Lab Notes 2

Scott Patterson

May 19, 2020

1 ADMIN

Hours: 9
Total: 62
Still waiting for update from purchasing department on the computer

2 Overview

Since the previous update, I have :

- Added UN regional groupings to the dataset. This includes dummy variable columns for AFRICA, ASIAPAC, WEOG, GRULAC, EASTEUROPE and a 'UN_REGION' variable that includes each region as a separate factor.
- Generated *topical prevalence* models at k values of 18, 24, and 30 that use 'UN_REGION' as a prevalence covariate.
- Used the 'findThoughts' function to isolate documents that best exemplify certain topics.
- Used the 'estimateEffects' function to investigate regional differences in the prevalence of topics relating to socioeconomic inequality. The preliminary findings suggest that Latin American states devote the most time to this topic while WEOG and Eastern European states play the topic down.
- Plotting: I had no issue generating a plot to show the general prevalence trend of inequality over time. I also plotted the difference in prevalence by region, albeit the graphic needs improvement. Visualizing regional differences over time has proven more challenging.
- Started a *topical prevalence* model using 'year' as an interaction term. This will allow for a deeper investigation into the relationship between time and region.

3 Findings

3.1 Covariate Addition

I had one minor issue while adding the regional covariates. After debugging several failed models, I noticed that 252 entries did not cluster into one of the defined regions. I haven't fully investigated why, but I suspect the NA cases were from countries that broke apart or no longer exist. This may merit further investigation down the road, as we will need to deal with the addition/subtraction of new countries since 1970.

NA cases aside, the regional covariates have been added to the preprocessing stream and are now available as a covariates.

3.2 Prevalence Models

This section contains the prevalence model word lists for the 18 topic model. There are several topics that are potentially related to socioeconomic inequality. This analysis zooms in on Topic 5.

Based on the noted keywords, Topic 5 warrants further exploration as a potential container for tokens relevant to socioeconomic inequality. This topic also appears to capture something related to drug conflicts in Latin America. Bearing this in mind, it will be useful to see if this topic breaks down into subtopics at higher K values.

