Man vs. Machine

A Novel Evaluation of Data Analytics Using Occupational Licensing as a Case Study

Conor Norris and Edward Timmons

MERCATUS WORKING PAPER

MERCATUS CENTER
George Mason University
3434 Washington Blvd., 4th Floor, Arlington, Virginia 22201
www.mercatus.org

Conor Norris and Edward Timmons. "Man vs. Machine: A Novel Evaluation of Data Analytics Using Occupational Licensing as a Case Study." Mercatus Working Paper, Mercatus Center at George Mason University, Arlington, VA, March 2021.

Abstract

For researchers of state regulatory policy, the difficulty of gathering data has long presented an obstacle. This study compares two new databases for state-level occupational licensing laws. The Knee Center for the Study of Occupational Regulation (CSOR) database uses traditional manual reading to gather data, while RegData uses a machine learning algorithm. We describe both data-gathering processes, weigh their costs and benefits, and compare their outputs. The CSOR database allows researchers to find specific licensing requirements typically used in the occupational licensing literature, but the traditional methodology is time and labor intensive. RegData provides researchers with a better overall measure of stringency and complexity in regulation that allows for comparisons across states. However, RegData cannot reach the level of detail in the CSOR database. The variables gathered by CSOR and RegData are useful for researchers and policymakers and can be used as a model to build databases for other state-level regulations.

JEL codes: C81, C63, J44

Keywords: complexity of law, data collection, occupational licensing, RegData

Author Affiliation and Contact Information

Conor Norris
Saint Francis University
cnorris@francis.edu

Edward Timmons
Saint Francis University
Mercatus Center at George Mason University
etimmons@francis.edu

This paper can be accessed at https://www.mercatus.org/publications/regulation/man-vs-machine-novel-evaluation-data-analytics-using-occupational-licensing

Man vs. Machine: A Novel Evaluation of Data Analytics Using Occupational Licensing as a Case Study

Conor Norris and Edward Timmons

Thanks to significant advances in the availability of data on federal regulations, we now know much more about their cumulative effects.[1] In the case of state-level regulation, a number of questions remain unanswered, primarily because of existing data limitations. With more comprehensive data in hand, researchers can help measure the effects and outcomes of state policy, a critical element in advising policymakers and informing the public.

One example of state-level policy with a considerable impact on both labor markets and consumers is occupational licensing. Research on the effects of occupational licensing has examined a range of outcomes. There is research that focuses on wage premiums for licensed professionals,[2] how licensing impacts recidivism,[3] racial disparities,[4] deterrence of crime,[5] and the quality of services provided.[6] However, researchers face difficulties gathering information about licensing requirements in all 50 states. Because licensing is passed at the state level, researchers must access each state's administrative code. Without a centralized database for

[1] O. Al-Ubaydli and P. A. Mclaughlin, "RegData: A Numerical Database on Industry-Specific Regulations for All United States Industries and Federal Regulations," *Regulation and Governance* 11, no. 1 (2015): 109–23.
[2] Morris M. Kleiner and Alan B. Krueger, "The Prevalence and Effects of Occupational Licensing," *British Journal of Industrial Relations* 48, no. 4 (2010): 676–87; Maury Gittleman and Morris M. Kleiner, "Wage Effects of Unionization and Occupational Licensing Coverage in the United States," *ILR Review* 69, no. 1 (2016): 142–72; Samuel J. Ingram, "Occupational Licensing and the Earnings Premium in the United States: Updated Evidence from the Current Population Survey," *British Journal of Industrial Relations* 57, no. 4 (2019): 732–63.
[3] Stephen Slivinski, "Turning Shackles into Bootstraps: Why Occupational Licensing Reform Is the Missing Piece of Criminal Justice Reform" (Policy Report, Center for the Study of Economic Liberty at Arizona State University, 2016).
[4] Peter Q. Blair and Bobby W. Chung, "Job Market Signaling through Occupational Licensing" (NBER Working Paper No. 24791, National Bureau of Economic Research, 2018).
[5] D. Deyo, B. Hoarty, C. Norris, and E. Timmons, "Licensing Massage Therapists in the Name of Crime: The Case of Harper v Lindsay," *Journal of Entrepreneurship and Public Policy* 10, no. 1 (2021): 1–14.
[6] Bradley Larsen, "Occupational Licensing and Quality: Distributional and Heterogeneous Effects in the Teaching Profession" (SSRN Paper 2387096, January 31, 2013).

licensing statutes and requirements, the process of gathering data from each state can make licensing research laborious and time-consuming.

