The Regulatory Review *In Depth*

Volume 3, Number 3 December 2014

WHY WE NEED TO MEASURE REGULATION

Omar Al-Ubaydli[†] & Patrick McLaughlin[‡]

How much should governments regulate? Economists attach huge significance to this question because the difference between rich and poor countries can be affected by variation in government quality, especially the extent to which the legal system protects property and encourages innovation and investment. But answering this question has been difficult due to limitations both in economic theory and available data. We have created a new dataset that can help overcome the limits in available data.

Admittedly, modern econometric methods have in recent decades allowed economists to advance significantly their ability to answer questions about governmental regulation. Researchers can propose sophisticated responses to questions about the effects of specific laws—e.g., what were the employment consequences of Germany's Renewable Energy Act?—and thereby help provide well-reasoned bases for future policy decisions.

Yet data limitations make it difficult if not impossible to answer more general questions about regulation. The key ingredient for the empirical analysis of regulation is *numerical* data, such as profits and costs. But regulation is typically measured in non-numerical, *categorical* terms.

Researchers use non-numerical data when categories cannot be ranked; for example, location or sector are typical non-ordinal, categorical variables. However, using categorical data forces researchers to narrow the scope of their analysis because these data constrain the researcher when the data exhibit gaps, which they always do. In contrast, numerical variables, such as unemployment or prices, allow researchers to cover gaps by interpolating or extrapolating.

To be more concrete, consider a researcher estimating the effect of mortgage rates on house prices—both numerical variables. According to the data on the Federal Reserve website, mortgage rates have varied substantially since 1971, but the data have gaps: mortgage rates have, for example, sometimes been exactly 3.60% and sometimes exactly 3.68%, but they have never been anything in between. Nor have they ever fallen below 3.35% or

[†] Director of International and Geo-Political Studies Program at the Bahrain Center for Strategic, International, and Energy Studies, and an Affiliated Associate Professor of Economics and an Affiliated Senior Scholar at the Mercatus Center at George Mason University.

[‡] Senior Research Fellow at the Mercatus Center at George Mason University.

exceeded 18.45%. By interpolating and extrapolating, however, we can still make educated guesses about what happens to housing prices when mortgage rates fall within these gaps. For example, a reasonable guess about what happens when rates are 3.64% is that they are halfway between what happens at 3.60% and 3.68%. In this sense, the data's numerical nature permits the researcher to expand the scope of her analysis beyond the range of the original data.

On the other hand, if the researcher wants to look at the effect of a state jurisdiction—a categorical variable—on something like housing prices, and if she is missing data on 20 states, then she has no reasonable basis for drawing conclusions about those missing states. She could "geographically" interpolate or extrapolate by, for example, assuming that house prices in South Dakota are halfway in between those in North Dakota and Nebraska (its northern/southern neighbors), but that requires a much larger leap of faith. That is why categorical data are less helpful than numerical data.

Government regulations have, at least until recently, been treated as non-ordinal, categorical variables. As a result, it has made little sense to say something like "mining is three times more regulated than farming." Even if indirect cardinal measures might be available, such as compliance expenditures, they can be misleading proxies for the degree of regulation because they measure more than one thing—both the extent of regulation as well as the costs to comply with or enforce that regulation. For example, in your office, your employer might impose the same constraints on the content of your speech and your email. Yet it is cheap to enforce the email constraints by using software, while enforcing the same constraints on oral speech may require spending huge amounts on proving someone said something inappropriate. If we used enforcement expenditure as a gauge of the degree of regulation, we may incorrectly infer a higher degree of "regulation" on speech than on email, even though the same standard applies to both.

Consequently, existing regulation research—while being valuable—usually has answered only narrow questions, such as, "What was the effect of the Obamacare on unemployment?" or "How do state variations in political affiliation affect the incidence of concealed carry laws?" In both of these examples, more ambitious studies would seek to examine questions that call for a cardinal representation of regulation in general: "What is the effect of regulation, writ large, on unemployment?" or "How does political affiliation in general affect the level of regulation?"

As a complement to empirical work, economists often call for the use of "theory." This means creating a simplified model of the primary actors, their choices, and the associated incentives, sometimes with the aim of informing the econometric model to be used. Economists often rely on theory because data limitations restrict the ability of purely empirical approaches to deliver definitive conclusions.

In the context of regulation, theory sets the stage for much of the controversy over government intervention. According to the British economist Arthur Pigou, markets can malfunction due to a variety of market failures, such as monopoly power or externalities. In these cases, a sufficiently informed

and benevolent policymaker can deploy regulation to enhance societal welfare, such as by adopting anti-trust or environmental laws.

Of course, economist Ronald Coase showed that Pigovian-motivated regulation might be rendered redundant by the organic desire of the affected actors to resolve market failure themselves, through a decentralized, multilateral bargain, aided by courts in the case of disputes. For example, residents concerned about local crime may choose to form a neighborhood watch rather than rely on government intervention. While the affected parties often have the strongest incentive to do something about a market failure, the large numbers may make it impractical: it would be difficult, for instance, to get all the residents of a suburb to bargain and reach a consensus on the optimal level of noise pollution. Hence, economic theory suggests that when "transactions costs" are prohibitive, Pigovian style government regulation may be an efficient alternative.