Figure 1: Topical Prevalence: k = 18, thresh = 10, weighting = prob

(a) Topic 5 keywords: develop, trade, social, debt

Figure 2: Topical Prevalence: k = 18, thresh = 10, weighting = frex

(a) Topic 5 keywords: debt, export, adjust, flow

Figure 3: Topical Prevalence: k = 18, thresh = 10, weighting = lift

(a) Topic 5 keywords: hyperinfl

Figure 4: Topical Prevalence: k = 18, thresh = 10, weighting = score

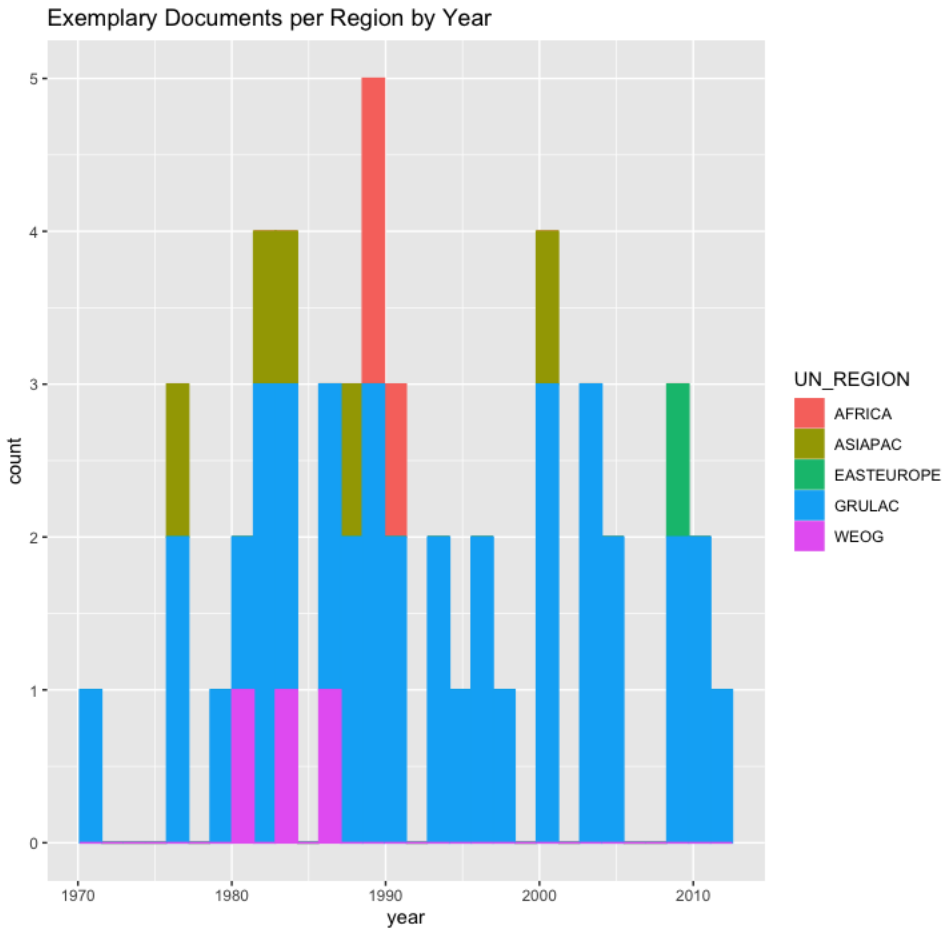
(a) Topic 5 keywords: develop, debt

3.3 findThoughts

findThoughts is a function that isolates documents that contain the highest prevalence of a specified topic. The output of the function is a vector of indices that correspond to the location of the specified documents within the corpus. Substantively, this function aims to identify documents from the corpus that best exemplify the topic in question.

This figure depicts the number of exemplary documents of Topic 5 per year by region. Interestingly, 38/50 speeches with the highest proportion of Topic 5 were delivered by Latin American (GRULAC) countries. The document frequency peaked in 1990 and there were relative peaks in the early 80s and 00s. The only WEOG documents appear in the early to mid-80s.

Figure 5: Exemplary Documents per year by region



UN_REGION	Number of Exemplary Documents
AFRICA	3
ASIAPAC	5
EASTEUROPE	1
GRULAC	38
WEOG	3

The best way to act on this finding is to qualitatively analyze a group of the exemplary documents to get a sense of the themes. I can provide a list of these documents at your request.

3.4 estimateEffect

estimateEffect is a function used to estimate the effect of a specified covariate on the prevalence of a topic. In this case, I specified the factor variable 'UN_REGION' plus the non-parametric 'year'. I can't yet fully explain the non-parametric "spline" that captures year in the model, but to my best understanding, this means that there is

no presumption of a linear relationship (thus non-parametric) between topic prevalence and year. Running these models made me realize that to capture the effect of year, I'll need to try it as an interaction term in a separate model.

Examining the effects table for the *magnitude* and *direction* of effects yields interesting results.

