-digit, five-digit, and finally the finest, six-digit.[7] Table 4 illustrates the gradation with an example. Table 5 shows the two-digit classification, and table 6 shows the three-digit classification.[8]

We created regulation data for both measures—restrictions and word counts—at four CFR levels of granularity: title, chapter, part, and paragraph. As stated earlier, there is no definitive mapping from any division of the CFR to NAICS codes based purely on title name. Our goal was to use text analysis to measure the applicability of the regulations contained in each specific unit of the CFR to a specific industry.

Table 4. 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

[7] See http://www.census.gov/eos/www/naics/ for more information.
[8] See the NAICS homepage for the larger tables corresponding to four-, five-, and six-digit classifications.

Table 5. Two-Digit North American Industry Classification System Industries

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	Accommodation and Food Services
81	Other Services (except Public Administration)
92	Public Administration

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

Table 6. Three-Digit North American Industry Classification System Industries, Codes 111–928

Code	Description
111	Crop Production
112	Animal Production
113	Forestry and Logging
114	Fishing, Hunting and Trapping
115	Support Activities for Agriculture and Forestry
211	Oil and Gas Extraction
212	Mining (except Oil and Gas)
213	Support Activities for Mining
221	Utilities
236	Construction of Buildings
237	Heavy and Civil Engineering Construction
238	Specialty Trade Contractors
311	Food Manufacturing
312	Beverage and Tobacco Product Manufacturing
313	Textile Mills

continued on next page

Code	Description
314	Textile Product Mills
315	Apparel Manufacturing
316	Leather and Allied Product Manufacturing
321	Wood Product Manufacturing
322	Paper Manufacturing
323	Printing and Related Support Activities
324	Petroleum and Coal Products Manufacturing
325	Chemical Manufacturing
326	Plastics and Rubber Products Manufacturing
327	Nonmetallic Mineral Product Manufacturing
331	Primary Metal Manufacturing
332	Fabricated Metal Product Manufacturing
333	Machinery Manufacturing
334	Computer and Electronic Product Manufacturing
335	Electrical Equipment, Appliance, and Component Manufacturing
336	Transportation Equipment Manufacturing
337	Furniture and Related Product Manufacturing
339	Miscellaneous Manufacturing
423	Merchant Wholesalers, Durable Goods
424	Merchant Wholesalers, Nondurable Goods
425	Wholesale Electronic Markets and Agents and Brokers
441	Motor Vehicle and Parts Dealers
442	Furniture and Home Furnishings Stores
443	Electronics and Appliance Stores
444	Building Material and Garden Equipment and Supplies Dealers
445	Food and Beverage Stores
446	Health and Personal Care Stores
447	Gasoline Stations
448	Clothing and Clothing Accessories Stores
451	Sporting Goods, Hobby, Book, and Music Stores
452	General Merchandise Stores
453	Miscellaneous Store Retailers
454	Nonstore Retailers
481	Air Transportation
482	Rail Transportation
483	Water Transportation
484	Truck Transportation
485	Transit and Ground Passenger Transportation
486	Pipeline Transportation
487	Scenic and Sightseeing Transportation
488	Support Activities for Transportation
491	Postal Service
492	Couriers and Messengers
493	Warehousing and Storage
511	Publishing Industries (except Internet)
512	Motion Picture and Sound Recording Industries
515	Broadcasting (except Internet)

continued on next page

Code	Description
517	Telecommunications
518	Data Processing, Hosting, and Related Services
519	Other Information Services
521	Monetary Authorities–Central Bank
522	Credit Intermediation and Related Activities
523	Securities, Commodity Contracts, and Other Financial Investments . . .
524	Insurance Carriers and Related Activities
525	Funds, Trusts, and Other Financial Vehicles
531	Real Estate
532	Rental and Leasing Services
533	Lessors of Nonfinancial Intangible Assets
541	Professional, Scientific, and Technical Services
551	Management of Companies and Enterprises
561	Administrative and Support Services
562	Waste Management and Remediation Services
611	Educational Services
621	Ambulatory Health Care Services
622	Hospitals
623	Nursing and Residential Care Facilities
624	Social Assistance
711	Performing Arts, Spectator Sports, and Related Industries
712	Museums, Historical Sites, and Similar Institutions
713	Amusement, Gambling, and Recreation Industries
811	Repair and Maintenance
812	Personal and Laundry Services
813	Religious, Grantmaking, Civic, Professional, and Similar Organizations
814	Private Households
921	Executive, Legislative, and Other General Government Support
922	Justice, Public Order, and Safety Activities
923	Administration of Human Resource Programs
924	Administration of Environmental Quality Programs
925	Administration of Housing Programs, Urban Planning, and Community Development
926	Administration of Economic Programs
927	Space Research and Technology
928	National Security and International Affairs

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

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 to transform NAICS descriptions into multiple search strings. The decision of what rules to create and to follow necessarily required subjective judgment. In the interest of transparency, we fully explain these rules in Appendix A. We give all search strings created with this approach on the website, along with the rule used to create each string. Thus, if another researcher disagrees with any particular rule, the researcher can remove all strings based on that rule.

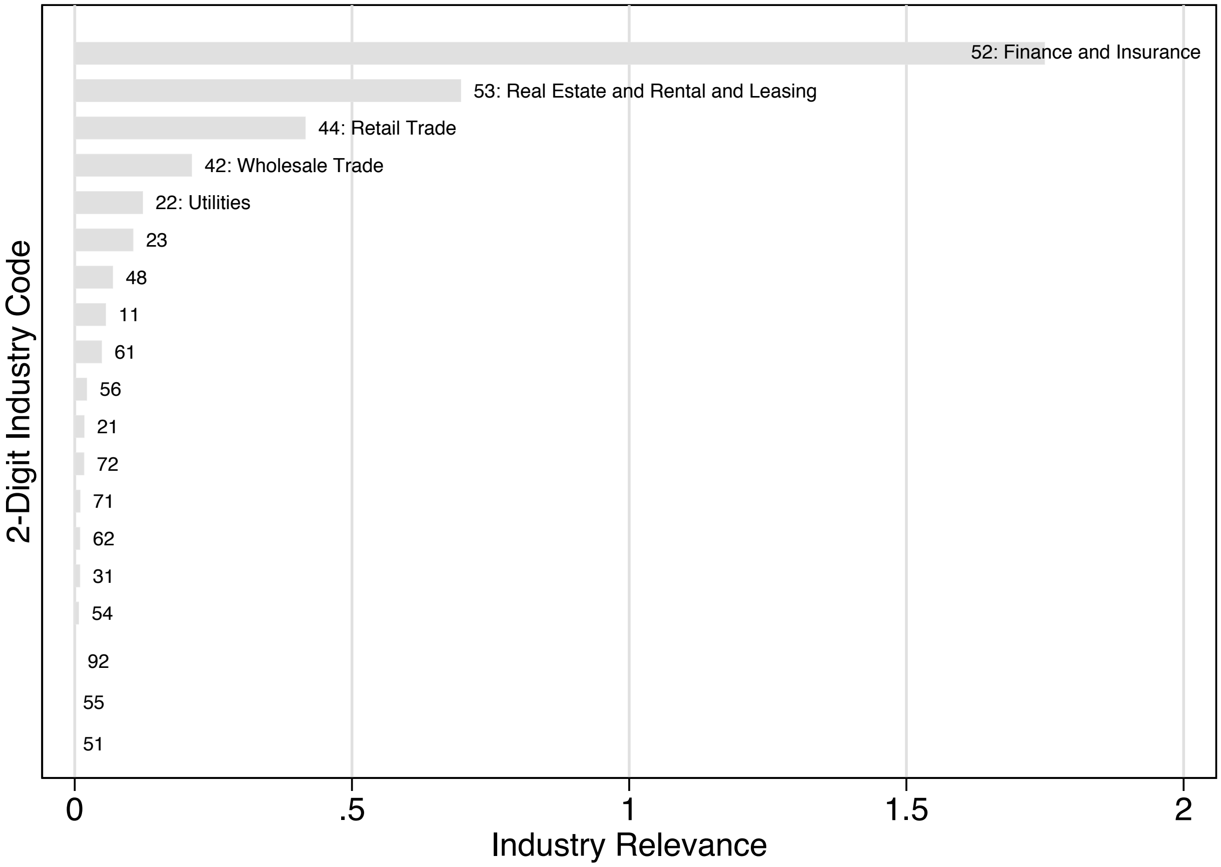
After forming each code's search strings, we counted the occurrences of each search string for each two-, three-, and four-digit industry in each division of the 1997–2012 CFR.[9] The resulting data give industry-specific measures of relevance—that is, measures of the extent to which a CFR division in a given year relates to specific industries as defined in the corresponding NAICS classifications. Our suggested measure of industry relevance is deflated by the number of words in the same CFR unit; 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.

We offer a few ways to visualize the results of our measurements of industry relevance, using titles as the CFR division for illustrative purposes. Figure 6 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 found in Title 12 in the year 2012, divided by Title 12's word count. As

[9] 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, 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.

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).

Figure 6. Relevance of *Code of Federal Regulations* Title 12, “Banks and Banking,” to All Two-Digit North American Industry Classification System Industries



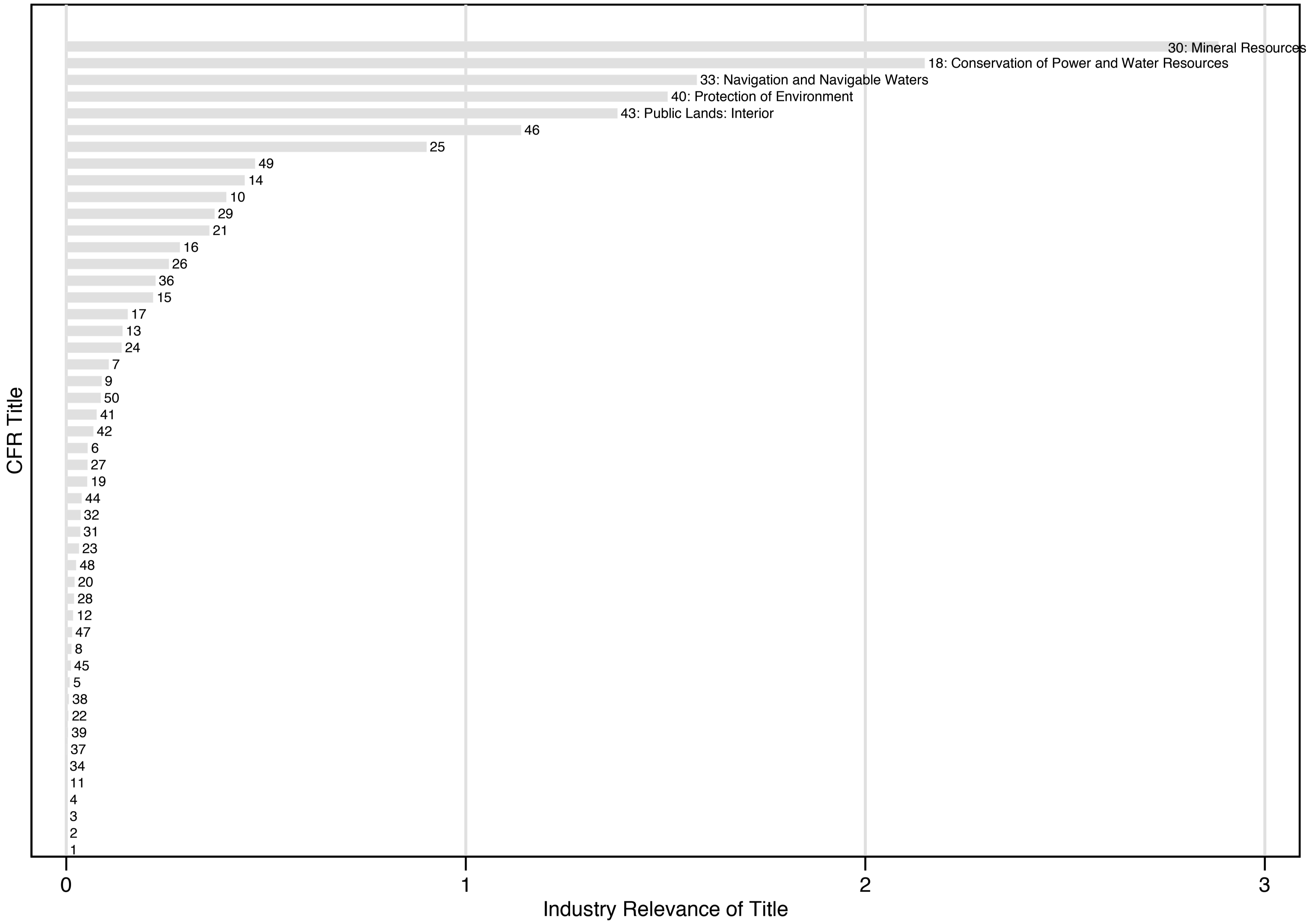
Note: Data are from 2012. The top five industries labeled; see table 5 for a list of all two-digit North American Industry Classification System industries and codes.

