Draft June 2020

Patrick Schulze, Simon Wiegrebe

June 2020

Contents

1 Introduction

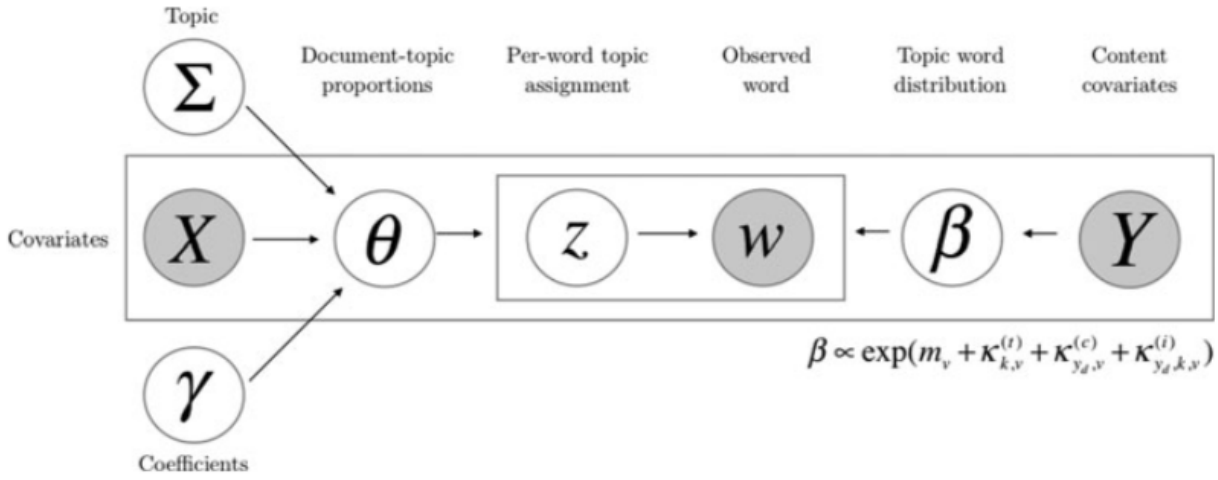
TBD

2 Theoretical Framework

2.1 STM - Introduction

The Structural Topic Model (STM) is a topic model which extends classical topic models such as Latent Dirichlet Allocation (LDA) by incorporating information of covariates. Topic models can be used to infer topics from a large text corpus grouped into documents. In topic modelling it is assumed that this corpus is generated from a small number of distributions over words, the topics. The proportions of these topics are document-specific. In contrast to simpler topic models such as LDA, the STM relates topic proportions to document-level covariates. Furthermore, each distribution over words, i.e. each topic, can vary for different documents dependent on the covariate values of this document.

The detailed mechanism underlying the STM can be illustrated using its graphical model representation (see Figure 1). As outlined above, for each document indexed by $d \in \{1, \ldots, D\}$ there exists a $K-1$-dimensional vector θ_d of topic proportions. Topic proportions are assumed to depend on P document-specific so-called topical prevalence covariates $X \in \mathbb{R}^{D \times P}$, by following a logistic normal distribution with mean $X_d\Gamma$, where $\Gamma = [\gamma_1|\ldots|\gamma_k]$, and covariance Σ. Each of the N_d words in document d is subsequently assigned to one of the K topics dependent on the topic proportions θ_d; this per-word topic assignment is captured by the latent variable $z_{d,n} \sim \text{Multinomial}_K(\theta_d)$, where $n \in \{1, \ldots, N_d\}$ denotes the word index. As stated, the distribution over words that characterizes a topic can vary for each document dependent on document-specific covariates $Y \in \mathbb{R}^{D \times A}$, the so-called topical content covariates. A word $w_{d,n}$ is then the result of the assigned topic, expressed by $z_{d,n}$, the content covariates Y_d, and their interactions. More precisely, this last step is intuitively best understood as a multinomial logistic regression of the words on the latter variables. A word $w_{d,n}$ then ultimately follows a multinomial distribution with probabilities $\beta_{d,n} := \beta(z_{d,n}, Y_d) \in \mathbb{R}^V$, i.e. $w_{d,n} \sim \text{Multinomial}_V(\beta_{d,n})$, where V denotes the total number of distinct words in the corpus; for details on the exact specification of $\beta_{d,n}$ see M. E. Roberts, Stewart, and Airoldi (2016), p. 991. Thus the occurence of a word (which is equivalent to being drawn from a corresponding multinomial distribution) depends on the topic assignment as well as on the topical content covariates, where the topic assignment itself is a function of the topical prevalence covariates.



