Is there an evidence base for evidence-based policy-making?

Motivation

There's generally been a big push in the UK in recent years towards evidence-based decision-making but relatively little research (as far as I'm aware) into whether or how providing evidence changes decisions of policy-makers making real decisions. For example, does evidence reduce confirmation bias or are decision-makers selective about which evidence they use and how much they scrutinise evidence they disagree with? Do decision-makers use a different process to come to a decision when there are evidence available to them? How do decision-makers arbitrate between conflicting evidence? How do decision-makers deal with a cacophony of evidence? How do they interpret the uncertainty and caveats associated with the evidence? I believe there's a bit more research on best ways to present evidence (e.g. the importance of visualisations) but there's still a long way to go before most of the evidence found gets presented this way.

So I'm generally interested in this question of understanding how providing evidence affects decision-making from both a self-interested perspective of whether I personally was making any difference, and also from a broader perspective of how to do it better so that we can get more evidence-based policy.

My intended plan was to compare green papers (initial policy documents in the UK) to ideals of evidence-based documents. Unfortunately, green papers are not easily accessible via scrapping or an API (as far as I could investigate - please let me know if you know otherwise!). So I decided to focus my efforts on debates in parliament - a slightly different group of people who are less focused on selecting an evidence-based implementation of a policy but who are nonetheless involved in selecting which issues to focus on. This prioritisation process no doubt involves other considerations than which issue they could have the most impact on, for example, what is likely to get them re-elected but I think is a useful exercise nonetheless.

Goal

My goal was to get an idea of how evidence-based each of the speeches were. So I compared how similar each of them was to 'ideal' evidence-based speeches, and used these similarity scores as a proxy for 'evidence-basedness'. I then used evidence-basedness as an input into quantitative models to investigate whether:

a) Debates had become more evidence-based over time
b) Specific interventions designed to increase the evidence-basedness of policy had affected the evidence-basedness of debates.

Accessing the Data

Theyworkforyou provides an API to make parliamentary debates more easily accessible. I initially used this API but, on advice from Theyworkforyou, used rsync access all of the xml

files from 1935 to the present day as the rsync method is faster. I stored all of the xml files in an AWS volume. I parsed the data using BeautifulSoup and then stored it in a csv file once it was in tabular format. The format of the xml files changes over the years and so it was a little fiddly extracting the necessary information and required considerable error handling. (I did store a small subset of the data in a MongoDB on my local machine to play around with this storage format). I selected debates from the years 2000 - 2018 to focus on the current evidence-based movement (although I would like to investigate how the evidence-basedness has changed over time too). This gave me about 800,000 speeches.

Preprocessing

To prepare the text data for analysis, I stemmed and lematised the words. This made sure that plurals and conjugations of the words didn't show up as different features. I also removed stop words (e.g. 'the', 'a') which are effectively noise in the context of NLP. I then created a 'bag of words' with words and bigrams (two concurrent words) so that I could take into account negations (e.g. 'not good') and qualifiers (e.g. 'terribly bad'). I then translated this bag of words into a TF-IDF (term frequency inverse document frequency) matrix where each speech ('document') is represented by vector of words which each have a score. The score represents how frequent the word is in that speech relative to how frequent it is in other speeches. This gives an idea of how important the word is to define that document uniquely - words which are frequent across all documents in the corpora have a lower score than words which are frequent in that document but not elsewhere. In making the TF-IDF matrix, I experimented with the parameters (minimum frequency, maximum frequency, maximum number of features) to pick up as much signal as possible whilst avoiding running into memory errors (my 100GB AWS volume was still struggling!) This gave me about 1200 features and so I needed to reduce the dimensionality.

I used LSI to reduce the number of features to a more managable number of components. Not only would this allow me to calculate the similarity more quickly but it might also pick up more signal. I chose 300 components as a starting points for LSI, and would like to investigate further the optimal number using singular value elbow plots when I have more time. I would also like to try NMF, and also to do more manual inspection of the speeches said to be similar to the ideal scientific ones. I am not too concerned about the interpretability of the components but would be interested to see whether my human intuition of which speeches are similar matches up better with NMF as a test of the model.

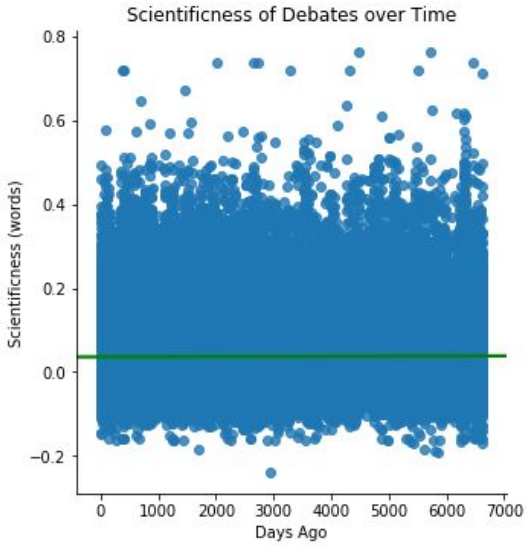
Following dimensionality reduction, I calculated a similarity score for each speech with my ideal evidence-based speeches. I then averaged across the similarity scores for each speech to all of the science lectures to take into account that there are different topics in the lectures and speeches (e.g.) and I'm interested in the kind of language and argumentation used rather than the topic per se. The ideal evidence-based speeches were the Christmas lectures from the Royal Institution which are given to help increase the public understanding of science. I also used a bag of science words such as 'research', 'data' and 'average' as a simple comparator.

Have debates become more evidence-based over time?