As an alternative visualization, figure 7 shows an example of the relevance of each CFR title to three-digit NAICS industry code 211, “Oil and Gas Extraction,” for the year 2012. Figure 7 shows that the search strings for the oil and gas extraction industry show up most often (after deflating for the number of words in a title) in Title 30 (Mineral Resources), Title 18 (Conservation of Power and Water Resources), Title 33 (Navigation and Navigable Waters), Title 40 (Protection of Environment), and Title 43 (Public Lands: Interior). These are the titles that common sense dictates should most intensively regulate this industry.

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 8, 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.

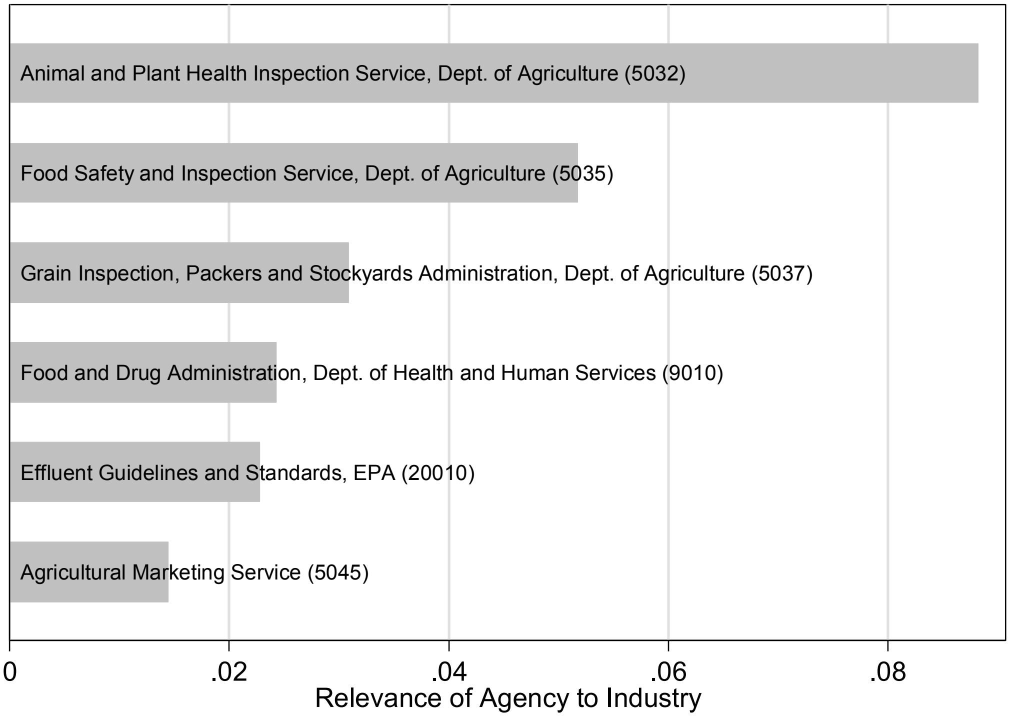
There are a variety of ways to interpret and use the data. For example, if one wants to compare Title 40's relevance to "Chemical Manufacturing" (code 325) with Title 40's relevance to "Motor Vehicle and Parts Dealers" (code 441) for the year 2000, one method is to directly compare the hits on the strings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

Figure 7. Relevance of All *Code of Federal Regulations* (CFR) Titles to North American Industry Classification System Industry 211, “Oil and Gas Extraction”



Note: Data are from 2012. The top five titles are labeled; see table 2 for a list of all titles.

Figure 8. Relevance of US Federal Agencies to North American Industry Classification System Industry 3116, "Animal Slaughtering and Processing"



Note: Thirteen other agencies are relevant but have industry relevance values of less than 0.01. Data are from 2012.

2.3.2. Limitations and solutions. One drawback of relying on NAICS industry descriptions to create search strings is that some search strings associated with a code are likely to occur more frequently for linguistic reasons unrelated to a title's relevance 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.

Another manifestation of this problem arises in industries bearing an NAICS name that is very uncommon or technical, especially compared to the words used in the CFR. For example, the "Credit Intermediation" industry (522) refers primarily to banking, but the word “intermediation” is used to avoid excluding savings institutions and credit unions. The string “credit intermediation” and its derivatives are quite uncommon in the CFR, even though Title 12 is called “Banks and Banking,” leading RegData to significantly understate the level of regulation in industry 522.

We have addressed this shortcoming in two ways. First, we have flagged those search strings that we deem likely to occur in irrelevant text (and therefore produce false positives), and we make this information available in the data on the website.

Second, we provide data on the probability of each search string occurring in written English and in legal English (written by lawyers). For written English, we used the data behind the Google Ngram Viewer to calculate these probabilities. The Google Ngram data offer counts of the number of times one- to five-word (one- to five-gram) strings were found in the Google Books corpora. We divide the number of occurrences of each search string in each year by the total count of Ngrams in the relevant corpus for that year to calculate the probability of each string occurring in written English. For the second—probabilities of a string occurring in legal text—we used the entire CFR (including appendixes and supplements) as our legal text corpus. As with the Google Books database, we calculated the number of times each search string was found in the legal text corpus and divided that by the total count of Ngrams in the corpus. These probabilities are given alongside each string in the downloadable datasets, and researchers can use these probabilities to weight search strings. Alternatively, humans can be employed to assess applicability for random subsets of occurrences of the words. In the interests of transparency and to promote fruitful experimentation, we make the entire database available along with string probabilities, and we invite users to customize the data in whatever way suits their purposes.

A related shortcoming that cannot be tackled by either of the methods described above concerns residual industries. A small subset of NAICS industries, many of which have a code that ends with a "9," usually start with one of the strings "other," "all other," "general," or "miscellaneous." For example, industry 4539 is "Other Miscellaneous Store Retailers." The current version of RegData is not equipped to handle these industries in a particularly enlightening way. One workaround is to combine the industry's "children" (though this option is not always available). For example, 4539 is composed of "Pet and Pet Supplies Stores" (45391), "Art Dealers" (45392), "Manufactured (Mobile) Home Dealers" (45393), and "All Other Miscellaneous Store Retailers" (45399). We create a dummy variable that flags those industries that are incompatible with RegData, and we exclude them from our analyses. They represent 10 of the 99 three-digit industries, and 37 of the 313 four-digit industries—that is, around 11 percent of NAICS industries.

Finally, the names of some NAICS industries result in a series of strings with zero hits in the CFR, possibly once RegData excludes terms that are deemed too widespread in contexts unrelated to the industry to warrant inclusion. For example, "Utility System Construction" (industry 2371) receives plenty of hits for strings such as "utility" and "system," but RegData does not allow them to count toward its industry relevance total since the numerous hits do not plausibly relate to the industry. As a result, it ends up with zero net hits. A primitive interpretation of RegData would imply that these industries have no text that is relevant to them in the year in question, but this interpretation is completely counter to common sense. To compensate for this problem, we classify any industry that yields zero hits in the entire CFR (possibly after string exclusion) as having missing data for industry relevance in each unit of the CFR rather than as having zero relevance. One can think of this choice as a "human correction" to the otherwise computerized algorithm for generating industry relevance. To give a sense of the prevalence of industries we have handled in this way, of the 89 “RegData compatible” three-digit industries, 78 of them (88 percent) result in positive industry relevance somewhere in the text and thus are not “missing observations.” The corresponding figure for the 276 “RegData compatible” four-digit industries is 218 (79 percent).

2.4. *Combining the Two Databases to Create a Panel*

Let i denote industry and y denote year; let I denote the set of industries and Y the set of years; let $S = \{I \times Y\}$. (In our case, I depends on which granularity of NAICS is used, while Y covers the period 1997–2012.) Title-, part-, and agency-specific measures of regulation—for example, restrictions or word counts—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 R_{py} 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 regulation receives in total regulations does not depend on the part, $R_y = \sum_p R_{py}$ is a measure of the total number of regulations in year y.

Let a_{pyi} 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 r_{pyi} measuring the regulations for industry i in part p in year y. The relationship will be of the form

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

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

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

alternatively, one could use a function of the form

$$f(a_{pyi}, R_{py}) = D(a_{pyi})R_{py},$$

where D is a dummy variable that takes the value 1 when a_{pyi} is above a threshold. Finally, assuming equal part weighting as above,

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

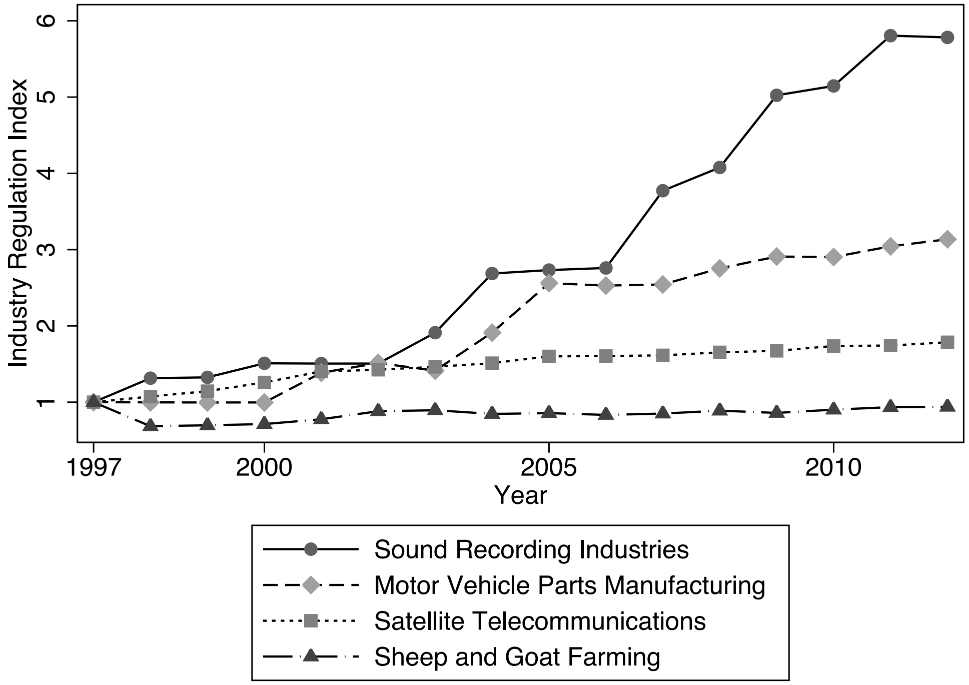
will be a measure of the regulations on industry i in year y. We provide

$$r_{pyi} = a_{pyi}R_{py}$$

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 a_{pyi} and R_{py}.

As an example, using our default, part-level method, figure 9 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 regulation since 1997. To give a broader look at the data, table 7 reports summary statistics for our default industry regulation level for a selection of two-, three-, and four-digit level industries.

Figure 9. 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.

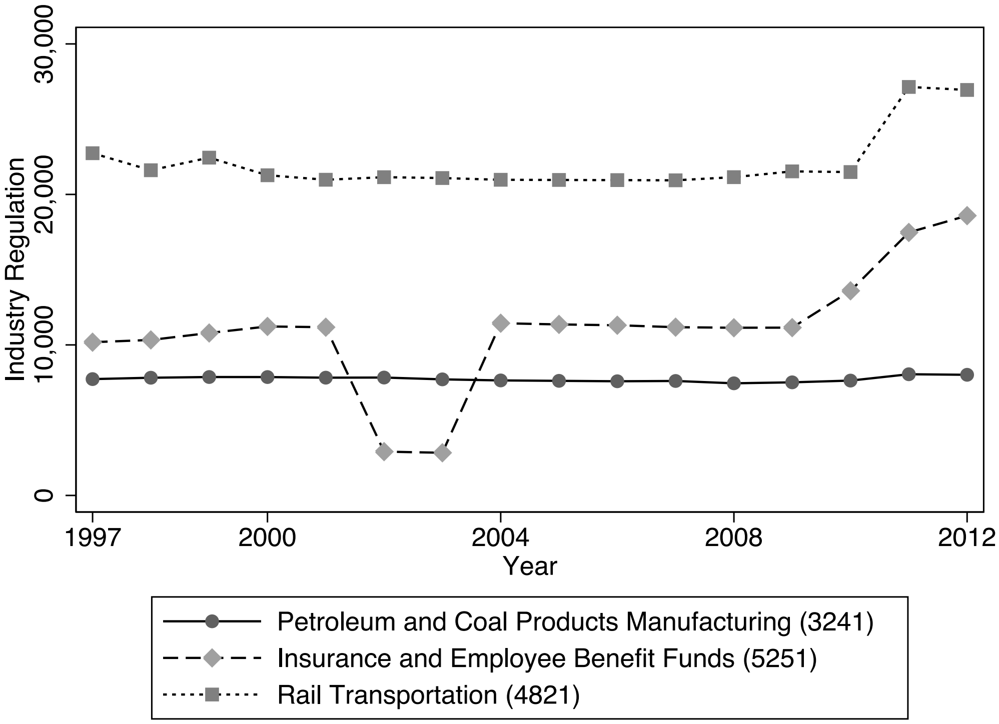
Table 7. Summary Statistics for the Regulation Level for a Selection of North American Industry Classification System Two-, Three-, and Four-Digit Industries

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).