Economist George Stigler took a more skeptical view of government motives. He argued that the Pigovian model—even one that accepts Coasian insights—assumes benevolent policymakers, when in practice regulations may in fact reflect policymakers' personal agendas. In the case of "regulatory capture," leading figures in the regulated industry co-opt the regulator, resulting in regulations that serve the interests of industry leaders at the expense of the industry's smaller players, potential entrants, or other industries. For example, U.S. car manufacturers might convince the government to impose a tariff on car imports to their benefit and at the expense of U.S. consumers. Historically, there are examples of policymakers taking things a step further and regulating for their own direct benefit, such as when a medieval baron erects a barrier across a river and charges travelers a toll. A hypothetical, modern incarnation would be a government passing stringent safety standards but allowing private actors to obtain discretionary exemptions in exchange for favors, such as financial kickbacks.

Good intentions are one thing, but according to economist Sam Peltzman, good information is another. Whether benevolently conceived or otherwise, regulations can backfire because regulators lack full information about the future consequences of the rules they adopt. For example, poor design and foresight meant that the Endangered Species Act motivated landowners to "shoot, shovel, and shut-up"—discreetly killing endangered species, rather than protecting them.

Who is right: Pigou or Stigler and Peltzman? Theory only goes so far in answering that question. Because there are sound reasons to anticipate both good and bad consequences from regulations, the burden shifts to empirical research. To refine our knowledge of the causes and consequences of government regulations, we must therefore develop better strategies for empirical inquiry.

As we noted at the outset, the inability to quantify regulation in numerical terms has limited researchers' ability to generalize about regulation. This is where a new dataset we developed, RegData, can help.[1] RegData is an attempt to loosen the ties that bind regulatory scholars. It is the first database that

[1] REGDATA, http://www.regdata.org.

provides users with an industry-level panel of U.S. federal regulation, turning regulation from a non-ordinal, categorical variable into a numerical one. RegData, which currently covers the period 1997-2012, is produced using custom-made text analysis software to measure, in numerical terms, how restrictive federal regulations are and which industries regulations are most likely to affect. Through RegData analyses, for example, researchers will now be able to compare the level of regulation in U.S. fishing in 2001 to the level of regulation in U.S. fishing in 2009.

Suffice it to say, with the kind of numerical data about regulation that RegData provides, researchers will be able to study a wider array of questions about regulations' causes and consequences—and perhaps finally begin to narrow the field's theoretical divide. Economic theory alone is incapable of resolving many of the controversies over optimal government regulation; therefore, sound empirical projects must play critical role in helping societies respond effectively to challenges like financial sector reform and climate change.

To advance the empirical study of industry regulation, we have developed the first meaningful numerical dataset of U.S. federal regulation, covering a period from 1997 to 2012. U.S. federal regulation is published in the *Code of Federal Regulations* (CFR).[2] This publication has led regulatory scholars over the years to use page-counts and word-counts to measure total economy-level regulation, resulting in valuable research at the level of the entire economy. Our new dataset, RegData, evolves this technique by providing more granular data on regulation at the industry level.[3]

How exactly does RegData work? In principle, it simply extends the method used for economy-level regulations: e.g., for each industry, it counts the number of words in the CFR that apply to the industry.

The easy part is getting a list of industries. The North American Industry Classification System (NAICS) provides an exhaustive list of industries.[4] In one version of NAICS, the U.S. economy is divided into approximately 20 industries, whereas in a finer-grained version of NAICS (a six-digit version), the economy is subdivided into over 1,000 industries.

The difficult part is determining whether any particular section of CFR text applies to a given industry. The CFR is not organized in a way that logically maps to a mutually exclusive list of industries. For example, one particularly large component of the CFR is called "Title 40: Protection of Environment," and the regulations in that title affect a wide variety of industries, from petrochemical manufacturing (NAICS code 32511) to shellfish fishing (NAICS code 114112).

To address the issue of what parts of the CFR apply to specific industries, we developed an algorithm (described more fully in our recent working paper) that assesses the extent to which a paragraph of CFR text applies to a given

---

[2] *Code of Federal Regulations*, https://www.ecfr.gov/.

[3] REGDATA, *supra* note 1.

[4] *North American Industry Classification System*, U.S. CENSUS BUREAU, https://www.census.gov/naics/.

industry.[5] It works by first creating a list of key words that are associated with each industry, and then seeing how often those words appear in each paragraph of rule text.