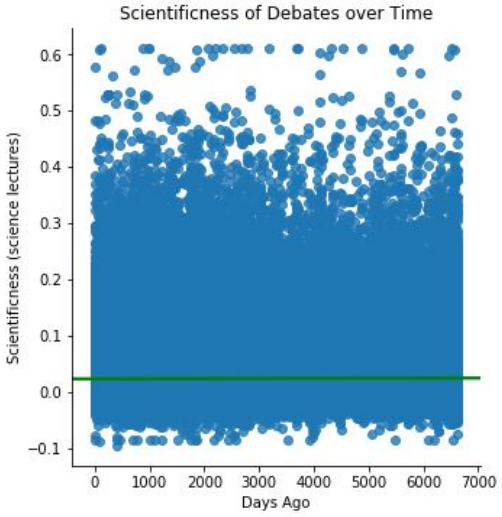
I plotted evidence-basedness against time, and also calculated the Pearson's correlation between evidence-basedness and time. There was a significant, negative correlation between 2000 and 20018 but it was small (0.001*** for both the science lectures and the science bag of words), especially in comparison to the amount of variance in evidence-basedness in speeches. However, I haven't included any other control variables, and this would be the next step. I would also be interested in looking over a longer timescale to understand how this trend has evolved over time.

(Please note that Graphs 1 and 2 have the number of days ago on the x-axis and so show increasing evidence-basedness as you go back in time.)

Graph 1



Graph 2



Have specific interventions to increase the evidence available helped?

There have been several bodies set up in the last 10-20 years with the aim of increasing the amount of evidence available to policy-makers and decision-makers. (The are collectively known as What Works Centres). For example, the Education Endowment Foundation was set up in 2011 to focus on increasing the evidence base in education. I was interested in whether such interventions make a difference in making their area of policy more evidence-based. I use the Education Endowment Foundation and the policy area of education as an example but would be interested in expanding this analysis to all of the What Works Centres in the future. I conduct a difference-in-difference analysis to try to estimate the causal impact of the EEF on the evidence basis of parliamentary debates talking about education. The intuition is that if the EEF has an effect on the evidence-basedness of education speeches, then the trend in evidence-basedness of non-education speeches will continue as before but the trend in the evidence-basedness of education speeches will change after the EEF has been set up. This assumes that the trends were the same before the period of analysis (an assumption I still need to test more rigourously).

I conducted OLS using the following features:

$Y = \beta0 + \beta1*[\text{Time Dummy (Before / After2011)}] + \beta2*[\text{Education Speech Dummy}] + \beta3*[\text{Interaction of time dummy and education dummy}] + \varepsilon$

The coefficient on the interaction term is the difference-in-difference coefficient. It was significantly negative in this case ('educ_time' in Table 1 below). This suggests that the EEF actually had a negative effect on how evidence-based the education-related parlimentary debates are. Although I note that I need to do more rigorous checking of the parallel trends assumption and that perhaps the impact of the EEF would be felt more in policy-making more directly, for example, in green papers (as I'd originally intended to study).

Table 1: Difference-in-difference regression: the effect of the EEF on the evidence-basedness of education-related speeches in parliament

	coef	std err	t	P>|t|	[0.025	0.975]
Education	0.0319	0.000	122.502	0.000	0.031	0.032
Time	0.0236	0.000	223.699	0.000	0.023	0.024
educ_time	-0.0211	0.000	-54.737	0.000	-0.022	-0.020

Omnibus:	374577.679	**Durbin-Watson:**	1.469
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	9263774.896
Skew:	3.139	**Prob(JB):**	0.00
Kurtosis:	22.546	**Cond. No.**	4.54

Recommending evidence champions

It may be that recognising and promoting those who use evidence in their speeches could improve use of evidence in others. To enable this as a strategy, I investigated which MPs or former MPs had speeches which were the most evidence-based. I identified two whose speeches were significantly more evidence-based than average: Joan Ruddock and Angela Eagle.

How evidence-based are specific topics of interest?

I tested whether debates which mentioned Brexit were more or less evidence-based (on a smaller subset of the data on my local machine). Contrary to my expectation, they were significantly and substantially *more* evidence-based. This made me think about what type of results this analysis can give me. It can tell me whether more scientific language was used but says nothing about whether the claims are true. For example, in the Brexit debate, lots of numbers were flung around which have since been found to have very little evidence base. It would be interesting to investigate whether I could pull in data from fact-checking websites to corroborate the facts which the MPs talk about. I would like to investigate whether these results hold using all of the data.

Conclusion

This project shows initial evidence that parliamentary debates in the UK became less evidence-based over the period 2000 - 2018, and that interventions designed to increase the evidence basedness actually decreased it. A major limitation of this research is access to what an ideal evidence-based speech would look like, and also that whilst parliamentary debates are interesting, it would be better to look directly at policy documents.

Future Steps

I would like to use this project as a proof of concept for the use of text data in analysing how people talk about evidence. In order to see whether the similarity scores are giving me an insight into the evidence-basedness of speeches, I would need to ask people with some expertise in evidence-based thinking to rate the evidence basedness on a subset of the speeches and see how well this aligns with the similarity scores. I would be interested in discussing what these experts also think is important in defining a speech as evidence-based and whether they can recommend other comparators as I believe the analysis would be highly sensitive to the comparators. It could be interesting also to look more broadly at whether such analysis can be applied to accessing the logic / rationality of someone's arguments.

I would be interested in investigating the impact of the other What Works Centres, and also looking at trends of evidence-basedness over a longer time period.

What I learnt

The biggest learning curve for me in this project was setting up an AWS instance and volume and trouble-shooting the associated problems. I learnt to deal with broken pipes (using screen or nohup), and also mitigating the effect of Jupyter Notebook crashing on my AWS instance by pickling all my data and models. I would like to spend more time tuning the parameters of my TF-IDF and dimensionality reduction to pick up more of an intuition around this.