RegData also permits the user to calculate how regulated an industry is by a specific department, agency, or set of departments and agencies. Figure 10, for example, shows agency-level regulation for three different industries. The agencies included in calculating the data used in figure 10 were agencies that are classified as creating "workplace" regulations in the Regulators' Budget (Dudley and Warren, 2013) (see the note below the figure for the list). All other metrics—words, restrictions, industry search terms, and industry relevance—are also available at the department and agency levels.

Figure 10. Industry Regulation 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. Those in the Department of Labor are the 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. The independent boards and commissions are the Architectural and Transportation Barriers Compliance Board, Equal Employment Opportunity Commission, National Labor Relations Board, and Occupational Safety and Health Review Commission.

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, though we still provide the user with all the choices.

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. We conduct an exploratory analysis in section 4.

3. Validation

RegData is a panel variable r_{iy} purportedly measuring regulation over a domain $S = \{I \times Y\}$. Assume that in principle one can indeed quantify regulation, and denote the resulting latent panel variable $\bar{r}_{iy}$. Assume that latent regulation is observable over some subset $\bar{S} \subset S$. Our claim is that r_{iy} is correlated with $\bar{r}_{iy}$. Can we validate such a claim? We have two broad validation strategies.

Ex-ante validation can be thought of as bottom-up and is deductive in nature. We explain how r_{iy} is constructed and deductively argue that it is likely to be correlated with $\bar{r}_{iy}$. It is the form of validation that we have employed thus far in the paper.

Ex-post validation is top-down and inductive. We treat the construction of r_{iy} as a black box and focus purely on the relationship between r_{iy} and $\bar{r}_{iy}$ over $\bar{S}$ (where $\bar{r}_{iy}$ is observable).

Needless to say, ex-post validation rests on the observability of the latent variable over some range. If, for example, we were to claim that we had devised a way of measuring the likelihood of a soccer team winning a game competing against aliens on Mars, then at the time of writing, we could only really validate such a claim ex ante.

In the case of regulation, we maintain that RegData is the first industry-specific panel series for regulation, and so one might initially conclude that ex-post validation is impossible. However, the existing data do in fact permit two limited forms of ex-post validation. First, there are discrete episodes of industry-specific regulation and deregulation that regulation experts generally acknowledge. We can examine the longitudinal aspects of our data to see whether they are consistent with these episodes. Second, Coates (2012) created a cross-sectional measure of regulation using the Fama and French (1997) 48-industry classification. We can collapse our panel into a cross section to compare it with the Coates-Fama-French regulation dataset.

A third, theoretically available validation option is to focus on the relationship between latent regulation, $\bar{r}$, and some other variable, z. If, for example, we are confident that despite its unobservability, latent regulation has a specific relationship to industry concentration, then we see if there is any evidence of the same relationship holding between RegData regulation and industry concentration and use such evidence to validate RegData.

To the best of our knowledge, this third validation option is not possible in our case due to the controversy over the relationship between an industry's *aggregate* regulation level and any other variable. Almost all economists would accept the law of demand, and such a law could be used to validate a new measure of the demand for a good. However, we are unaware of any relationship between regulation and another economic variable that economists would agree upon sufficiently to permit using it as a source of validation for the measure of regulation. If we were to focus on individual regulations, such as a health and safety regulation, then we could appeal to a consensus about its predicted (negative) effect on productivity. However, when we aggregate, we lose this ability because there are two convincing and somewhat opposing theories of regulation: regulatory capture (Stigler, 1971), which argues that regulation serves the interests of industry leaders often at the expense of competing groups such as workers, entrants, and other industries, and Pigouvian regulation (Pigou, 1938), which perceives regulation as the outcome of benevolent decision-making by politicians. For virtually any variable z that may be related to regulation, plausible models (invoking Pigou and Stigler, among others) can be put forward to explain a positive, negative, or nonexistent relationship between regulation and z.

Nevertheless, in section 4, we investigate the relationship between RegData regulation and some economic variables that are commonly thought to be related to it, such as employment and productivity, but these are not definitive demonstrations of RegData's accuracy as a measure of regulation because our priors on the signs and strengths of those relationships are weak at best.

3.1. *Ex-Post Validation Based on Episodic Regulation and Deregulation*

There are episodes of regulation or deregulation where a scholar could reasonably surmise that latent regulation, $\bar{r}_{iy}$, increased or decreased substantially for a certain industry over a certain period of time. In other words, certain episodes of regulation or deregulation are visible to the naked eye. For example, following the Clean Air Act, it is a stylized fact that several manufacturing industries experienced an increase in regulation. The subset $\bar{S}$ comprises regulation that is observable. (Dichotomizing the observability of $\bar{r}_{iy}$ is a substantial simplification since observability is more precisely expressed as a continuous variable, but it is useful for expositional simplicity.)

As intimated above, we do not have actual numerical data on $\bar{r}_{iy}$ over the subset $\bar{S}$; we can distill our knowledge of r_{iy} into M episodes of regulation or deregulation. Each episode is a quadruple (i, y_0, y_1, θ), where $\theta \in \{+, -\}$, y_0 denotes the (approximate) starting date of the regulation or deregulation episode, y_1 denotes the (approximate) ending date, and θ denotes whether the episode was regulation $(\theta = +)$ or deregulation $(\theta = -)$. In the case of regulation, $\bar{r}_{iy_1} > \bar{r}_{iy_0}$, and of deregulation, $\bar{r}_{iy_1} < \bar{r}_{iy_0}$. We can also assume weak monotonicity with respect to y:

$$\theta = + \Rightarrow \bar{r}_{iy} \geq \bar{r}_{ix}\ \forall\ (x, y) \in \{y_0, y_0 + 1, \dots, y_1 - 1, y_1\} \times \{y_0, y_0 + 1, \dots, y_1 - 1, y_1\} \text{ and } y > x$$

$$\theta = - \Rightarrow \bar{r}_{iy} \leq \bar{r}_{ix}\ \forall\ (x, y) \in \{y_0, y_0 + 1, \dots, y_1 - 1, y_1\} \times \{y_0, y_0 + 1, \dots, y_1 - 1, y_1\} \text{ and } y > x$$

Let $r_i^{y_0, y_1} = \left(r_{i,y_0}, r_{i,y_0+1}, \dots, r_{i,y_1-1}, r_{i,y_1}\right)$ and let $f\left(r_i^{y_0,y_1}\right) = +$ if $r_{iy_1} > r_{iy_0}$ and $-$ otherwise. Our ex-post validation strategy is a comparison of θ and $f\left(r_i^{y_0,y_1}\right)$ for each of the M triples (i, y_0, y_1) that $\bar{S}$ comprises. Equivalently, we will check if our purported measure of regulation, r_{iy}, actually describes a bout of regulation when conventional wisdom about real, latent regulation, $\bar{r}_{iy}$, describes a bout of regulation $(\theta = +)$ and conversely deregulation when conventional wisdom of $\bar{r}_{iy}$ describes a bout of deregulation $(\theta = -)$.

3.1.1. Nonstationarity and regulation. While RegData is the first industry-specific panel series on regulation, several aggregate measures already exist (Dawson and Seater, 2008; Crews, 2011). A common feature of all aggregate regulation series is that they describe increasing regulation at nearly all points in time, that is, regulation is nonstationary (according to the Office of the Federal Register, the CFR page count data series shows year-to-year total CFR pages increasing in 30 out of the 37 years since 1975). This description matches the general perception by economists and noneconomists alike that regulation is almost always increasing (Glaeser and Shleifer, 2003).

If it is indeed the case that most industry-level time series $\{\bar{r}_i\}$ describe increasing regulation, then weighting bouts of regulation $(\theta = +)$ the same as bouts of deregulation $(\theta = -)$ is a low-power strategy for testing the null hypothesis that RegData accurately measures latent regulation; after all, the number of computers being used in a given industry is almost certainly an increasing, nonstationary variable, and therefore as far as ex-post validation goes, it would be no worse than RegData, even though in terms of ex-ante validation it would seem utterly useless.

In the pursuit of testing power, therefore, we divide our ex-post validation strategy into two steps. The first is to check that the time series $\{r_i\}$ represents an increasing nonstationary variable for the overwhelming majority of industries. This component is a low-power way to check that RegData matches a coarse and vague perception that specialists and nonspecialists alike have about latent regulation.The second and more powerful step is to focus on what economists and legal experts perceive to have been bouts of deregulation and confirm that in these cases, $f\left(\hat{r}_i^{t_0,t_1}\right) = -$.

3.1.2. Data limitations. Deregulation in the United States is infrequent, and the most prominent bouts of deregulation all occurred before 1997. There exists a perception among laypeople that the financial sector was substantially deregulated in the late 1990s as part of the Gramm-Leach-Bliley Act, but a closer examination of the events suggests otherwise. The Gramm-Leach-Bliley Act eliminated the last vestiges of Glass-Steagall, pieces of which had already been eroded by regulatory supervisors. Glass-Steagall governed the relationship between banks and investment banks. However, no agencies were eliminated or radically downsized in the financial regulatory world until Dodd-Frank came along and eliminated the Office of Thrift Supervision in 2010 (and Dodd-Frank added several new agencies and offices). There were also some new regulatory initiatives in the 1990s like the Federal Deposit Insurance Corporation Improvement Act. See Calabria (2009) for more details.

The scarcity of episodes of deregulation poses a problem for our ex-post validation since the current version of RegData starts at 1997 (we intend to extend the data backward at a future point). To address this issue, we create hybrid pockets of RegData that cover some of the periods of alleged deregulation. The pockets are hybrid because current data limitations prevent us from applying the full RegData creation methods. We are in the process of acquiring character-recognizable electronic copies of the editions of the CFR published before 1997, but this process is as yet incomplete. Thus, we have to rely on measures of regulation at the industry level that depend only on page counts and not on the CFR's precise textual content.

We argued above that one cannot construct a mapping from the CFR's titles and volumes to industries in way that generates a convincing panel series of regulation. However, if one zooms down to the part level and links data produced at the part level to information about the agency in charge of producing the part, one can begin to link specific parts to specific industries. Such a strategy requires heavy input from the human eye, so the presence of thousands upon thousands of parts in the CFR means that this method is impractical for producing a broad panel. However, it is practical for looking at isolated industries for short periods of time.

Thus, suppose we convincingly argue that a substantial portion of the regulations governing industry i in time period $\{y_0, y_0+1, \dots, y_1-1, y_1\}$ are covered in a specific collection of CFR parts. Then we can use page counts for those parts to produce a hybrid measure of

$\{\bar{r}_{it}\}_{y=y_0}^{y=y_1}$, and this measure in turn permits us to perform an ex-post validation test along the lines described above.

In practice, this straightforward plan is complicated because government agencies mutate, cease to exist, or are replaced by new agencies, meaning that the links between parts and agencies are not always stable (setting aside the links between a part and an industry). Thus, a discerning human eye is required to identify possible mutations in agencies.

3.1.3. RegData and increasing regulation. The first step in our ex-post validation strategy is an examination of the trend in regulation by industry. RegData allows us to measure regulation at the economy level and at multiple NAICS levels. Table 8 shows the proportion of industries at each level that demonstrate an increase in regulation over the period 1997–2012 according to RegData. At all levels, RegData supports the stylized fact that regulation is rising in almost all industries. Thus, RegData passes the first, modest, test of ex-post validity.

