Appendix: Text Analysis

The second approach to estimate the ideological positions of the legislators is through speeches. For this, we will be using text analysis tools to scale words to ideal points. Language-based models have gained popularity, as speech also reflect political actors ideological preferences. There are certain expectations and costs associated with speech related to reputation, party loyalty and representation, which represent the values of the political agent.

Table 1: Number of speeches by legislators

Legislatura	Mean	SD	Max	Min	N*
LX	11.31	14.54	113	1	471
LXI	20.69	61.71	923	1	464
LXII	20.74	38.91	483	1	467
LXIII	16.05	17.35	108	1	473
LXIV	18.63	20.4	285	1	440

N = number of legislator with at least one speech recorded in the parlaimentary debate. There are 500 legislators per term

1.0 Text Model

To extract the political positions from the speeches, we used a scaling technique called **wordfish** developed by Slapin and Proksch (2008), which estimate policy positions based on word frequencies in texts. This is an unsupervised approach, so we do not require the use of reference text to classify the positions. Instead it assumes an statistical distribution of word counts. A text is represented as a vector of word counts, individual words are assumed to be distributed at random (the probability that each word occurs in a text is independent of the position of other words in the text). In Wordfish, the frequency with which politician i uses word k is generated by a Poisson process:

$$y_{ik} \sim Poisson(\lambda_{ik}) \tag{1}$$

$$\lambda_{ik} = exp(\alpha_i + \psi_k + \beta_k * \omega_i) \tag{2}$$

With ω_i as the politician's ideological position and β_k as the discrimination parameter of word k (the weight capturing the importance of word k in discriminating between ideological positions) the other parameters refer to actor and words fixed effects. The main idea is that words wth negative β_k will tend to be used more often by politicians with negative ω_i. So, politicians with similar ideological positions, will tend to use similar words. The parameters of interest are the ideological position of the actors and the β_k because they allow us to analyze which words differentiate between party positions. Wordfish uses an expectation maximization algorithm to retrieve maximum likelihood estimates for all parameters.

2.0 Document Processing

Document processing is a fundamental step in every text analysis. Every text was pre-processed by the standard steps in text analysis:

(i) Removing punctuation

(ii) Removing numbers

(iii) Apply lowercase

(iv) Removing stop words (most common words in a language such as *a, but, of, if...*)

(v) Porter stemming Algorithm to reduce words to their common roots: the advantage is that similar words will be captured as one (*A stemming algorithm might also reduce the words such as cats, catlike, and catty to cat*).

(vi) Remove infrequent words with Sparsity criteria of 0.99 (words that appear at least in 1

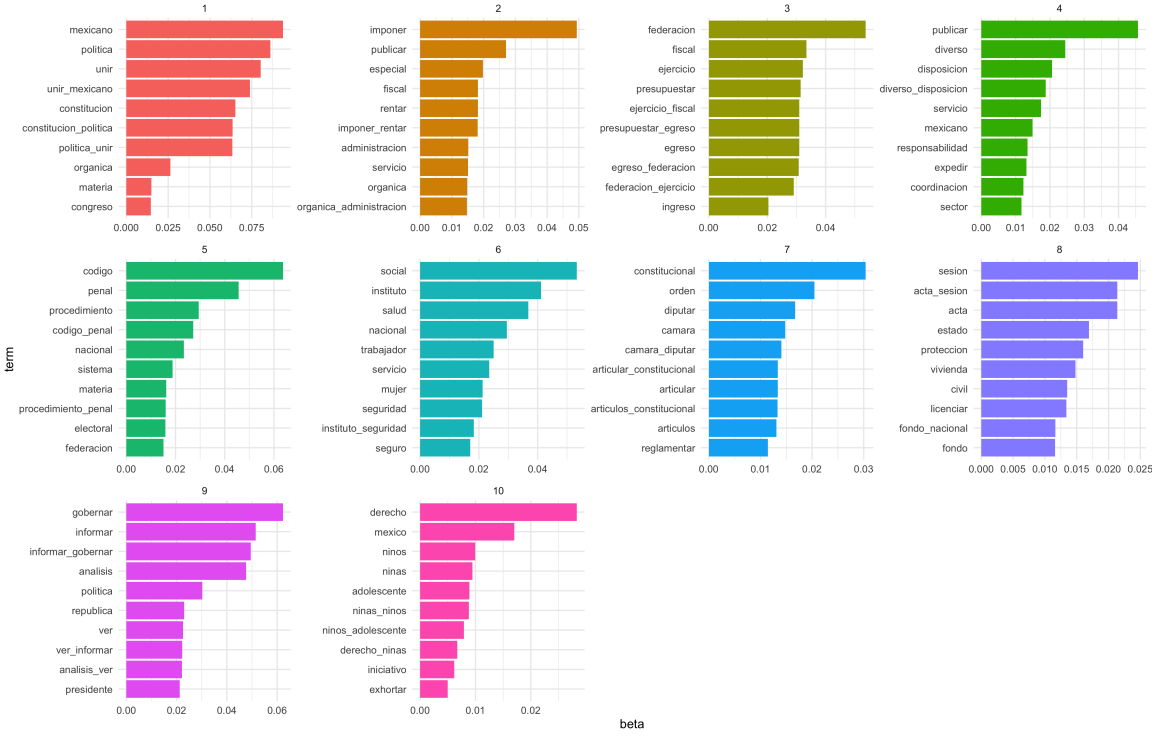
2.1 Policy Dimension

Legislators in the parliamentary debate are able to speak about any topic. This might be a problem for the estimation because will be much harder to identify words common in every topics that can differentiate ideology positions. For this purpose we are going to estimate the ideal point for different policy positions.