Historically, teams of researchers or research assistants have had to gather the data by hand from each state. Gathering data manually allows for more detailed information to be collected. With recent improvements in machine learning technology, researchers are able to use algorithms to read administrative codes and pull out the requirements, reducing the time and effort it takes to gather data. However, reliance on machine learning comes with concerns about accuracy at this stage of development; sophisticated programming capabilities and human assistance are also needed to refine the algorithm.

In this paper, we compare a database that uses traditional methods, the Knee Center for the Study of Occupational Regulation (CSOR) database, and another that uses machine learning, the Occupational Licensing RegData database.[7] We begin with a discussion of the process behind building both databases, comparing both methods and finding the strengths and shortcomings of each in practice. Then we compare the outputs from both methods, providing detailed information from both databases for four states for the dental hygienist profession. We then compare rankings of each state's level of stringency and discuss the results. We find that both datasets have potential in helping to answer different empirical questions.

CSOR Database

In 2016, the CSOR created a nationwide database of licensing requirements at the state level. Before then, researchers studying the effect of occupational licensing had to read through state statutes for each occupation being studied, because no comprehensive database existed. State

[7] Additional datasets are scheduled to be forthcoming in 2021. See Morris Kleiner and Edward Timmons, "Occupational Licensing: Improving Access to Regulatory Information," *Journal of Labor Research* 41, no. 4 (2020): 333–37.

governments generally do not compile licensing standards in a place that allows for easy comparison. Additionally, gathering specific requirements for occupations meant undertaking the labor-intensive manual reading of 50 separate state codes.

The CSOR database allows researchers and policymakers to compare the stringency of occupational licensing between states. These standards tend to vary state by state, sometimes substantially. The variables in the database are objective and typically quantifiable, including the amount of fees, days of training, and years of education required to obtain a license. Additionally, because the CSOR database contains over 330 separate professions, it allows for comparisons of licensing standards between professions.

The CSOR database is compiled by a team of undergraduate student fellows who must read and collect data from a large number of sources. The current team is composed of 15 students who are active fellows. Each fellow is assigned a set of occupations and given a deadline to complete tasks. For each occupation, team members are assigned all 50 states. This is done to prevent inconsistencies when collecting data from different states in the same profession and to limit the effect of human subjectivity, thereby improving the precision of the data within occupations.

The student fellows are on a team monitored by a staff member with two graduate degrees and three graduate assistants. They ensure that student fellows meet deadlines and the variables in the database are accurate through extensive checking. Each student fellow is given four to eight occupations per year. The data exclusively come from government sources, typically the legislation and state regulatory codes, but sometimes from other state licensing board communications, where the boards are given the authority to set certain requirements.

RegData Database

RegData products come from the QuantGov platform created by the Mercatus Center. RegData was developed in 2012 to provide a measure of federal regulations to be used for empirical analysis, something previously not possible. Researchers had been forced to rely on proxy variables, like the number of pages in the Code of Federal Regulations. Although simple and easy to gather, these variables could be inaccurate when measuring the stringency of regulations. RegData improves on past methods by providing the word count, the number of words that indicate a prohibited or required activity, and the complexity of regulations. This data can be separated into industries using the six-digit North American Industry Classification System (NAICS) code. Recently, the RegData algorithm has been used to measure the state regulatory codes as well, providing the same type of variables to allow comparisons between states.

Researchers relied on proxy variables because of the time and cost of using human effort to read and create data about the stringency of regulatory policy. To overcome this limitation, RegData uses machine learning algorithms to capture the extent of regulations. The text analysis creates a database of variables to compare regulations, including the volume, the restrictiveness, and the complexity of the regulations. RegData counts the number of restrictions, which improves the accuracy of data and avoids counting deregulatory actions as regulations. Restrictions are measured using words that are a requirement to comply, such as “shall,” “must,” “may not,” “required,” and “prohibited.”

The Occupational Licensing RegData (OL RegData) used the same machine learning algorithm as the original RegData. The algorithm is trained by searching for sections that contain language that regulates participation in an occupation. The training includes 1,200 regulations

from the code, 200 of which are related to licensing and 1,000 of which are other regulations. Once the algorithm correctly identifies each licensing regulation, it is then deployed on the entire state code to create an occupational licensing database in two steps. First, the algorithm uses text analysis to identify occupational licensing regulations. Next, text analysis is used to identify the number of restrictions and determine the complexity of the code. OL RegData allows researchers to break down the data by occupation using the Standard Occupational Classification system.