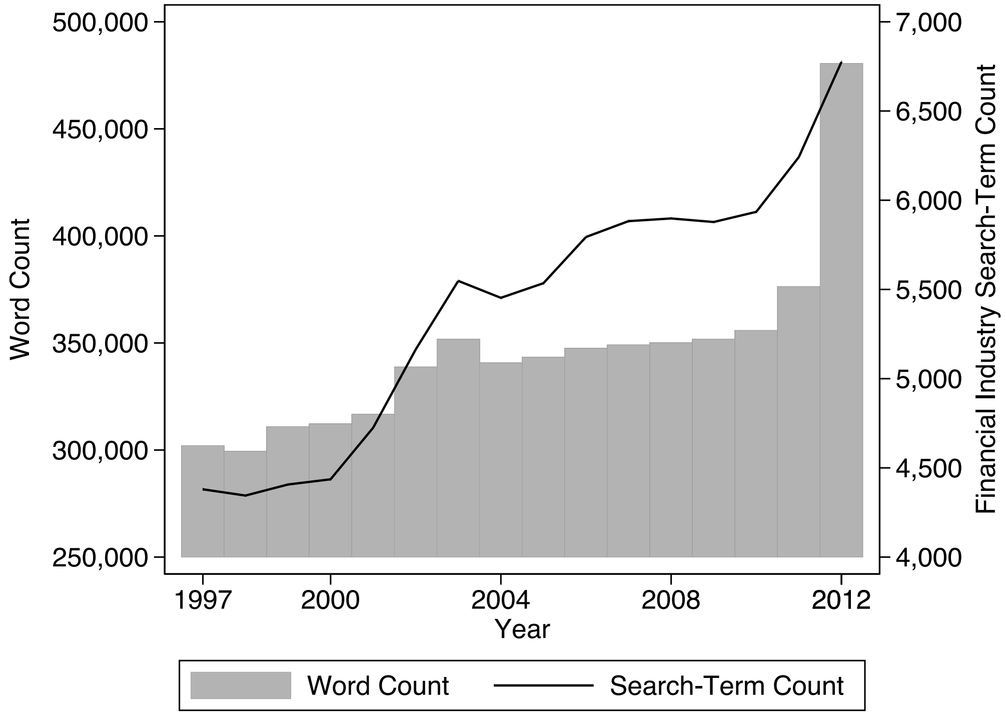
Table 8. North American Industry Classification System (NAICS) Industries Experiencing Regulatory Growth, 1997–2012, According to RegData

NAICS granularity	Two-digit	Three-digit	Four-digit
Number of industries	18	83	231
% of industries that exhibit more regulation in 2012 than in 1997	94.4	83.1	77.4

RegData also permits the examination of specific episodes widely considered to have created a large quantity of new financial regulations, such as the Dodd-Frank Act. McLaughlin and Greene (2012) used RegData methodology to estimate the total number of restrictions the Dodd-Frank Act would add to CFR titles 12 and 17, which the Securities and Exchange Commission and Commodity Futures Trading Commission (CFTC) publish, respectively. They estimate that, when finished, the rulemakings induced by Dodd-Frank will add approximately 26 percent more restrictions to those CFR titles, thereby showing that Dodd-Frank is a major regulatory event when measured using RegData.

Restriction counts, word counts, industry search terms, industry relevance, and the industry regulation index calculated from the regulatory text of specific agencies also may show agency responses to perceived crises. For example, figure 11 shows both word counts and industry search-term counts for financial industries for one of the major agencies, the CFTC, that oversees many industries at the heart of the most recent financial crisis. Although RegData does not yet attribute specific agency actions to their authorizing statutes (another feature that we hope to add in a future update), it seems plausible that the discrete jump in word count and search-term counts in 2011 and 2012 for financial industries is attributable to responses to the financial crisis. Those responses would at least partially be caused by the Dodd-Frank Act, which was signed into law on July 21, 2010.

Figure 11. Word Count and Financial Industry Search-Term Count, Commodity Futures Trading Commission



Note: This figure shows the word count of the Commodity Future Trading Commission's regulatory text (gray bars, measured on the left-hand axis), with the financial industry search-term count overlaid (black line, right-hand axis). The industries included in the financial industry search-term count are Agencies, Brokerages, and Other Insurance Related Activities; Funds, Trusts, and Other Financial Vehicles; Insurance Carriers; Insurance Carriers and Related Activities; Insurance and Employee Benefit Funds; Monetary Authorities—Central Bank; Securities and Commodity Contracts Intermediation and Brokerage; Securities and Commodity Exchanges; and Securities, Commodity Contracts, and Other Financial Investments and Related Activities.

3.1.4. The transport deregulation episodes of the late 1970s and early 1980s. We examine three known, major episodes of transportation deregulation: the Airline Deregulation Act of 1978, which deregulated the air transportation industry; the Staggers Rail Act of 1980, which deregulated the railroad industry; and the Motor Carrier Act of 1980, which deregulated the trucking industry. Since these episodes occurred before 1997, our first task is to produce the hybrid form of RegData for each the three industries. Then we can examine whether the data confirm the general perception of deregulation.

The process by which we produce the hybrid RegData is highly idiosyncratic at the industry and time-period levels, requiring a lot of industry-specific knowledge.[10] Trying to validate this process by, say, repeating it for the period 1997–2012 (when the original RegData is available) and comparing the hybrid RegData with the original RegData in a conventional statistical manner is not feasible, since too many details of the method of creating the hybrid data change with the time period. We thus rely on a mixture of first principles, detailed industry knowledge, and common sense.

The three acts largely targeted agencies responsible for setting rates and creating other so-called economic regulations related to the transport of goods and passengers by air, rail, or truck, respectively. For airlines, the responsible agency was the Civil Aeronautics Board, while for railroads and motor carriers, the Interstate Commerce Commission held sway.

We suspected that each of these industries was primarily regulated by one department publishing in one CFR title. For airlines, it was CFR Title 14, where the Civil Aeronautics Board and the Federal Aviation Administration both published. For railroads and motor carriers, we anticipated that CFR Title 49 mattered the most because that was where the Interstate Commerce Commission and relevant safety agencies published. Although we were unable to calculate industry relevance for each title for these years for the aforementioned technical reasons, we relied on editions of the CFR Index and Finding Guide to confirm that Titles 14 and 49 were the most important for the air transportation, railroad, and motor carrier industries.

The CFR Index and Finding Guide was first published in 1977. In its earliest form, the index was little more than a table of contents for agencies. It did not catalog the relevance of CFR parts to topics like "air transportation" except to point to a few agencies that regulated the

[10] One of this paper's authors, Patrick McLaughlin, has a background in applied transport regulation, furnishing him with the detailed knowledge necessary for the task.

industry. However, in 1979, the index became more comprehensive. That was the first year that the topics “air transportation,” “railroads,” and “motor carriers” listed dozens of CFR title-parts as relevant to the topics. We captured all the listings for those three topics in 1979 and again in 1982, and calculated the percentage of the total number of CFR parts listed under each topic that was published in Title 14 for air transportation and in Title 49 for railroads and motor carriers. Table 9 reports our results, confirming that the majority of relevant parts were in Title 14 for air transportation and Title 49 for rail transportation and motor carriers. Table 10 breaks down the number of parts listed in the 1979 and 1982 indexes according to the agency that published the part.

The deregulation of these industries was directly precipitated by the aforementioned three acts, and economists broadly consider it to have led to increased innovation, industry efficiency, and consumer welfare (see Winston [1998] for a further discussion of the effects of deregulation). In the case of railroads, deregulation also permitted the industry to transform from one returning less than 3 percent on equity in the 1970s to more than 8 percent in the 1990s (Winston, 1998). In other words, these deregulatory episodes were consequential, and they offered an opportunity to test whether our approach to measuring regulation would show deregulation as such.

Table 9. Transport Relevance of Title 14 and Title 49

	1979	1982
% of air transportation–relevant parts in Title 14	0.58	0.66
% of railroad-relevant parts in Title 49	0.60	0.59
% of motor carrier-relevant parts in Title 49	0.65	0.62

Table 10. Parts and Publishing Agencies in the 1979 and 1982 *Code of Federal Regulations* Index and Finding Guides

	Agency	Agency parts	Count in index, 1979	Count in index, 1982
Air transportation	Federal Aviation Administration, DOT*	1–199	60	60
	Civil Aeronautics Board	200–399	14	59
	National Aeronautics and Space Administration	1201–END	2	2
	All others		56	63
Motor carriers	Office of the Secretary of Transportation	0–99	0	0
	Research and Special Programs Administration	100–199	2	2
	Federal Railroad Administration	200–299	0	0
	Federal Highway Administration	300–399	13	15
	Coast Guard	400–499	3	3
	National Highway Traffic Safety Administration	500–599	30	32
	Urban Mass Transportation Administration	600–699	3	3
	National Railroad Passenger Corporation	700–799	Did not exist	0
	National Transportation Safety Board	800–899	0	0
	United States Railway Association	900–999	0	0
	Interstate Commerce Commission	1000–1399	61	62
	All others		62	72
Railroads	Office of the Secretary of Transportation	0–99	1	1
	Research and Special Programs Administration	100–199	2	2
	Federal Railroad Administration	200–299	30	33
	Federal Highway Administration	300–399	0	0
	Coast Guard	400–499	0	0
	National Highway Traffic Safety Administration	500–599	0	0
	Urban Mass Transportation Administration	600–699	1	1
	National Railroad Passenger Corporation	700–799	Did not exist	0
	National Transportation Safety Board	800–899	1	1
	United States Railway Association	900–999	3	4
	Interstate Commerce Commission	1000–1399	62	57
	All others		66	71

* Department of Transportation.

3.1.4.1. Air transportation. The Airline Deregulation Act was the first of the three acts. Its primary target, the Civil Aeronautics Board, published regulations in parts 200 to 1199 of CFR Title 14. The act may have also targeted the Federal Aviation Administration, but to a lesser degree, because this administration was not responsible for economic regulations. As table 11 shows, in years 1974–1975, the Civil Aeronautics Board published in Volume 3 alongside the National Transportation Safety Board and the National Aeronautics and Space Administration. In 1976, the National Transportation Safety Board's regulations were transferred to Title 49, and the National Aeronautics and Space Administration was given its own binding. These changes left the Civil Aeronautics Board as the only agency publishing in Volume 3 until 1981. In 1981, another volume was added to Title 14 to accommodate the growth of the Federal Aviation Administration's regulations. As a result, the Civil Aeronautics Board's pages were pushed into Volume 4, while Volumes 1–3 were reserved for the Federal Aviation Administration, and the National Aeronautics and Space Administration's regulations were moved to Volume 5. This configuration was maintained until 1985, when the Civil Aeronautics Board was formally dissolved. The regulations that remained after the deregulatory actions induced by the Airline Deregulation Act were transferred to the Office of the Secretary of the Department of Transportation, which kept them in Volume 4 of Title 14.

Table 11. Parts and Publishing Agencies for Air Transportation

Years	Vol.	Volume chapters	Volume parts	Agency*	Agency chapters	Agency parts
1974–1975	1	I	1–59	Federal Aviation Administration, DOT	I	1–199
1974–1975	2	I	60–199	Federal Aviation Administration, DOT	I	1–199
1974–1975	3	II, III, V	200–END	**Civil Aeronautics Board**	II	200–399
1974–1975	3	II, III, V	200–END	National Transportation Safety Board	III	400–1199
1974–1975	3	II, III, V	200–END	National Aeronautics and Space Administration	V	1200–END
1976–1981	1	I	1–59	Federal Aviation Administration, DOT	I	1–199
1976–1981	2	I	60–199	Federal Aviation Administration, DOT	I	1–199
1976–1981	3	II	200–1199	**Civil Aeronautics Board**	II	200–399
1976–1981	4	V	1200–END	National Aeronautics and Space Administration	V	1201–END
1982–1984	1	I	1–59	Federal Aviation Administration, DOT	I	1–199
1982–1984	2	I	60–138	Federal Aviation Administration, DOT	I	1–199
1982–1984	3	I	140–199	Federal Aviation Administration, DOT	I	1–199
1982–1984	4	II	200–1199	**Civil Aeronautics Board**	II	200–399
1982–1984	5	V	1200–END	National Aeronautics and Space Administration	V	1201–END
1985	1	I	1–59	Federal Aviation Administration, DOT	I	1–199
1985	2	I	60–138	Federal Aviation Administration, DOT	I	1–199
1985	3	I	140–199	Federal Aviation Administration, DOT	I	1–199
1985	4	II	200–1199	**Office of the Secretary, DOT**	II	200–399
1985	5	V	1200–END	National Aeronautics and Space Administration	V	1201–END

* DOT stands for Department of Transportation.

Because the Civil Aeronautics Board's regulations were published in a single volume from 1976 onward, we were able to examine how the total number of pages of regulations published by this agency changed around the time of deregulation by looking at page totals for the volume. We also compared the number of pages published by the Civil Aeronautics Board with the page counts of the other agencies publishing in Title 14 in this period. However, two other major changes to Title 14, besides deregulation of air transportation, occurred over this period and affect our comparisons. First, because the National Transportation Safety Board was removed from Title 14 at the end of 1975, and it had previously been included in Volume 3, that volume loses a significant number of pages between 1975 and 1976. Second, the National Aeronautics and Space Administration's pages were removed from Volume 3 in 1976, at which point its regulations were published in a separate volume.