To obtain the relevant topics across documents, we applied a **Latent Dirichlet Allocation** (LDA). This is o is a particularly popular method for fitting a topic model.It treats each document as a mixture of topics, and each topic as a mixture of words, allowing documents to overlap each other in terms of content. Each term has an associated probability of being generated from each topic.

To apply the LDA we aggregate all speeches of all legislative terms and and based on the law they were talking about in the speech, we applied the LDA (this information is provided by the official website of the congress). Initially we determine a classification of 10 topics across all documents. The resulting terms that better describe each topic are the following: (i) constitution, (ii) public administration, (iii) fiscal policy, (iv) accountability, (v) justice, (vi) health, (vii) legislative studies, (viii) labor, (ix) governance, (x) human rights.

The following graph contains the the terms with the highest probability of being generated for each topic. These 10 terms are the most common within each topic.

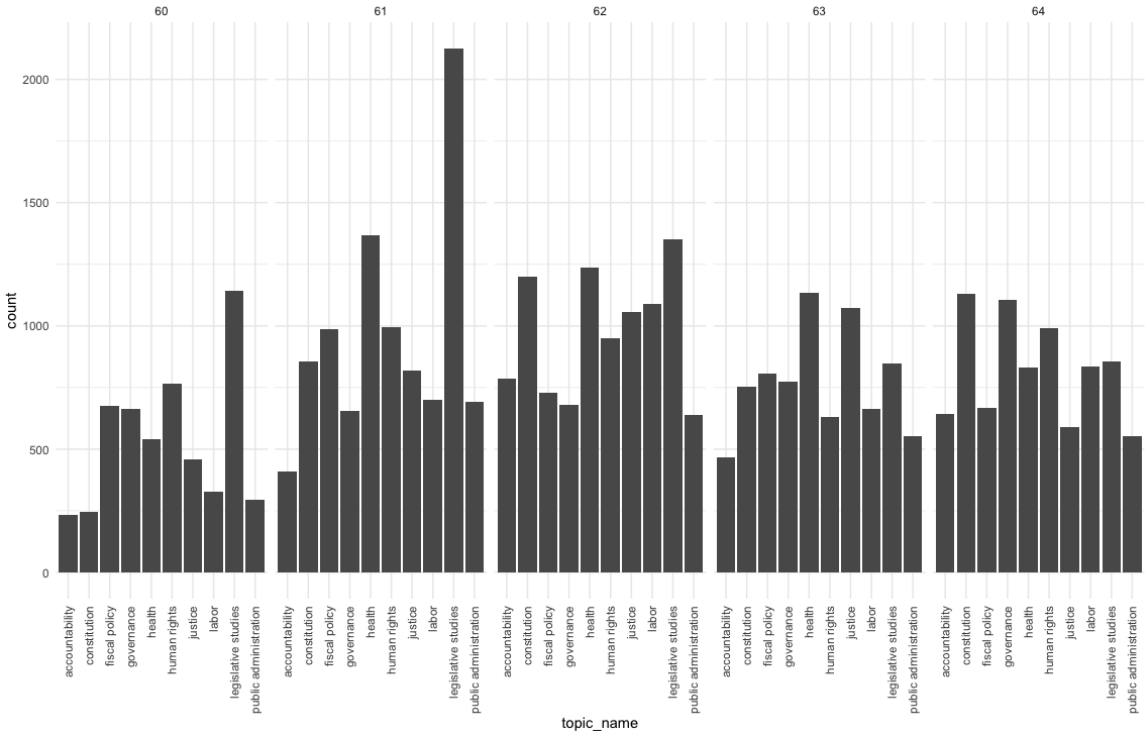


Topic 3: **fiscal policy**

Topic 2: **Health**

The Wordfish algorithm will be applied to each topic for each congress and members. With this, we are going to be able to obtain a ideal point for each member in each policy dimension.

Distribution of speeches in each Congress



2.2 Document-term matrix

After document processing and classifying the speeches to each topic, we will generate a **document-term matrix**. This object is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. A row correspond to a document (in this case the aggregated speech of one legislator in one topic) and columns correspond to terms. For this case, we considered unigrams and bigrams to construct the vocabulary of terms (n-gram is a contiguous sequence of n items (words) from a given sample of text).

2.3 Ideal Point Estimation

References

Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts" American Journal of Political Science 52(3): 705-722.