Assessing Strengths and Weaknesses

Both methods of gathering data have relative strengths and weaknesses worth exploring in more detail, as both provide valuable information for researchers and policymakers.

The CSOR method has advantages over OL RegData. CSOR is able to compare specific burdens of occupational licensing with more granular detail. Instead of assessing overall stringency, CSOR database shows the specific licensing requirements that are present. This allows us to compare which restrictions exist and how stringent they are between states. Cross-state comparisons are made easier by the quantified burdens used. Researchers can estimate the effect of specific occupational licensing requirements between states—for instance, estimating the impact of the required years of education across states. The variables are standardized between states, even if different terminology or units are used by different states.

Human data gathering allows the database to provide a greater level of detail regarding specific requirements. The ability to include notes in the dataset can help explain anomalies that may cause researchers to exclude a data point as an outlier or an error. Manual data gathering allows team members to find new variables to include while in the process of gathering data, even if the variable is specific to one or a small number of occupations.

However, this level of detail comes at a cost. Types of variables included in the dataset are restricted by the method of data collection. While easily quantifiable variables can be gathered with little effort, other variables are impractical. For instance, counting the total number of restrictions is not reasonable. Developing an index of restrictiveness relies on an additional researcher creating a methodology to rank and assess the level of stringency.

Legislative codes are large and difficult to read, requiring a large team to compile a database. Team members require periods of training, and there is a learning curve involved with gathering and inputting data. Manual data gathering by humans also requires that team members make judgment calls about what to include or how to interpret complex legal codes. Even though occupations are gathered by one team member to minimize these judgments biasing the data across states, and the process is overseen by a single staff member, the concern remains for comparisons across occupations.

Human data gathering also introduces the possibility of human error. Mistakes entering numbers, misreading requirements, and reading wrong parts of the regulatory code are all possible and present accuracy challenges. Supervision and training can reduce some of these issues, yet they still persist at some level. Furthermore, these solutions themselves are not costless.

The RegData method using machine learning text analysis improves on the weaknesses of gathering data manually. It substantially reduces the number of people necessary to compile a dataset and the amount of time needed. Machine learning removes human error or judgment calls between occupations or states, providing more accurate and precise data.

While the variables in the RegData dataset present challenges for traditional occupational licensing research, the RegData method opens new research possibilities that previously were not

feasible. Machine learning analysis of text provides the ability to easily gather data that measures the overall burden of licensing laws. Stringency measures are uniform across occupations and include measuring the length of the relevant portion of the state code, the number of restrictions, and the difficulty of reading the text. This allows researchers to compare the overall stringency of the regulatory environment between states.

In addition to state-level data, RegData can be broken down by Standard Occupational Classification (SOC), allowing researchers to easily link up with data from the Census Bureau and the Bureau of Labor Statistics that use SOC classifications.

There are some weaknesses with this approach in its early use. Accuracy issues still exist. Writing and training the algorithm is a lengthy process that requires testing to ensure that it is reliable enough to use. Unlike the original RegData, which uses the Code of Federal Regulations, the OL RegData uses 50 different state regulatory codes, which requires ensuring that the results are accurate for each state. The variables in the dataset are less specific than gathering data by hand, so to answer some research questions, researchers still must use traditional data-gathering methods.

A Simple Comparison

To better explain both approaches, we present a simple comparison. In tables 1 and 2 (see the appendix), we compare the outputs for OL RegData and the CSOR database. The occupation selected was dental hygienist, because it is licensed in all 50 states and can be isolated using the CSOR database classification of occupations and the SOC system that RegData uses. We

selected the 4 states that were included in the first policy brief for OL RegData:[8] Indiana, Maryland, Ohio, and Pennsylvania. We also took data for the same states from the CSOR database.[9]

The OL RegData includes three variables: the number of restrictions, the word count, and the average sentence length. The number of restrictions varies substantially, from a high of 1,052 in Ohio to a low of 254 in Indiana. Likewise, the word count varies from 85,187 in Ohio to 22,615 in Indiana. The average sentence length was 127 words in Indiana, while it was only 17 words for Pennsylvania. These findings suggest that Ohio has the most burdensome licensing regulations, Maryland has the next most burdensome regulations, and Pennsylvania and Indiana have the least burdensome regulations.