Graphical model representation of the STM (from M. E. Roberts, Stewart, and Airoldi (2016), p. 990)

2.2 STM - Scope

Topic models are unsupervised learning methods, since the true topics from which the text was generated are not known. Thus, traditionally topic models have been used as an exploratory tool providing a concise summary of topics, where it is hoped that the posterior induces a good decomposition of the corpus. Topic models have also been used for tasks such as collaborative filtering and classification (see e.g. Blei, Ng, and Jordan (2003)). In particular, they can be used as a dimensionality reducing method in semi-supervised learning methods. Such a process can in general be described as a two-stage approach, where in the first stage topic proportions and content are learned, and in the second stage a supervised method such as regression takes this learned representation as input.

The fundamental idea of STMs is to combine these two steps: Topics and their relation to covariates are jointly estimated. For instance, the estimated effect of topical prevalence covariates X_d on topic proportions is reflected in the estimate of Γ. However, since the topic proportions θ_d are random variables, it is a better approach to incorporate the uncertainty of θ_d, accesible through the estimated approximation of the posterior $p(\theta_d|\Gamma, \Sigma, X)$, when determinig the effect of covariates on topic proportions. This is achieved by what is called the "method of composition" in social sciences: By sampling from the approximate posterior for θ and

subsequently regressing these topic proportions on X it is possible to integrate out the topic proportions (since these are latent variables!) and obtain an i.i.d. sample from the marginal posterior of the regression coefficients for the topical prevalence covariates.

A problem we see with this approach is, however, that the same covariates and in general the same data used to infer the topical structure are subsequently used to determine effetcs of the former on the latter (or vice versa). This problem has recently also been adressed by Egami et al. (2018). In practice, in case of the regression coefficients for the topical prevalence covariates (obtained using the method of composition as outlined above), due to the regularizing priors for Γ we have found that the prevalence covariates have almost no influence on the estimated topic proportions. Thus the regression coefficients (with the topic proportions as the dependent variable) should not be largely affected by this problem. However, the question then appears why the covariate variables have been used to obtain the topical structure in the first place. In an empirical evaluation M. E. Roberts, Stewart, and Airoldi (2016) showed that the STM consistently outperformed other topic models such as LDA, when comparing the respective heldout likelihoods in different settings. This indicates that the STM performs better at predicting the topical structure by incorporating covariates, regardless of the concrete specification of these covariates.

Nevertheless, it should in each case be investigated whether the relationship of variables implied by the STM is valid. For instance, we have split our data into training and test sets and found that the topical structure predicted on the test set differs starkly from the structure on the training set. This could of course be caused by a misspecification of the topical prevalence and content variables. However, since the topical prevalence covariates have almost no influence on the estimated topic proportions on the training set due to the regularizing priors (and e.g. likewise on the heldout likelihood that can be used for validation), it is practically impossible to validate a good prevalence specification.

2.3 Posterior Distribution

The posterior given on p. 992, M. E. Roberts, Stewart, and Airoldi (2016), can be derived as follows:

