Occupational Licensing RegData 1.1 User Guide

Kofi Ampaabeng, Jonathan Nelson, Walter Stover,
and Stephen Strosko

October 10, 2020

1 Purpose

The Occupational Licensing RegData (OL RegData) belongs to the Mercatus Center's RegData suite of products. Using the QuantGov platform, the OL Data catalogs the list of occupations that require licensing in 44 US states. The dataset includes the following output, like all RegData products: the probability that the unit of regulation is OL, the total number of restrictions, including the types of restrictive terms (shall, must, may not, required, prohibited), word counts, the complexity of the text, and the industry relevance.

2 Content

OL RegData identifies the occupations and licensing regulatory restrictions in US states. Version 1.1 of OL RegData now includes all states with data available in State RegData 2.0. In building OL RegData, Mercatus researchers used the QuantGov platform to train an algorithm to predict the probability that a unit of regulatory text imposes occupational licensing regulations in a jurisdiction. The training involved defining a unit of regulation for each state. This task was necessary because states organize their regulatory/administrative codes differently.

Occupational licensing regulations typically include multiple facets – the requirement to be licensed in order to operate in the regulated area, educational requirements, fees, professional obligations, reciprocity across states, criminal history requirements, etc. Thus, a unit a unit of regulation is considered to pertain to occupational licensing if it addresses any of these facets. This approach often required smaller units of regulations to be aggregated to form a coherent document regulating an occupation.

3 OL Probability

The key metric of OL RegData is the probability that a unit of regulation imposes occupational licensing restrictions on an individual or establishment.

MERCATUS CENTER
George Mason University
QuantGov

The OL RegData includes only those units of regulation that meet a minimum probability threshold. We didn't establish the threshold a priori. Rather, each state's threshold was calculated from the probability distribution [Technical notes 1, 2].

4 Technical Notes

1. The performance of the classification algorithm varied across states. However, within each state, there is a clear bimodal distribution of predicted OL probabilities. This allowed for the development of state-specific thresholds for inclusion.

2. The threshold for a state is calculated as the mode OL probability (for a continuous variable, this is the highest value) minus twice the standard deviation of all predicted probabilities above the F1 score.

Table 1: Variable Descriptions

Variable	Description Definition
document reference	The reference for the document used as a source for the prediction algorithm.
document title	The title of the document if available.
ol probability	The probability that the unit of regulation imposes some form of occupational licensing restrictions/requirements.
restrictions	The total number of restrictions, comprises the sum of "shall", "must", "may not", "required", "prohibited".
shall	Occurrences of the word "shall" in the unit.
must	Occurrences of the word "must" in the unit.
may not	Occurrences of the words "may not" in the unit.
required	Occurrences of the word "required" in the unit.
prohibited	Occurrences of the word "prohibited" in the unit.
words	The total number of words in the unit.
industry code	Industry classification code (NAICS).
industry probability	The probability that the regulatory unit pertains to industry identified by "industry" variable.

Table 2: Change Log

Version	Release Date	Release Type	Notes
1.1	October 2020	Minor	Includes all states in State RegData 2.0. Increased the accuracy of state thresholds.
1.0	March 2020	Major	Includes only 37 of the 50 U.S. States.