The CSOR data contain a greater number of variables. The initial licensing fees range from a high of $325 in Maryland to a low of $75 in Pennsylvania. An associate degree is required in all four states. Indiana, Maryland, and Ohio all require three exams to obtain a license, while Pennsylvania requires just two exams. Ohio is the lone state with a minimum age requirement (18). All four states require good moral character to work in this occupation. Maryland requires 30 hours of continuing education biannually, while Indiana requires just 14 hours over the same period. Renewal fees range from $182 in Maryland to $42 in Pennsylvania. Finally, Indiana allows licensure by endorsement—a uniform pathway for a licensee to obtain a license in Indiana without restarting the licensing process—while Maryland, Ohio, and Pennsylvania accept out-of-state credentials on a more limited basis.

[8] Kofi Ampaabeng, Conor Norris, and Edward J. Timmons, "A Snapshot of Occupational Licensing Regulation in the Midwest and Mid-Atlantic States" (Mercatus Policy Brief, Mercatus Center at George Mason University, Arlington, VA, March 2020).
[9] CSOR Occupational Regulation Database, accessed January 5, 2021, https://csorsfu.com/find-occupations/.

Comparing the two datasets, some key differences emerge. Ohio has the most burdensome regulatory code related to dental hygienist licensing in the sample. However, fees and continuing education requirements in Maryland are more onerous than in Ohio. OL RegData provides information about Ohio's overall stringency and the difficulty of complying with the longer, more detailed code. On the other hand, the CSOR database provides a more detailed picture for individual dental hygienists, who may find that some key barriers to entry into the profession are more restrictive in Maryland. We should note again that Ohio is the only state with a minimum age requirement. It is not clear whether this restriction is binding (dentists may already choose not to hire dental hygienists under the age of 18). All of these differences are important, and there are instances where some questions are better answered by one method over the other or where the methods are best suited to answer different questions. As more data become available and as machine learning methods improve, further assessment will be needed to enhance our understanding of using each method.

Conclusion

Questions about state regulatory policy have been difficult to explore empirically because of the difficulty of gathering data across 50 states. Some research has explored specific regulations, but little has been done on the overall regulatory approach. Recent advances in machine learning have allowed researchers to read and gather data on entire state regulatory codes, opening new research possibilities.

We compare machine learning text analysis from OL RegData to an existing database, CSOR, that relies on data gathering done by hand by a team of people. We describe the process of each. While the CSOR method can obtain specific requirements that individuals face, it may introduce human error and judgment, and the speed of the dataset assembly is slow. The OL

RegData allows researchers to gather broad state or profession-specific data on the overall burden of licensing regulations, but this method sacrifices some specific measures affecting individuals. The direct comparison of both methods applied to one occupation, dental hygienists, suggests that the datasets are useful for exploring different research questions. As the data collection methodology improves and more data become available, it will be interesting to reassess this question moving forward.

Appendix

Table 1. Occupational Licensing RegData

State	Occupation (SOC Code)	Restrictions	Word Count	Average Sentence Length (Words)
Indiana	Dental hygienists (29-2020)	254	22,615	127
Maryland	Dental hygienists (29-2020)	840	73,174	29
Ohio	Dental hygienists (29-2020)	1,052	85,187	37
Pennsylvania	Dental hygienists (29-2020)	349	35,419	17

Source: Data from Kofi Ampaabeng, Conor Norris, and Edward J. Timmons, "A Snapshot of Occupational Licensing Regulation in the Midwest and Mid-Atlantic States" (Mercatus Policy Brief, Mercatus Center at George Mason University, Arlington, VA, March 2020).

Table 2. Knee Center for the Study of Occupational Regulation Database

State	Occupation	Type of Regulation	Licensing Fees	Degree	Number of Exams	Minimum Age	Good Moral Character	Continuing Education	License Renewal Fee	Endorsement
Indiana	Dental hygienists	Licensed	$100	Associate	3		Yes	14 hours	$70	Endorsement
Maryland	Dental hygienists	Licensed	$325	Associate	3		Yes	30 hours	$182	Limited
Ohio	Dental hygienists	Licensed	$184	Associate	3	18	Yes	24 hours	$144	Limited
Pennsylvania	Dental hygienists	Licensed	$75	Associate	2		Yes	20 hours	$42	Limited

Source: Data from the CSOR Occupational Regulation Database, accessed January 5, 2021, https://csorsfu.com/find-occupations/.