Figure 6: estimateEffect: Regional Grouping v. Topic 5 Prevalence

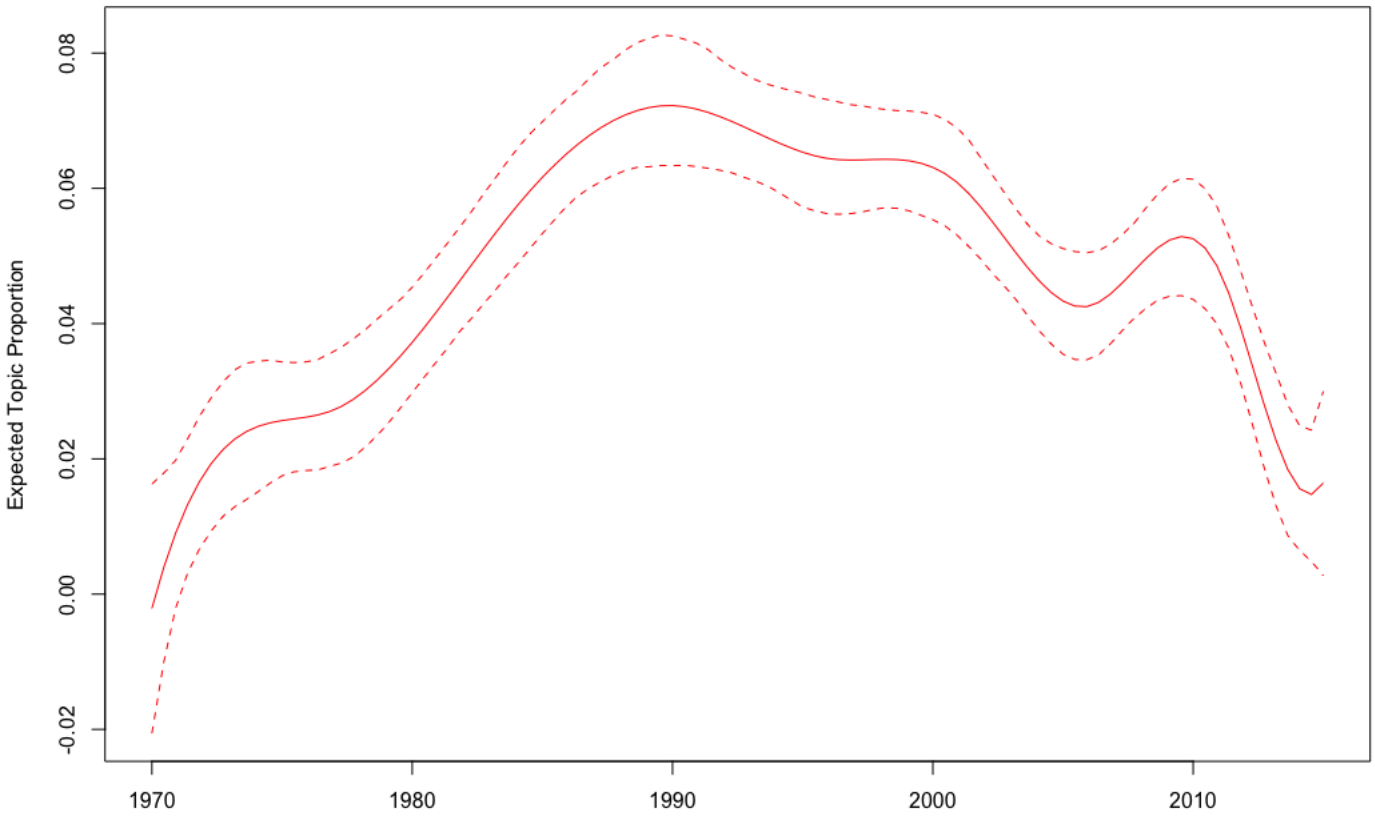
```
Topic 5:

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.002363   0.009419  -0.251  0.80191
UN_REGIONASIAPAC     -0.002227   0.003090  -0.721  0.47116
UN_REGIONEASTEUROPE -0.022634   0.004186  -5.407 6.63e-08 ***
UN_REGIONGRULAC       0.089397   0.003866  23.125  < 2e-16 ***
UN_REGIONWEOG        -0.010798   0.003704  -2.916  0.00356 **
s(year)1              0.033800   0.017320   1.951  0.05104 .
s(year)2              0.018112   0.012693   1.427  0.15362
s(year)3              0.061975   0.013369   4.636 3.62e-06 ***
s(year)4              0.080586   0.012033   6.697 2.28e-11 ***
s(year)5              0.062605   0.012472   5.020 5.30e-07 ***
s(year)6              0.071133   0.011628   6.117 1.00e-09 ***
s(year)7              0.029733   0.012557   2.368  0.01792 *
s(year)8              0.076799   0.013732   5.593 2.31e-08 ***
s(year)9              0.009838   0.013112   0.750  0.45307
s(year)10             0.018592   0.011732   1.585  0.11308
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There appears to be a negative relationship between Topic 5 prevalence and the WEOG and Eastern Europe groupings.

The relationship between GRULAC and Topic 5 prevalence appears positive and at higher magnitude than with any other grouping. Again, it appears that the Latin American states devote much more of their UNGA floor time talking about Topic 5 than any other regional grouping.

I was able to generate a few plots from the estimateEffect object. The first one depicts the general prevalence of Topic 5 over time. For this figure, the y-axis depicts the estimated proportion of the overall corpus devoted to topic 5 over time.

Figure 7: Topic 5 Prevalence over time



I also generated a plot that depicts Topic 5 prevalence by region and a plot for the prevalence of Topic 5 relative to other topics. However, I haven't yet been able to simultaneously capture regional differences and time, which leads me to think I need to come up with a new STM model that uses time as an interaction term not as a non-parametric additive.

Figure 8: Topic 5 Prevalence by Regional Grouping

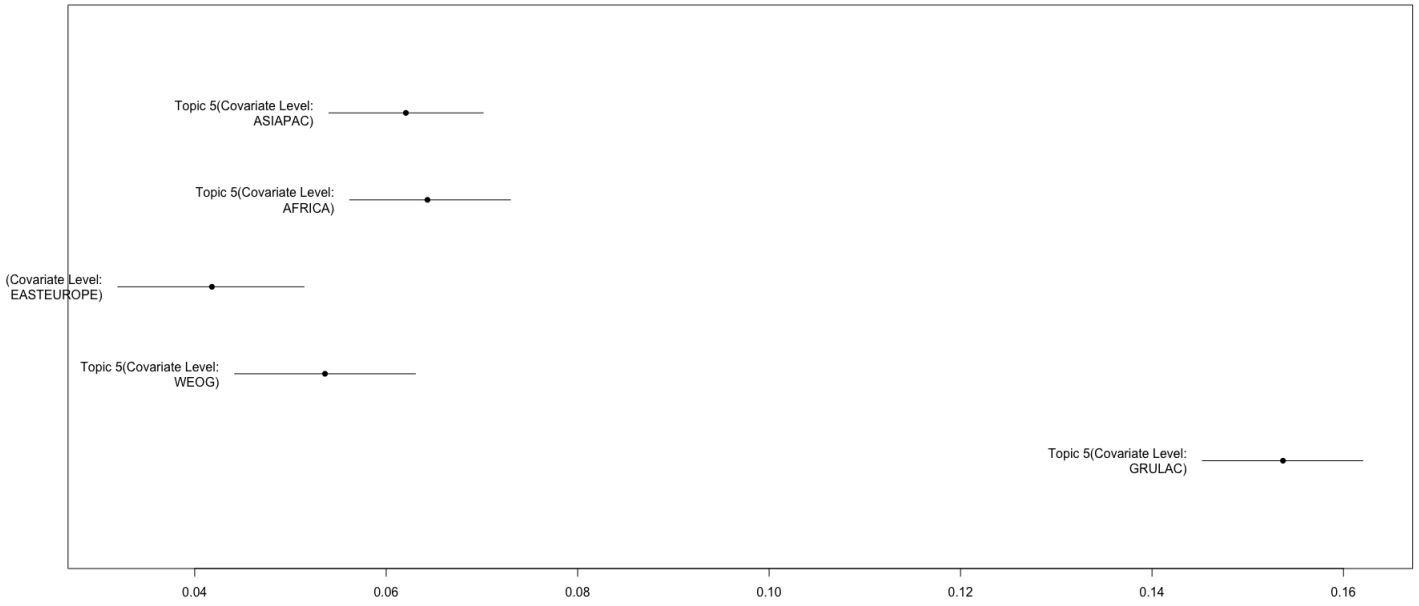
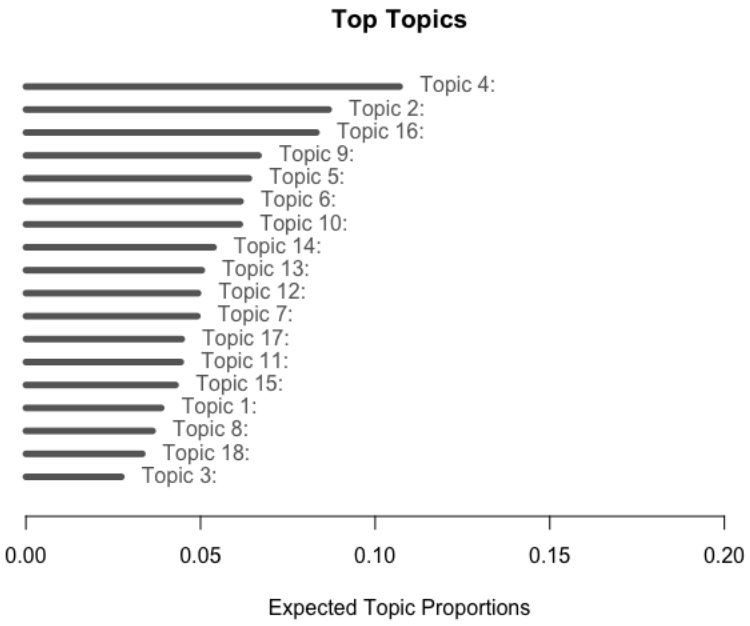


Figure 9: General Topical Prevalence

(a) Note that Topic 5 is the 5th most prevalent topic overall



4 Forward