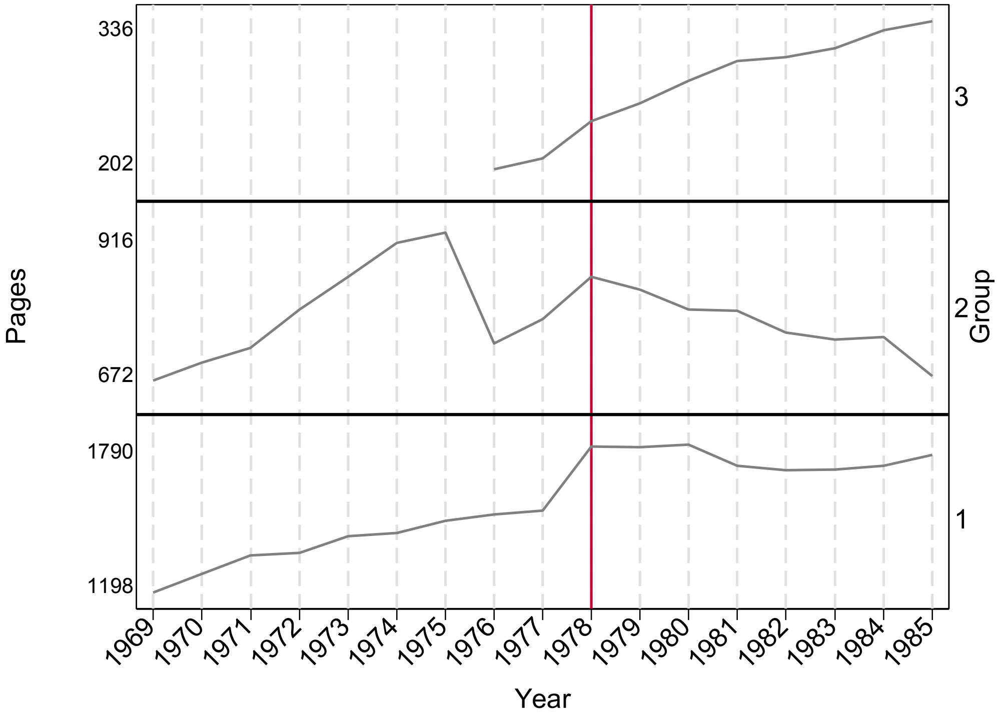
Despite the confounding factors that caused a decrease in pages in Volume 3 at the end of 1975, we examined the period from 1969 to 1985. Because of the reconfigurations of the volumes that occur over the period, we organized the agencies into groups based on the original configuration. Table 12 shows the three groups we created for the sake of comparison. Group 1 contains parts 1–199, which are published by the Federal Aviation Administration, throughout the period. Group 2 initially comprised all the agencies that published in the 1974 Volume 3, which included the Civil Aeronautics Board, the National Transportation Safety Board, and the National Aeronautics and Space Administration. However, the reconfiguration at the end of 1975 left the Civil Aeronautics Board as the sole agency in its volume, so the agency of primary interest can be carefully examined before, during, and after the deregulation caused by the 1978 Airline Deregulation Act. Group 3 did not exist until the National Aeronautics and Space Administration's pages were taken out of Volume 3 and placed in a new volume in 1976. For the remainder of the period, Group 3 consists of the National Aeronautics and Space Administration's pages.

Table 12. Parts and Agencies for Air Transportation

Group	Years	Group parts	Agency
1	1974–1975	1–199	Federal Aviation Administration, Department of Transportation
1	1976–1981	1–199	Federal Aviation Administration, Department of Transportation
1	1982–1984	1–199	Federal Aviation Administration, Department of Transportation
1	1985	1–199	Federal Aviation Administration, Department of Transportation
2	1974–1975	200–END	Civil Aeronautics Board, National Transportation Safety Board, National Aeronautics and Space Administration
2	1976–1981	200–1199	Civil Aeronautics Board
2	1982–1984	200–1199	Civil Aeronautics Board
2	1985	200–1199	Office of the Secretary of Transportation
3	1974–1975	None	None
3	1976–1981	1200–END	National Aeronautics and Space Administration
3	1982–1984	1200–END	National Aeronautics and Space Administration
3	1985	1200–END	National Aeronautics and Space Administration

We recorded the total number of pages published in the final version of the CFR in each year from 1969 to 1985 for all volumes in Title 14. We then grouped the volumes' pages according to the scheme described in table 12. The total number of pages in each of these groups in each year over the period is shown below in figure 12.

Figure 12. Page Counts of Title 14's Air Transportation Regulatory Agencies by Groups Specified in Table 11

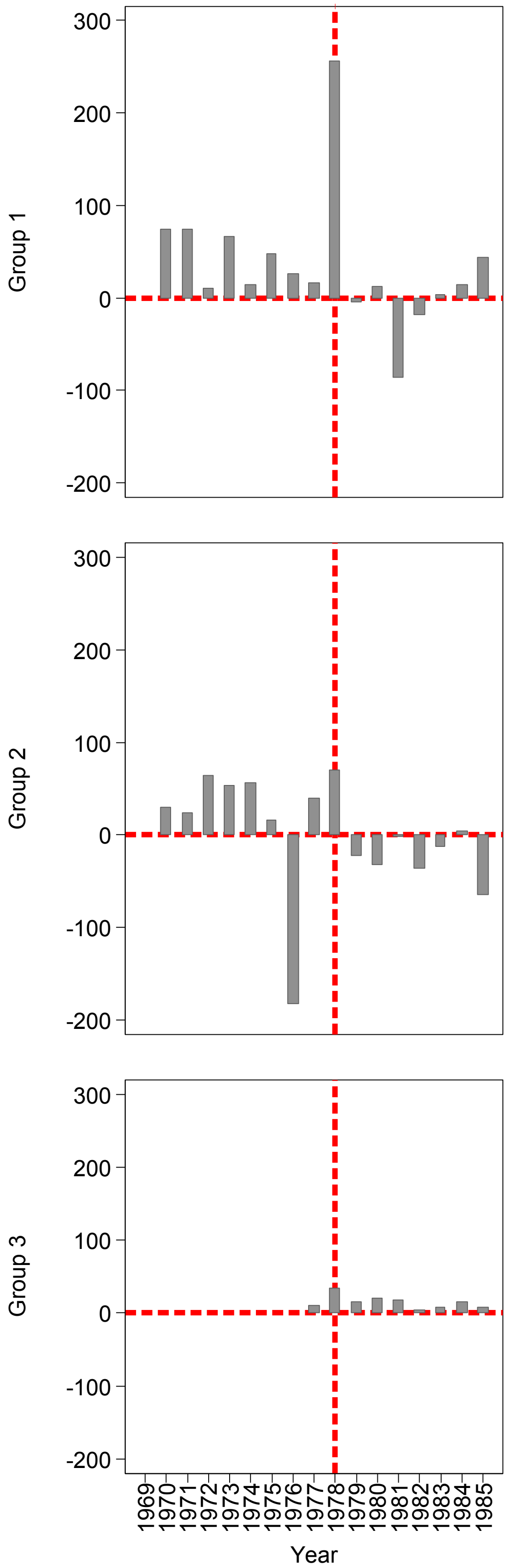


Note: The vertical line shows the passage of the Airline Deregulation Act of 1978.

Regulatory actions stemming from legislation can take several years. A search of the *Federal Register* for rulemakings related to the Airline Deregulation Act of 1978 resulted in a list of rulemakings that spanned at least five subsequent years. So it should not be expected that the act would result in an immediate reduction of Civil Aeronautics Board pages, but instead that the trend following the 1978 act would turn negative (or more negative, if it was already negative). Figure 12 shows a clear, negative trend for Group 2 from 1978 onward, while the other two groups either stay relatively static or grow. It is reasonable to conclude that the Airline Deregulation Act precipitated this change in Group 2.

We also have charted the change in total pages in each group from year to year, shown in figure 13. In this figure, each bar represents the number of pages added to or deleted from a group's total when compared with the previous year. Again, we see the consistent negative trend

Figure 13. Title 14 Year-to-Year Changes in Pages by Group and Year



in Group 2 after 1978. Recall that 1976 witnessed the reconfiguration of Title 14, with the National Transportation Safety Board being transferred to Title 49 and the National Aeronautics and Space Administration's pages being shifted to a new volume. This reconfiguration explains the loss of pages shown in 1976 for Group 2. Other than that confounding (and nonregulatory) action, the simplest explanation for the patterns seen in Group 2 appears to be the Airline Deregulation Act of 1978, shown with the vertical line. It is also noteworthy that in 1981 and 1982, Group 1 sees a decrease in pages. Group 1 contained the Federal Aviation Administration's regulations, and it is possible that the Airline Deregulation Act affected those regulations as well.

3.1.4.2. Railroads and trucking. In 1980, two acts of Congress resulted in the removal or modification of many economic regulations of the railroad and motor carrier industries. These were the Staggers Rail Act of 1980 and the Motor Carrier Act of 1980. The target of these acts was the Interstate Commerce Commission, which had been responsible for setting rates, among other economic regulatory roles, in these transportation industries. Its regulations were published in Title 49 (Transportation).

In 1976, Title 49 (Transportation) comprised six volumes. Table 13 lists the various configurations used in Title 49 over the period 1976–1985. Volume 1 contained rules from the Office of the Secretary and Volume 2 contained rules from the Material Transportation Bureau. Volume 3 included several agencies' rules: the Federal Railroad Administration, Federal Highway Administration, Coast Guard, National Highway Traffic Safety Administration, Urban Mass Transportation Administration, National Transportation Safety Board, and United States Railway Association. The Interstate Commerce Commission took up the remaining three volumes, Volumes 4–6. This configuration lasted through 1978.

Table 13. Parts and Agencies for Railroad and Truck Transportation

Years	Vol.	Vol. subtitles/ chapters	Volume parts	Agency	Agency sub-titles/chapters	Agency parts
1976–1978	1	Subtitle A	1–99	Office of the Sec. of Transportation	Subtitle A	0–99
1976–1978	2	Chapter I	100–199	Materials Transportation Bureau	Chapter I	100–199
1976–1978	3	Chapter II–IX	200–999	Fed. Railroad Admin.	Chapter II	200–299
1976–1978	3	Chapter II–IX	200–999	Fed. Highway Admin.	Chapter III	300–399
1976–1978	3	Chapter II–IX	200–999	Coast Guard	Chapter IV	400–499
1976–1978	3	Chapter II–IX	200–999	Natl. Highway Traffic Safety Administration	Chapter V	500–599
1976–1978	3	Chapter II–IX	200–999	Urban Mass Transportation Admin.	Chapter VI	600–699
1976–1978	3	Chapter II–IX	200–999	National Transportation Safety Board	Chapter VIII	800–899
1976–1978	3	Chapter II–IX	200–999	US Railway Association	Chapter IX	900–999
1976–1978	4	Chapter X	1000–1199	Interstate Commerce Commission	Chapter X	1000–1399
1976–1978	5	Chapter X	1200–1299	Interstate Commerce Commission	Chapter X	1000–1399
1976–1978	6	Chapter X	1300–END	Interstate Commerce Commission	Chapter X	1000–1399

continued on next page

Years	Vol.	Vol. subtitles/ chapters	Volume parts	Agency	Agency sub-titles/chapters	Agency parts
1979–1981	1	Subtitle A	1–99	Office of the Sec. of Transportation	Subtitle A	0–99
1979–1981	2	Chapter I	100–177	Research and Special Programs Admin.	Chapter I	100–199
1979–1981	3	Chapter I	178–199	Research and Special Programs Admin.	Chapter I	100–199
1979–1981	4	Chapter II–III	200–399	Fed. Railroad Admin.	Chapter II	200–299
1979–1981	4	Chapter II–III	200–399	Fed. Highway Admin.	Chapter III	300–399
1979–1981	5	Chapter IV–IX	400–999	Coast Guard	Chapter IV	400–499
1979–1981	5	Chapter IV–IX	400–999	Natl. Highway Traffic Safety Administration	Chapter V	500–599
1979–1981	5	Chapter IV–IX	400–999	Urban Mass Transportation Admin.	Chapter VI	600–699
1979–1981	5	Chapter IV–IX	400–999	National Transportation Safety Board	Chapter VIII	800–899
1979–1981	5	Chapter IV–IX	400–999	US Railway Association	Chapter IX	900–999
1979–1981	6	Chapter X	1000–1199	Interstate Commerce Commission	Chapter X	1000–1399
1979–1981	7	Chapter X	1200–1299	Interstate Commerce Commission	Chapter X	1000–1399
1979–1981	8	Chapter X	1300–END	Interstate Commerce Commission	Chapter X	1000–1399
1982–1985	1	Subtitle A	1–99	Office of the Sec. of Transportation	Subtitle A	0–99
1982–1985	2	Chapter I	100–177	Research and Special Programs Admin.	Chapter I	100–199
1982–1985	3	Chapter I	178–199	Research and Special Programs Admin.	Chapter I	100–199
1982–1985	4	Chapter II–III	200–399	Fed. Railroad Admin.	Chapter II	200–299
1982–1985	4	Chapter II–III	200–399	Fed. Highway Admin.	Chapter III	300–399
1982–1985	5	Chapter IV–IX	400–999	Coast Guard	Chapter IV	400–499
1982–1985	5	Chapter IV–IX	400–999	Natl. Highway Traffic Safety Administration	Chapter V	500–599
1982–1985	5	Chapter IV–IX	400–999	Urban Mass Transportation Admin.	Chapter VI	600–699
1982–1985	5	Chapter IV–IX	400–999	National Railroad Passenger Corporation (Amtrak)	Chapter VII	700–799
1982–1985	5	Chapter IV–IX	400–999	National Transportation Safety Board	Chapter VIII	800–899
1982–1985	5	Chapter IV–IX	400–999	US Railway Association	Chapter IX	900–999
1982–1985	6	Chapter X	1000–1199	Interstate Commerce Commission	Chapter X	1000–1399
1982–1985	7	Chapter X	1200–1299	Interstate Commerce Commission	Chapter X	1000–1399
1982–1985	8	Chapter X	1300–END	Interstate Commerce Commission	Chapter X	1000–1399

In 1979, two new volumes were added to Title 49. These two new volumes split up 1978's Volume 3 into 1979's Volumes 3, 4, and 5. As before, the Interstate Commerce Commission used the final three volumes.