$$
\begin{aligned}
p(\eta, z, \kappa, \Gamma, \Sigma | w, X, Y) &\propto \underbrace{p(w|\eta, z, \kappa, \Gamma, \Sigma, X, Y)}_{=p(w|z,\kappa,Y)} p(\eta, z, \kappa, \Gamma, \Sigma | X, Y) \\
&\propto p(w|z, \kappa, Y) p(z|\eta) p(\eta|\Gamma, \Sigma, X) \prod p(\kappa) \prod p(\Gamma) p(\Sigma) \\
&\propto \Big\{ \prod_{d=1}^{D} p(\eta_d|\Gamma, \Sigma, X_d) \Big(\prod_{n=1}^{N} p(w_n|\beta_{d,n}) p(z_{d,n}|\theta_d) \Big) \Big\} \prod p(\kappa) \prod p(\Gamma) p(\Sigma) \\
&\propto \Big\{ \prod_{d=1}^{D} \text{Normal}(\eta_d|X_d\Gamma, \Sigma) \Big(\prod_{n=1}^{N} \text{Multinomial}(z_{n,d}|\theta_d) \\
&\quad \times \text{Multinomial}(w_n|\beta_{d,n}) \Big) \Big\} \times \prod p(\kappa) \prod p(\Gamma) p(\Sigma),
\end{aligned}
$$

where $\beta_{d,n} := \beta(z_{d,n}, Y_d) \in \mathbb{R}^V$ with entries $\beta_{d,k,\nu} \propto \exp(m_\nu + \kappa^{(t)}_{k,\nu} + \kappa^{(c)}_{y_d,\nu} + \kappa^{(i)}_{y_d,k,\nu})$, $\nu \in \{1, \dots, V\}$, and $\theta_d := \text{softmax}(\eta_d)$.

3 Data

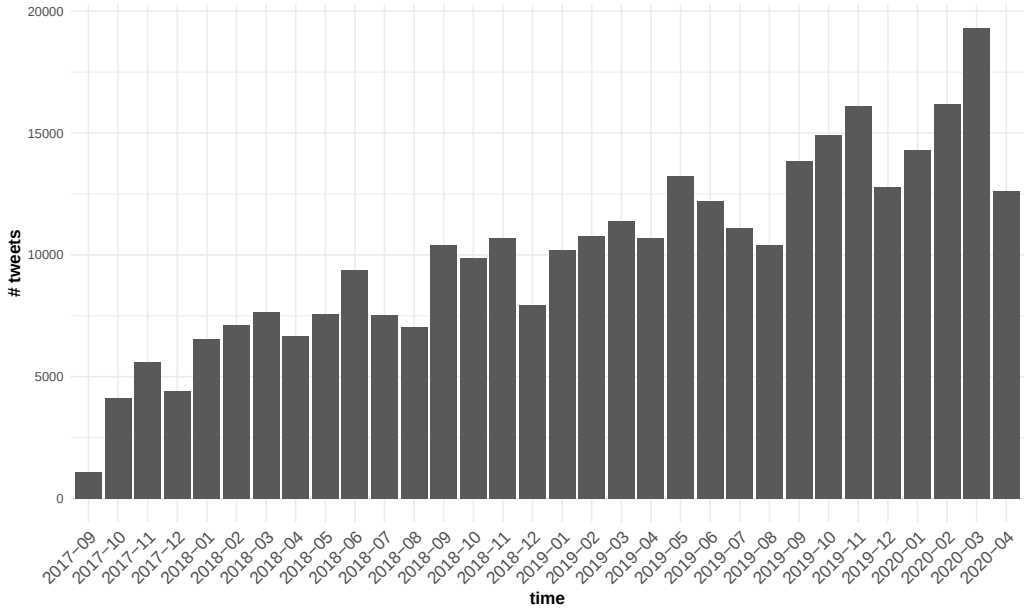
3.1 Data Collection

As a first step towards applying the STM to German political entities, we constructed a database with personal information about all German MPs. Using Python's BeautifulSoup web scraping tool as well as a

selenium webdriver, we gathered data such as name, party, and electoral district from the official parliament website for all of the 709 members of the German parliament during its 19th election period, elected on September 24, 2017. (Footnote: MPs who resigned or passed away since this date were also listed on the website and thus included initially; they were manually excluded from further analysis.)