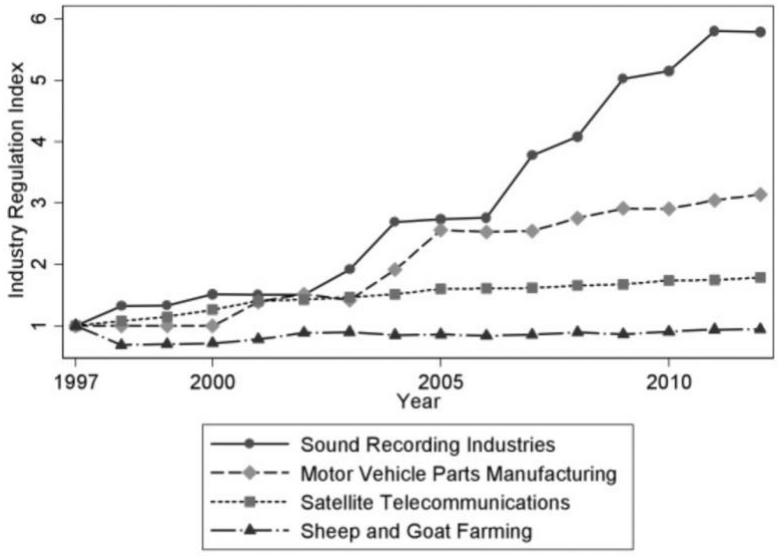
In the case of petrochemical manufacturing, unsurprisingly, one of the key words that the algorithm associates with the industry is "petrochemical." The more times a paragraph has the word "petrochemical" appear in it, the more applicable that paragraph is to petrochemical manufacturing, according to RegData.

Once the algorithm has computed an applicability rating for each industry and each paragraph (within the millions of paragraphs in the CFR), it multiplies the industry's applicability rating by the number of words in the paragraph. It then totals that figure up across all the paragraphs. That generates a final index of the level of regulation in that industry.

As a further refinement, the algorithm also takes into account how many words in the paragraph are associated with regulations in general, such as "must," "shall," and "required." Put simply, when words like "petrochemical" appear close to words such as "must," then the section of text is likely to be regulating petrochemicals.

Figure 1 below demonstrates the algorithm's results for four merely illustrative industries. One can see that with the advent of digital rights management, sound recording industries faced dramatic increases in the level of regulation through the turn of the new millennium, especially compared to the stability of regulation imposed on the more mundane sheep and goat farming industry.

Figure 1: Industry Regulation Index for
a Selection of NAICS Four-Digit Industries



[5] *See* Omar Al-Ubaydli & Patrick A. McLaughlin, *RegData: A Numerical Database on Industry-Specific Regulations for All U.S. Industries and Federal Regulations, 1997-2012* (Geo. Mason U. Mercatus Center, Working Paper No. 12-20, 2014).

RegData allows the researcher to probe deeper on the source of regulation. The text in the CFR is written by the dozens of departments and agencies that comprise the federal government, and, for the most part, the CFR indicates the department or agency that wrote each passage of text. The RegData algorithm can therefore be modified to determine the extent to which a given department or agency regulates a given industry. For example, to what extent does the Department of Labor regulate petrochemical manufacturing?

RegData also offers potentially fresh insight on "hot" debates, such as financial sector reform. In the wake of the Great Recession, two camps emerged. The first camp blamed the financial crisis on unregulated, free markets overcome by greed and shortsightedness, and that camp called for tighter regulation as a solution. The second camp considered excessive government involvement, be it in the form of support for housing loans or post-collapse bailouts, as the source of economic woes, and it saw true liberalization as the solution.

RegData provides researchers with a rich set of tools for assessing the extent and consequences of financial sector regulation. Researchers might gain insights, for example, by looking back across time to see if various indicators of financial market performance – from bankruptcies to new start-ups – correlate with changing levels of regulation.

To be sure, RegData is not a perfect measure of regulation: if you want the "true" measure, you will have to read the entire CFR! Anytime you summarize, you lose important details. But the RegData algorithm is publicly available and adaptable. Similar to the open source community, to maximize the rate at which RegData improves, we subscribe to principles of openness and customizability.

In the often emotional debates over the nature of optimal government intervention, RegData's imperfections do mean that it can be exploited to support certain agendas. That is another important reason for our openness and customizability policy. We want to avoid the proverbial baby being thrown out with the bathwater by skeptical users. Thus, if some business lobby should try to use RegData to milk favors from the government and opponents grow suspicious of RegData's algorithm, we would encourage the skeptics to dig deeper, ask questions, and experiment for the benefit of all.

RegData looks to emulate other data collection projects that have facilitated the advancement of economic knowledge. For example, once upon a time, economists could not conduct cross-country comparisons of economic welfare because there were no international GDP datasets. Now, thanks to projects such as the Penn World Table, economists can study the reasons behind international differences in economic development with unprecedented levels of econometric sophistication. Similarly, thanks to the Barro and Lee database, for 30 years now scholars have been able to study international variation in educational attainment.[6]

---

[6] *Barro-Lee Educational Attainment Dataset*, BARRO AND LEE DATASET, http://www.barrolee.com/.

We aspire to have RegData have much the same impact on the study of industry-level regulation. Rather than being an improvement on previous measurement efforts, RegData is the *first* true measurement effort beyond mere page counts, meaning that it has the potential to open many new doors.

We hope that many scholars and professionals will find the data useful and will contribute to the further transparent development of new versions of RegData.