In 1982, the National Railroad Passenger Corporation (Amtrak) was created, and the regulations related to Amtrak were inserted into the previously empty parts 700–799 of Volume 5. Again, the Interstate Commerce Commission printed in Volumes 6–8. Because both the Staggers Rail Act and the Motor Carrier Act targeted regulations from the Interstate Commerce Commission, we examined the pages of rules printed in the Interstate Commerce Commission's volumes over the 10-year period, 1976–1985, surrounding the acts' passage in 1980. For the sake of comparison with other Department of Transportation rules, we also examined the pages printed in the other volumes. We arranged the volumes into six groups based on the original 1976 configuration. As more volumes were added to Title 49, we assigned the new volumes to the group that seemed most appropriate based on the part numbers and agencies publishing in the new volumes. Table 14 lists the groups, the parts contained in the groups, and the agencies publishing in the parts and thus in each group.

Table 14. Parts and Agencies for Railroad and Truck Transportation

Group	Years	Group parts	Agency
1	1976–1978	1–99	Office of the Secretary of Transportation
1	1979–1981	1–99	Office of the Secretary of Transportation
1	1982–1985	1–99	Office of the Secretary of Transportation
2	1976–1978	100–199	Materials Transportation Bureau
2	1979–1981	100–199	Research and Special Programs Administration
2	1982–1985	100–199	Research and Special Programs Administration

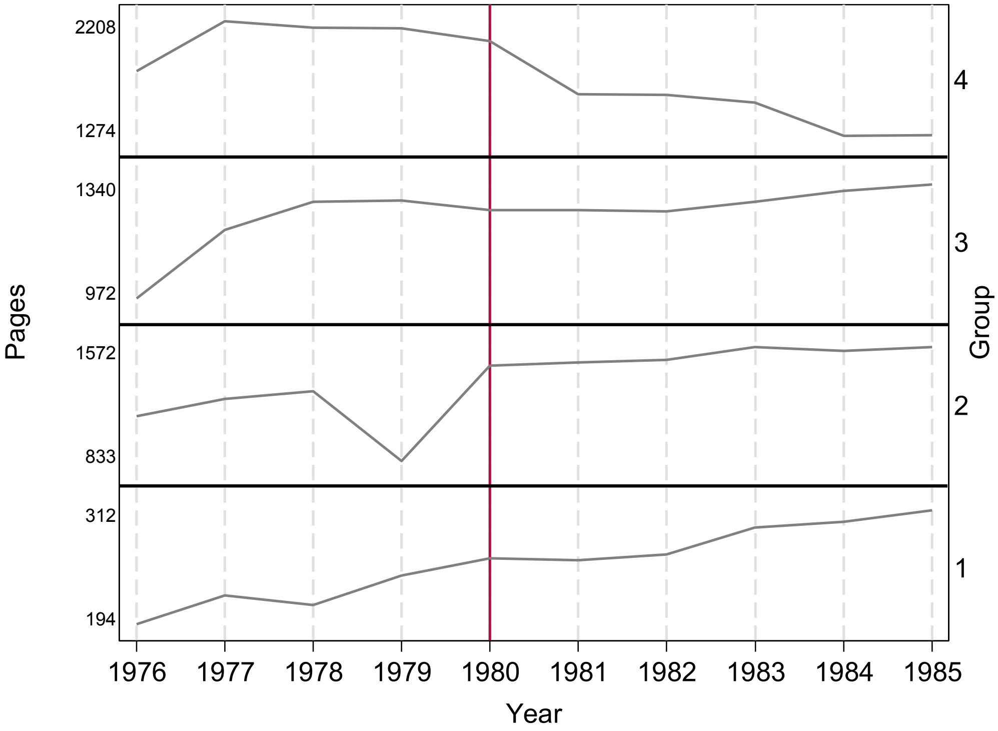
continued on next page

Group	Years	Group parts	Agency
3	1976–1978	200–999	Federal Railroad Administration
3	1976–1978	200–999	Federal Highway Administration
3	1976–1978	200–999	Coast Guard
3	1976–1978	200–999	National Highway Traffic Safety Administration
3	1976–1978	200–999	Urban Mass Transportation Administration
3	1976–1978	200–999	National Transportation Safety Board
3	1976–1978	200–999	United States Railway Association
3	1979–1981	200–999	Federal Railroad Administration
3	1979–1981	200–999	Federal Highway Administration
3	1979–1981	200–999	Coast Guard
3	1979–1981	200–999	National Highway Traffic Safety Administration
3	1979–1981	200–999	Urban Mass Transportation Administration
3	1979–1981	200–999	National Transportation Safety Board
3	1979–1981	200–999	United States Railway Association
3	1982–1985	200–999	Federal Railroad Administration
3	1982–1985	200–999	Federal Highway Administration
3	1982–1985	200–999	Coast Guard
3	1982–1985	200–999	National Highway Traffic Safety Administration
3	1982–1985	200–999	Urban Mass Transportation Administration
3	1982–1985	200–999	National Railroad Passenger Corporation (Amtrak)
3	1982–1985	200–999	National Transportation Safety Board
3	1982–1985	200–999	United States Railway Association
4	1976–1978	1000–1199	Interstate Commerce Commission
4	1979–1981	1000–1199	Interstate Commerce Commission
4	1982–1985	1000–1199	Interstate Commerce Commission
5	1976–1978	1200–1299	Interstate Commerce Commission
5	1979–1981	1200–1299	Interstate Commerce Commission
5	1982–1985	1200–1299	Interstate Commerce Commission
6	1976–1978	1300–END	Interstate Commerce Commission
6	1979–1981	1300–END	Interstate Commerce Commission
6	1982–1985	1300–END	Interstate Commerce Commission

Because Title 49 is more voluminous than Title 14, and because it saw more agencies added to it or taken away from it during this period, dividing the title into comparable groups is slightly more complicated. Group 1 consisted of the Office of the Secretary of Transportation's regulations throughout the period. Group 2 was initially composed of the Materials Transportation Bureau, but, in 1978, the Transportation Systems Center and three other subagencies previously under the Office of the Secretary of Transportation were consolidated with the Materials Transportation Bureau into the Research and Special Programs Administration, which continued publishing where the Materials Transportation Bureau had been. Group 3 contained regulations from the Federal Railroad Administration, Federal Highway Administration, Coast Guard, National Highway Traffic Safety Administration, Urban Mass Transportation Administration, National Transportation Safety Board, and United States Railway Association, because this group jointly shared Volume 3 in the initial 1976 configuration. In 1982, Amtrak was added to this group. Finally, Group 4 contained only the Interstate Commerce Commission's regulations during the entire period.

We calculated the total number of pages published in each of these groups in each year from 1976 to 1985. We expected the deregulatory Staggers Rail Act of 1980 and the Motor Carrier Act of 1980 to diminish the total page count of Group 4, which consisted entirely of Interstate Commerce Commission regulations. In contrast, we expected Groups 1, 2, and 3 to continue along more or less the same pattern throughout the period. Figure 14 shows the total pages of each group over this 10-year period.

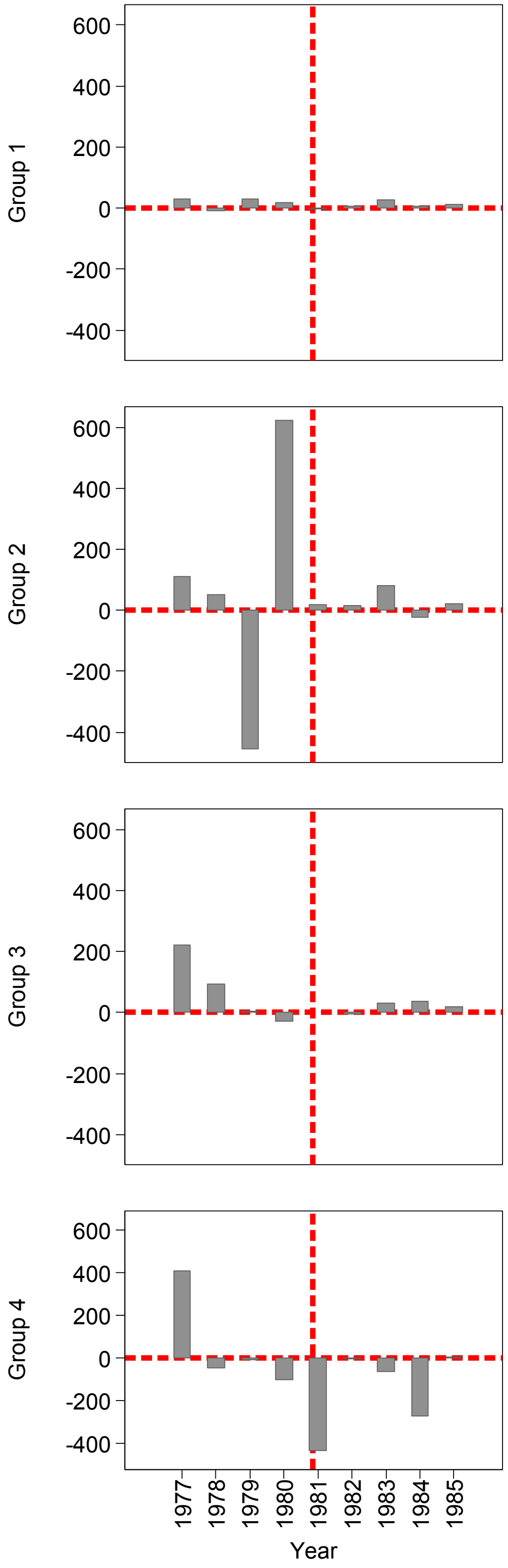
Figure 14. Page Counts of Title 49's Railroad and Motor Carrier Transportation Regulatory Agencies by Groups Specified in Table 14



Note: The vertical line shows the passage of the Staggers Rail Act of 1980 and the Motor Carrier Act of 1980.

As anticipated, Group 4 exhibits a strong, negative trend following the passage of the Staggers Rail Act and the Motor Carrier Act in 1980. Figure 15 shows the changes in pages for each group over this period. Each bar shows how many pages were added or subtracted relative to the previous year. Group 2 exhibits some outstanding additions and subtractions likely related to the restructuring of the Materials Transportation Bureau into the Research and Special Programs Administration. Other than that, the only substantial negative changes occur in Group 4, and mostly after the 1980 threshold.

Figure 15. Title 49 Year-to-Year Changes in Pages by Group and Year



3.2. *Ex-Post Validation Based on Coates's Cross-Sectional Data*

NAICS's forerunner was the Standard Industrial Classification (SIC) system. Fama and French (1997), as part of their research on equity prices, divided the economy into a nonexhaustive list of 48 industries, where each industry was a composite of a selection of SIC industries. Coates (2012) subsequently created a new data series crudely measuring regulation in the 48 industries. Each industry was classified as unregulated, moderately regulated, or heavily regulated. We have been unable to definitively determine the methods used for establishing the level of regulation in an industry.

In Coates (2012), the cross-sectional regulation data are used in a panel that covers the time range 1998–2010. In principle, if we match Fama-French industries to NAICS industries, we can collapse the RegData panel into a cross section for the period 1998–2010 and compare it with the Fama-French-defined industries' regulation data created by Coates (hereafter called Coates-Fama-French regulation). Assuming that the Coates-Fama-French regulation is indeed an accurate measure of the regulation, then a positive relationship between Coates-Fama-French regulation and RegData would constitute ex-post validation of RegData.

3.2.1. Matching Fama-French industries to NAICS industries. Similar to the NAICS system, SIC industries come in different granularities, with the finest being four-digit. There exist well-developed, officially endorsed techniques for transforming SIC industries to NAICS industries. However, Fama and French (1997) combine the finest grain of SIC-defined industries in a way that does not correspond to the coarser grains; that is, neither SIC nor Fama-French are nested within the other. For example, Fama-French "Agriculture" and Fama-French "Beer and Liquor" intersect with both SIC "Agricultural Production Crops" and SIC "Manufacturing."

This inconsistency alone is not particularly problematic since no four-digit SIC industry corresponds to more than one Fama-French industry. However, there are occasions where more than one four-digit SIC industry corresponds to the same NAICS industry. More importantly, RegData is applied at the four-digit NAICS level (see footnote 9 for an explanation), meaning that *many* four-digit SIC industries from different Fama-French groups correspond to the same four-digit NAICS industry.