Since information on social media profiles was scarce and incomplete on the official parliament website, we scraped official party homepages for each of the six political parties represented in the current parliament. MPs who did not provide a Twitter account either on the official parliament website or on their party's official homepage were excluded. Using Python's tweepy library to access the official Twitter API, we scraped all tweets by German MPs from September 24, 2017 through April 24, 2020, i.e., during a total of 31 months. (Footnote: tweepy restricts the total number of retrievable tweets to 3,200. For those MPs with a larger number of tweets, the most recent 3,200 tweets are taken into account. However, this only affects two MPs.) This initially yielded 342542 tweets from a total of 470 members of parliament.

To complement personal data, we also gathered socioeconomic data such as GDP per capita and unemployment rate as well as 2017 election results on an electoral-district level for all of the 299 electoral districts, from the official electoral website. After removing independent MPs as well as MPs without a specific electoral district assigned to them (for matchability with socioeconomic data), the final dataset counted 450 MPs. The corresponding total number of tweets amounted to 323740. The table below shows total monthly tweet frequencies for our period of analysis, September 24, 2017 through April 24, 2020. As can be seen, tweet frequencies - though fluctuating - increase over time, peaking at almost 20,000 in March 2020.



Next, data were grouped and tweets concatenated on a per-user level (thus aggregating tweets across the entire 31 months) as well as on a per-user per-month level, yielding a user-level and a (user-level) monthly dataset. This means that a document represents the concatenation of *all* of a single MP's tweets for the user-level dataset and a single MP's *monthly* tweets for the monthly dataset. This also means that MP-level metadata such as personal information and socioeconomic data (through the electoral district matching) can be used as document-level covariates. For the monthly dataset, the temporal component (year and month) constitutes an additional covariate. At this point, the data preparation was completed, thus marking the starting point of the preprocessing required for topic analysis, which is identical for both datasets.

3.2 Data Preprocessing

We used the *quanteda* package within the R programming language for preprocessing. As a first step, we built a quanteda corpus from all documents, already transcribing German umlauts *ä/Ä*, *ö/Ö*, *ü/Ü* as well as

German ligature *ß* as *ae/Ae*, *oe/Oe*, *ue/Ue*, and *ss* and removed hyphens. Next, we transformed the text data into a quanteda document-feature matrix (DFM), which essentially tokenizes texts, thereby convering all characters to lowercase. From the DFM, we removed an extensive list of German stopwords, using the stopwords-iso GitHub repository, as well as English stopwords included in the *quanteda* package. Moreover, hashtags, usernames, quantities and units (e.g., *10kg* or *14.15uhr*), interjections (e.g., *aaahhh* or *ufff*), terms containing non-alphanumerical characters, meaningless word stumps (e.g., *innen* from the German female plural declension or *amp*, the remainder left after removing the ampersand sign, *&*) were removed. Terms with less than four characters and terms with a term frequency (overall number of occurrences) below five or with a document frequency (number of documents containing the word) below three were excluded. Finally, we manually removed over-frequent terms that would diminish the distinguishability of topics, such as *bundestag* or *polit*.

We also performed word-stemming, which means cutting off word endings to remove discrepancies arising purely from declensions or conjugations - of particular importance for the German language. Due to the nature of the German language, the additional gains of lemmatization (which aims at identifying the base form of each word) would only be small as compared to the large increase in complexity, which is why we decided to use stemming only. Another issue when dealing with German language documents are compound words, which are sometimes hyphenated, basically leading to a distinction where semantically there is none. We addressed this issue by removing hyphens in the very beginning of the preprocessing and converting all terms to lowercase, thus "gluing together" compound words; this way, terms like *Bundesregierung* and *Bundes-Regierung* are both transformed into *bundesregierung* (and, after stemming, into *bundesregier*). Finally, automatic segmentation techniques were not necesssary for the German language (Lucas et al. (2015)).