My first priority for the next stretch will be to continue working on the k18 model while altering the underlying effect estimate to account for the interaction between regional groupings and time. The low hanging fruit here

will be the regional dummies, which will allow me to visualize the difference between each region and the average topical prevalence. I have already found some samples that will help me to write the syntax for this as soon as I generate the new model. The more difficult step will be to visualize the relative prevalence of each regional grouping simultaneously. This will be more difficult because the 'estimateEffect' object generated by STM is not easily transferable to other R packages. STM contains several basic visualization features as well as some highly advanced ones, not much in the middle. In any case, this is nothing that can't be figured out without a bit of tinkering.

When I have finished with the step above, I'd like to replicate all of these results with the k24 and k30 models. I expect that in these models, Topic 5 will devolve into several more precise topics. This may isolate the inequality tokens from the war on drugs tokens. If not, it could also be possible that inequality and the war on drugs are rhetorically-proximate, at least in the Latin American case. We can also explore this possibility by reading through some of the documents that popped up from the 'findThoughts' test.

I zeroed in on Topic 5 because it seemed like a good fit at a close glance. Some of the other topics also contained tokens that may make reference to socioeconomic inequality. Now that I've built the pipes to run these tests, they can quickly be replicated on other topics.

The regional covariates should also be useful when returning to the *topical content* models. One issue that I've encountered before is that Topic 5 won't necessarily appear in exactly the same way for a content and prevalence model. I think that I can now remedy this with an intermediate fix that I learned from Prof Erlich. When I have the new computer, I can use the extra computational power for a guaranteed remedy here, as I can generate both a content and prevalence model simultaneously.

Finally, I'm also working on a more dynamic approach to visualization that will allow you to tinker with some of the STM specifications from your web browser. I think this will be better than these snapshots that I am presenting here.

I look forward to discussing these results at your convenience.