For example, "Paving and Roofing Materials" (SIC 2951-2952, part of Fama-French "Construction Materials") and "Petroleum Refining" (SIC 2900-2912, part of Fama-French "Petroleum and Natural Gas") both correspond to "Petroleum and Coal Products Manufacturing" under four-digit NAICS (3241), among other NAICS industries. Should NAICS 3241 be included in both "Construction Materials" and "Petroleum Refining," one of the two, or neither? Perhaps it should be weighted? To add to the complication, some Fama-French industries match up well to combinations of two-, three-, and four-digit NAICS industries, while others match up to exclusively two-, three- or four-digit industries. Is straightforward aggregation appropriate, or is some weighting in order?

These technicalities do not render futile the exercise of transforming Fama-French to NAICS, but they do introduce subjectivity. Since there are at most 48 Fama-French industries, and therefore at most 48 observations, it is almost inevitable that the results will be sensitive to some of the researcher's judgment calls. It is therefore prudent to put comparatively little weight on these data as a source of validation. Table 15 contains the summary data, including the Coates-Fama-French regulation designations according to our judgment of how best to match Fama-French industries to NAICS industries.

Table 15. Coates-Fama-French vs. RegData Industry Matches and Regulation Data

Coates-Fama-French			RegData			
Code	Industry	Reg.	Code	Industry	Reg. (gross)	Reg. (weight.)
5	Tobacco Products	High	3122	Tobacco Mfg.	79,000	79,000
13	Pharmaceutical Products	High	3254	Pharma. & Medicine Mfg.	3,600	3,600
24	Aircraft	High	3364	Aerospace Products & Parts Mfg.	18	18
31	Utilities	High	221	Utilities	100,000	31,000
32	Communication	High	517	Telecommunications	49,000	14,000
40	Transportation	High	48	Trans. & Warehousing	580,000	60,000
44	Banking	High	522	Credit Intermediation	73	22
45	Insurance	High	5241	Insurance Carriers	240,000	240,000
27	Precious Metals	Med.	2122	Metal Ore Mining	79	79
30	Petroleum and Natural Gas	Med.	211	Oil & Gas Extraction	430,000	130,000
47	Trading	Med.	523	Securities, Commodities, Etc.	380,000	110,000
1	Agriculture	Low	11	Ag., For., Fish. & Hunt.	230,000	24,000
2	Food Products	Low	311	Food Mfg.	250,000	75,000
3	Candy & Soda	Low	3113	Sugar & Confec. Prod. Mfg.	300	300
6	Recreation	Low	3343	A/V Equip. Mfg.	1,600	1,600
7	Entertainment	Low	5121	Motion Picture & Video Ind.	5,500	2,800
8	Printing and Publishing	Low	3231	Print. & Related Support	26,000	26,000
10	Apparel	Low	3152	Cut & Sew Apparel Mfg.	27	27
11	Healthcare	Low	62	Health Care & Social Assist.	34,000	3,500
12	Medical Equipment	Low	3391	Medical Equip. & Suppl. Mfg.	2,300	2,300
14	Chemicals	Low	325	Chemical Mfg.	130,000	39,000
15	Rubber and Plastic Products	Low	326	Plastics & Rubber Prod. Mfg.	46,000	14,000
16	Textiles	Low	313	Textile Mills	15,000	4,600
17	Construction Materials	Low	327	Nonmetallic Mineral Prod. Mfg.	100,000	30,000
18	Construction	Low	23	Construction	290,000	30,000
19	Steel Works, Etc.	Low	331	Primary Metal Mfg.	83,000	25,000
20	Fabricated Products	Low	332	Fabricated Metal Prod. Mfg.	94,000	28,000
21	Machinery	Low	333	Machinery Mfg.	41,000	12,000
22	Electrical Equipment	Low	335	Electr. Equip. . . . Mfg.	30,000	9,100
23	Automobiles and Trucks	Low	3363	Motor Vehicle Parts Mfg.	180	180
25	Shipbuilding, Railroad Equip.	Low	3365	Railroad Rolling Stock Mfg.	690	690
28	Non-Metal. & Ind. Metal Min.	Low	2123	Nonmetal Min. & Quarry	800	800
29	Coal	Low	2121	Coal Mining	96,000	96,000
33	Personal Services	Low	812	Personal & Laundry Services	15,000	4,600
35	Computers	Low	3341	Comp. & Periph. Equip. Mfg.	440	440
36	Electronic Equipment	Low	3344	Semiconductor . . . Mfg.	2,500	2,500
37	Measuring & Control Equip.	Low	3345	Control Instruments Mfg.	510	510
38	Business Supplies	Low	322	Paper Manufacturing	66,000	20,000
39	Shipping Containers	Low	3324	Boiler, Tank & Ship. Cont. Mfg.	660	660
41	Wholesale	Low	42	Wholesale Trade	290,000	29,000
42	Retail	Low	44	Retail Trade	300,000	31,000
43	Restaurants, Hotels, Motels	Low	72	Accommodation & Food Services	270,000	28,000
46	Real Estate	Low	531	Real Estate	24,000	7,100

Note: Gross regulation is unweighted; weighted regulation adjusts by a factor so that two-digit, three-digit, and four-digit industries all share the same mean level of regulation. Some Coates-Fama-French industries correspond to two NAICS industries, but the table only reports the primary one; the secondary ones are Recreation (1125, 3343), Entertainment (711, 5121), Printing and Publishing (5111, 3231), Apparel (3152, 3162), Construction Materials (327, 332), Automobiles and Trucks (3362, 3363), Shipbuilding, Railroad Equipment (3365, 3366), Petroleum and Natural Gas (211, 213), Personal Services (811, 812), Electronic Equipment (3342, 3344), Business Supplies (322, 337), Shipping Containers (3222, 3324), and Trading (523, 533).

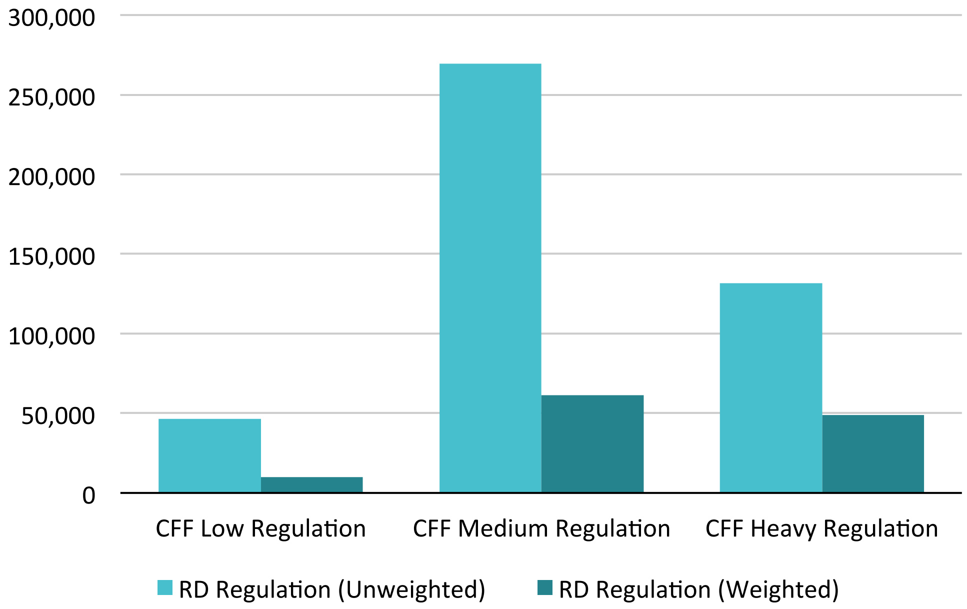
Of the 48 industries, only 43 have reasonable NAICS matches. Some of the matches involve combinations of NAICS industries; for example we match Fama-French industry 25 (Shipbuilding, Railroad Equipment) with NAICS 3365 (Railroad Rolling Stock Manufacturing) and 3366 (Ship and Boat Building). In such cases, case we simply sum the relevant regulation metrics from the matching NAICS industries.

As mentioned above, the matched NAICS industries are drawn from all three levels. In RegData, industries at the two-digit level tend to be more regulated than those at finer levels for linguistic reasons (coarser industries tend to have simpler names that use more common words). We therefore create a weighted version of RegData, where each industry's regulation level is multiplied by a factor that renders the mean regulation level for two-digit industries equal to that for three- and four-digit industries. The last two columns of table 15 contain the gross regulation and the weighted regulation, respectively, of the Fama-French industry according to the RegData-NAICS match.

3.2.2. RegData vs. Coates-Fama-French regulation. RegData regulation (weighted and unweighted) is virtually continuous and numerical, whereas Coates-Fama-French regulation takes three ordered, categorical values (unregulated, medium regulated, and heavily regulated), so there are a variety of ways to explore the relationship. Figure 16 is a graphical exposition with three notable features. First, the 31 unregulated Fama-French industries exhibit lower amounts of RegData regulation than the three medium regulated and eight heavily regulated Coates-Fama-French industries. Second, heavily regulated Coates-Fama-French industries exhibit lower levels of RegData regulation than the medium regulated Coates-Fama-French industries; that is, the relationship between Coates-Fama-French regulation and RegData regulation is nonmonotonic.

Third, the relationship is much closer to monotonicity when using weighted rather than unweighted RegData regulation.

Figure 16. RegData (RD) Regulation vs. Coates-Fama-French (CFF) Regulation



Note: Weighted regulation adjusts by a factor so that two-digit, three-digit, and four-digit industries all share the same mean level of regulation.

Table 16 contains a formal statistical analysis of the relationship. Model 1 is a regression where the dependent variable is RegData (unweighted) regulation and where there are two explanatory variables: a dummy for Coates-Fama-French medium regulation and a dummy for Coates-Fama-French heavy regulation. The results confirm the visual evidence in figure 16; moreover, the estimated coefficient on medium regulation is statistically significant ($p < 1\%$), and that on heavy regulation is marginally significant ($p < 8\%$). Model 2 repeats the exercise with RegData weighted regulation as the dependent variable. Now both medium (p < 5%) and heavy (p < 5%) regulation are significant at conventional levels.[11]

Table 16. OLS Regression Results for RegData (RD) Regulation vs. Coates-Fama-French (CFF) Regulation

	Model 1	Model 2
Dependent variable	RD regulation (unweighted)	RD regulation (weighted)
CFF medium regulated (dummy)	223,000*** (71,800)	51,300** (23,800)
CFF heavily regulated (dummy)	85,200* (47,000)	40,000** (15,600)
Constant	46,400** (21,000)	9,970 (6,970)
Observations	43	43
R^2	0.23	0.19

* denotes statistical significance at the 10% level, ** at the 5% level, *** at the 1% level.
Note: Weighted regulation adjusts by a factor so that two-digit, three-digit, and four-digit industries all share the same mean level of regulation. Figures in parentheses denote standard errors. All numbers are rounded to three significant figures.

With a paltry 43 observations, the results will almost inevitably be sensitive to outliers. A simple way to examine robustness is to repeat the regressions 43 times, dropping each observation in turn (we omit these results for parsimony). It turns out that the results in table 16 are somewhat sensitive to the omission of the Coates-Fama-French industries “Petroleum and Natural Gas” (medium regulated), “Insurance” (heavily regulated), and “Trading” (medium regulated), where the latter refers to securities, investments, funds, trusts, and so on. In all three

[11] Similar results, omitted for parsimony, emerge with an ordered logit using Coates-Fama-French regulation as the dependent variable.

cases, the estimated coefficient on either medium regulation or heavy regulation is no longer statistically significant when the dependent variable is RegData unweighted regulation. However, glancing through the list of industries in table 15, few would argue with the classification of either petroleum and natural gas or trading as being medium regulated (if anything, one might have expected them to be classified as heavily regulated), or with insurance as being heavily regulated, and so we are unconcerned by these results.

In summary, a comparison of RegData to Coates-Fama-French data does deliver some limited ex-post validation. However, these are not results that we wish to emphasize for three reasons: first, the sample size is small; second, the industry-matching procedure requires considerable subjectivity; and third, we have virtually no information on the origins of Coates-Fama-French regulation, meaning that we have very little basis for treating it as a sound benchmark.

4. Applying RegData to Variables of Interest

RegData is most useful when combined with other databases to investigate economic questions that relate to regulation—specifically, the causes and consequences of regulation. The most convenient databases for this purpose are those that include industry-level data where industries are classified according to NAICS. Two US agencies whose databases include such data are the Bureau of Labor Statistics and the Bureau of Economic Analysis.

In this section, we briefly and superficially investigate the econometric relationship between regulation (taken from RegData) and a handful of important economic variables taken from complementary databases. The goal is *not* to present a thorough analysis of the posited relationships since that would require substantial specialized knowledge and would extend the paper's scope well beyond its chief intended purpose, which is to introduce the new data. Rather, the intention is to demonstrate RegData's usefulness in posing and answering important economic questions. We hope that the exploratory regressions we report will motivate specialists to investigate the relationships more thoroughly.