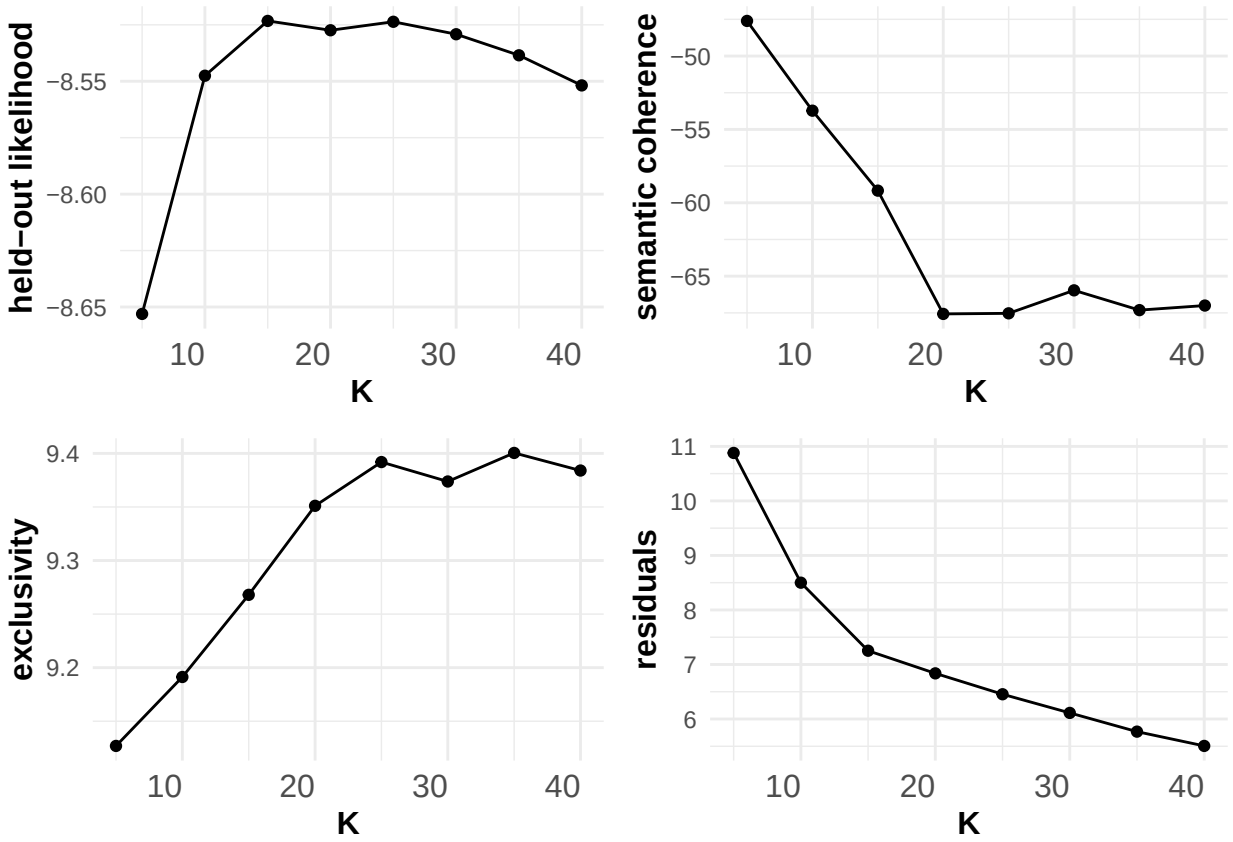
As the result of preprocessing, one empty MP-level document was dropped, so that a total of 10998 MP-level documents were eventually analyzed, each one associated with 90 covariates.

4 Results

4.1 Hyperparameter Search and Model Fitting

Throughout the topic analysis, we use the *stm* package, which is implemented in the R programming language (M. E. Roberts, Stewart, and Tingley (2019)). The most important choice when fitting an STM is the number of topics, K. While there is no *true* or *optimal* number of topics, we explore the hyperparameter space using the *searchK* function to get an understanding of the impact of K on model fit. We use four of the metrics that come with this function, *held-out likelihood*, *semantic coherence*, *exclusivity*, and *residuals*. As for the first one, the *searchK* function randomly holds out a proportion of some of the documents. This set of held-out words is then used to evaluate their probability given the trained model, giving rise to the *held-out likelihood*. Regarding the second metric, a model with K topics is *semantically coherent* whenever those words that characterize a specific topic (i.e., the most frequent words within a topic) also do appear in the same documents. *Exclusivity* basically tells us to which degree a topic's word *only* occur in that topic (for more detail, see the *FREX* methodology further below). Finally, *residuals* is a metric based on residual dispersion, which theoretically should be equal to one; so if the observed residuals exceed this value, the number of topics was most likely chosen *insufficiently small*.