4.1. ***Basic Method***

Let z_{iy} be a panel dependent variable that is posited to be causally related to regulation, r_{iy} (which we measure using RegData). Variation in both z_{iy} and r_{iy} is naturally occurring rather than experimentally randomized. Thus, an econometrician investigating the relationship between the two will look to control for the most obvious sources of unobserved variation in an attempt to limit endogeneity bias. Moreover, since the data are in panel form, an econometrician will also make efforts to correct for nonspherical disturbances to maximize the accuracy of the proceeding inference.

The choice of controls usually requires specialized knowledge about the specific relationship in question, and this is particularly true when using industry- or economy-level data. However, estimating the relationship without controls typically constitutes an instructive first step in an investigation. With this in mind, for each pair of variables (z_{iy}, r_{iy}), we estimate the following model:

$$z_{iy} = \alpha_i + \mu_y + \beta r_{iy} + \varepsilon_{iy},$$

where α_i is an industry fixed effect, μ_y is a year effect, and ε_{iy} is a disturbance that has a covariance matrix that allows for intraindustry correlation.

There can be a significant lag between the time a regulation is declared (either in intent by an act of Congress, in a proposal in the *Federal Register*, or even in discussion between regulators and stakeholders) and the time its final version appears in the CFR. Furthermore, there can be a significant lag between the time a regulation appears in the CFR and the time it is actually enforced (and generally perceived to be enforced). Thus, the true relationship between z_{iy} and regulation may be best captured by either backward- or forward-lagged regulation. We therefore also estimate the following six models:

$$z_{iy} = \alpha_i + \mu_y + \beta r_{i,y+k} + \varepsilon_{iy}, k \in \{-3, -2, -1, +1, +2, +3\}.$$

All dependent variables and regulation are logged, meaning that the coefficient on regulation can be interpreted as an elasticity. Finally, our regressions pool three- and four-digit industries.

4.2. *Data*

The US Bureau of Labor Statistics provides data on a wide variety of variables related to employment and labor productivity at the three- and four-digit NAICS industry levels. We were particularly interested in the following labor-related series:

- labor productivity (indexes of output per worker and output per hour)
- employment (number of workers and number of hours)
- labor compensation (gross payroll plus supplemental benefits)

Details about the data collection methods can be found at the BLS website (http://www.bls.gov /data/).

Compustat provides US stock market data for thousands of firms, also classified according to the NAICS system. We use the series on market value to calculate the gross nominal market capitalization of each industry at the three- and four-digit levels, and we then calculate real market capitalization using CPI data.

Beyond the convenience factor of NAICS, we chose these data because they are all variables that are likely to causally respond to regulation. In fact, the debate on the employment and productivity effects of regulations is particularly intense, and there have been many studies of the effect of a narrow range of regulations (see, e.g., Greenstone [2002]).

4.3. *Exploratory Results*

Table 17 contains the regression results where regulation is neither forward nor backward lagged (those detailed results are available upon request). We term the following results as "exploratory" because they feature no controls (beyond time and industry fixed effects); we make no claims about their robustness to what a specialist might consider to be a "standard" set of controls.

Table 17. OLS Regression Results for RegData Regulation vs. Key Economic Variables

Dependent variable	Model	Coefficient (SE) on regulation		No. of obs.	No. of industries	R^2
Output/worker (index)	1	0.00755	(0.0292)	2,178	146	0.277
Output/hour (index)	2	0.0108	(0.0285)	2,178	146	0.250
Total employment (persons)	3	−0.0871***	(0.0289)	4,500	286	0.0101
Total employment (hours)	4	−0.0888***	(0.0296)	4,500	286	0.0115
Total labor compensation ($)	5	−0.146***	(0.0458)	2,178	146	0.107
Stock market value ($)	6	−0.0709	(0.153)	3,175	260	0.0102

* denotes statistical significance at the 10% level, ** at the 5% level, *** at the 1% level.
Note: All regressions include industry fixed effects and year dummies (panel is 2007–2011), and standard errors are robust. All numbers are rounded to three significant figures.

Exploratory Result 1: According to rudimentary regressions, the causal effect of regulation on labor productivity is insignificantly different from zero.[12]

The coefficient on regulation in the labor productivity models (1 and 2) is miniscule (an elasticity of approximately 1 percent) and insignificantly different from zero. Moreover, the estimated sign is not robust when looking at the forward- and backward-lagged data (results omitted).

Exploratory Result 2: According to rudimentary regressions, there is suggestive evidence of a negative effect of regulation on employment, whether employment is measured in number of workers or total hours worked.

The coefficient on regulation in the employment models (3 and 4) is −9 percent in both employment measures (including all lags; results omitted) and is statistically significant ($p < 1\%$).

Exploratory Result 3: According to rudimentary regressions, there is suggestive evidence of a negative effect of regulation on total labor compensation.

The coefficient on regulation in the compensation model (5) is −15 percent (also in all lags; results omitted) and is statistically significant ($p < 1\%$).

Loosely speaking, decreasing total labor compensation is either the result of decreased wages/salaries, decreased employment, or a combination of the two. In light of the preceding results, Exploratory Result 3 seems to be more the result of decreased employment than decreased wages, since productivity (and hence wages/salaries implicitly) is apparently not adversely affected by regulation.

[12] However, Davies (2014) used the first version of RegData to more carefully explore the relationship between labor productivity and regulation, and finds that higher levels of regulation tended to cause slower labor productivity growth.

Exploratory Result 4: According to rudimentary regressions, the causal effect of regulation on the stock market value of firms is insignificantly different from zero.

The coefficient on regulation in the market capitalization model (6) is small (approximately −8 percent) and negative in all specifications (including all lags; results omitted) and is statistically insignificant. For researchers planning a more rigorous study, it is worth noting that the coefficient increases in magnitude (up to −24 percent) and attains conventional significance as one shifts from contemporaneous regulation to three-year, forward-lagged regulation as the explanatory variable. In principle, this result is consistent with stock prices shifting immediately in response to regulation announcements, which themselves take a few years to enter the CFR. However, a thorough investigation of this claim is beyond the scope of this paper's exploratory analysis.

5. Closing Remarks

This paper introduces RegData. The version of RegData described here is the second 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 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 to 2012. The first database contains three metrics of regulation quantity: CFR page-count data, digitized CFR file-size data, and a novel measure called "restrictions" that counts the number of legally binding words (e.g., "shall" or "must") contained in regulatory text.

In the second dataset, we offer the first measure of the relevance of units of the CFR (such as titles, chapters, and parts) to industries in the United States. This measure was created by searching each unit of the CFR for text strings that describe each industry in the United States, 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 (see the Bureau of Economic Analysis website for examples of such data).

RegData offers users numerous choices that we hope will permit maximum experimentation and minimize any subjectivity or idiosyncrasy inherent in the database's creation. Users can decide how to combine the databases (e.g., whether and how to weight restrictions 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.

We have employed two strategies to validate RegData: 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. On the flip side, its novelty increases its potential value, since its novelty means that it holds much promise in helping elucidate the causes and consequences of regulation. 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. There is suggestive evidence, for example, that regulation negatively affects employment and total labor compensation.

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—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.

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.

References

Aghion, P., Algan, Y., Cahuc, P., & Shleifer, A. (2010). Regulation and Distrust. Quarterly Journal of Economics, 125, 1015–49.

Baker, S., Blook, N., & Davis, S. (2013). Measuring Economic Policy Uncertainty. Working Paper, Stanford University.

Botero, J., Djankov, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2004). The Regulation of Labor. Quarterly Journal of Economics, 119, 1339–82.

Calabria, M. (2009). Did Deregulation Cause the Financial Crisis? Cato Policy Report, 31(4).

Coates, J. (2012). Corporate Politics, Governance, and Value before and after Citizens United. Journal of Empirical Legal Studies, 9, 657–96.

Coffey, B., McLaughlin, P. A., & Tollison, R. D. (2012). Regulators and Redskins. Public Choice, 153, 191–204.

Coglianese, C. (2002). Empirical Analysis and Administrative Law. University of Illinois Law Review, 4, 1111–38.

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

Davies, A. (2014). Regulation and Productivity. Mercatus Research. Mercatus Center at George Mason University. Available online: http://mercatus.org/publication/regulation-and-productivity

Dawson, J., & Seater, J. (2008). Federal Regulation and Aggregate Economic Growth. Working Paper.

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

Fama, E., & French, K. (1997). Industry Costs of Equity. Journal of Financial Economics, 43, 153–93.

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, 401–25.

Greenstone, M. (2002). The Impacts of Environmental Regulations on Industrial Activity. Journal of Political Economy, 110, 1175–217.

McChesney, F. S. (1987). Rent Extraction and Rent Creation in the Economic Theory of Regulation. Journal of Legal Studies, 16, 101–18.

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

McLaughlin, P., & Greene, R. (2012). Measuring the Regulatory Impact of the Dodd-Frank Act. Mercatus Center Working Paper.

Mulligan, C., & Shleifer, A. (2005). The Extent of the Market and the Supply of Regulation. Quarterly Journal of Economics, 120, 1445–73.

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

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

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

Winston, C. (1998). US Industry Adjustment to Economic Deregulation. Journal of Economic Perspectives, 12, 89–110.

Appendix A: Construction of Industry Relevance Metric

1. Industry Name Structure

The NAICS industry description is a collection of words or phrases linked by conjunctions or commas, for example, "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 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 be absent [*Construction*]. We ignore prepositions.

2. Rules for Strings

Each of the following 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 from 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 to group 1 and *gas extraction* is assigned to group 2.

3. 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, for example, "mining (except oil and gas)." Our solution is to simply ignore the parenthetical comment.

The second is the case of "other," "support," or "related" activities, for example, "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.

Finally, reasonable people can agree to differ about whether a certain string is relevant to a certain industry. The database is constructed according to our judgment; our judgment should be taken as neither definitive nor binding.

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

4. Calculating Industry Relevance

Each industry description is associated with a collection of strings. The strings are classified according to group and rule. For each group in each industry, each rule in the range 1 to 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

In the above, 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 or excluding, though we still measure the occurrence of every string to allow readers to judge for themselves.

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 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. We also include code that avoids double counting the rule 9 strings, given that they are necessarily included in rule 7.

In some cases, a string is shared by multiple groups in the same industry, for example, *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 we sum the total occurrences of the included strings in the unit of aggregation (part/chapter/title). We then divide by the number of words in the unit to obtain a measure of **industry relevance per word**.

What we have described above is the **standard/direct** approach. To address the shortcomings described in section 3 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 sub-industries. Future iterations of this database will include industries at the five-digit and six-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 50 CSV (comma separated values) files. There are also three 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_restrictions.csv: This file contains the frequency of each command string by year/agency.

- *year*: year from {1997, 1998, . . . , 2012}
- *agency*: regulatory agency from {1, 2, . . . , 280}
- *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; for example, the word "muster" will not result in a hit for the string "must." Restriction counts are applied to agencies according to their location in the CFR. Each agency publishes regulatory text in specified CFR parts. The agency-level datasets available in RegData 2.0 are the summations of the part-level data for each agency. Some text is published in appendixes, supplements, and other portions of the CFR that are not specifically assigned to agencies. This unassigned text is grouped together into a single "agency" called "all other agencies."

data_word_count.csv: This file contains the number of words by year/agency.

- *year*: year from {1997, 1998, . . . , 2012}
- *agency*: regulatory agency from {1, 2, . . . , 280}
- *word_count*: word count for year/agency (positive integer)

Word counts are applied to agencies according to their location in the CFR. Each agency publishes regulatory text in specified CFR parts. Some text is published in appendixes, supplements, and other portions of the CFR that are not specifically assigned to agencies. This unassigned text is grouped together and called "all other agencies."

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/agency.

- *year*: year from {1997, 1998, . . . , 2012}
- *agency*: regulatory agency from {1, 2, . . . , 280}
- *string*: string derived from NAICS industry description according to the rules specified above
- *count*: the number of times the string appears in the year/title/volume (positive integer)
- *code*: two-digit industry code according to the NAICS
- *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*: a 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*: a 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*: a 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 must be matched; for example, 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/agency. Everything else is identical to **naics2_X.csv**.

naics4_X.csv (where X is an element of {1, 2, . . . , 30}): This file contains the frequency of each four-digit industry-relevance string by year/agency. 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 industry code according to NAICS
- *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_naics4.csv: This file contains the full descriptions of the four-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 agencies' parts and presents two-digit industry data at the year/agency level.

- *year*: year from {1997, 1998, . . . , 2012}
- *title*: CFR title from {1, 2, . . . , 50}
- *title_name*: the CFR title name
- *agency*: regulatory agency from {1, 2, . . . , 280}
- *code_2*: two-digit industry code according to the NAICS
- *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}
- *word_count*: year/agency word count (positive integer)

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 agencies' parts and presents three-digit industry data at the year/agency 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 4).

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

---

cleaning_naics4.do: This Stata.do file cleans and combines the above data files. It aggregates over agecnies' parts and presents four-digit industry data at the year/agency level. All variables are identical or analogous to **cleaning_naics2.do**, except that we include two-digit and three-digit industry codes and names in case the user wants to use a bottom-up approach (see Appendix A, section 4).

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

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 Yazigi, 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.