Another aspect to be taken into account when choosing K (or, to be precise, when choosing a search grid for searchK) is interpretability. While a large K certainly allows for a more fine-grained determination of topics, the resulting topics might be rather difficult to label. Furthermore, for large K we would obtain many topics which could be considered sub-topics of the topics we would obtain when using a smaller value for K. The graph below shows the four metrics, as introduced above, for values of K between 5 and 40 (in steps of 5).



Both 15 and 20 topics seem to be good trade-offs between the metrics used. Taking into account the interpretability aspect, we opt for K = 15.

Before fitting the model, we need to choose the document-level covariates we want to include. Since a topic model is explorative by definition, we simply include those covariates that seem to be most influential *a priori*: party and state (both categorical), date (as smooth effect), as well as percentage of immigrants, GDP per capita, unemployment rate, and the 2017 election results of the MP's respective party (the last four as smooth effects and on an electoral-district level).

4.2 Labelling

As a starting point subsequent to model fitting, we visually inspect the resulting topics, in particular their best words, as demonstrated in the output below. There are different metrics for evaluating which words are most representative of a topic. The STM comes with four such metrics: *highest probability*, *FREX*, *Lift*, and *Score*. The first one, *highest probability*, simply outputs for each topic those words in the topic-specific word vector, $\beta_{d,n}$, with the highest corpus frequency, i.e, those with the highest absolute frequency across all documents. *FREX* takes into account not only how frequent but also how exclusive words are: for a given topic, it is calculated by taking the harmonic mean of i) the word's rank by probability within the topic (frequency) and ii) the topic's rank by word frequency across all topics (exclusivity). Further information on the estimation of *FREX*, *Lift*, and *Score* can be found in Bischof and Airoldi (2012), in the R package *lda* (J. Chang and Chang (2010)), and in Taddy (2012), respectively.

The output below shows, for each topic, the 5 top words for each of the four topic-word evaluation metrics.

```
## Topic 1 Top Words:
##       Highest Prob: buerg, link, frau, merkel, gruen
##       FREX: altpartei, islam, linksextremist, asylbewerb, linksgru
##       Lift: eitan, 22jaehrig, abdelsamad, abgehalftert, afd'l
```

```
##       Score: altpartei, islam, linksextremist, frauenkongress, boehring
## Topic 2 Top Words:
##       Highest Prob: frag, genau, einfach, find, gern
##       FREX: geles, quatsch, sorry, versteh, satz
##       Lift: duitsland, freiraumkonto, garn, kombo, lieblingss
##       Score: tweet, fuerstenberg, sorry, haushaelt, geles
## Topic 3 Top Words:
##       Highest Prob: brauch, gruen, klimaschutz, wichtig, leid
##       FREX: emissionshandel, erneuerbar, klimaziel, fossil, emission
##       Lift: bahnunternehm, betriebskonzept, bewaesser, biogas, biokraftstoff
##       Score: emissionshandel, co2limit, emission, kraftstoff, erneuerbar
## Topic 4 Top Words:
##       Highest Prob: sozial, miet, berlin, brauch, arbeit
##       FREX: miet, mieterinn, wohnung, wohnungsbau, vermiet
##       Lift: baugrundstueck, baustaatssekreta, behrendt, billigflieg, bodenwertzuwachssteu
##       Score: miet, mietendeckel, mieterinn, wohnung, bezahlbar
## Topic 5 Top Words:
##       Highest Prob: europaeisch, thank, good, great, wichtig
##       FREX: important, foreign, policy, discussion, clos
##       Lift: abroad, acknowledg, across, activity, addressed
##       Score: important, need, great, thank, right
## Topic 6 Top Words:
##       Highest Prob: gruen, frag, euro, geld, minist
##       FREX: scheu, verkehrsminist, autoindustri, nachruest, verkehrsministerium
##       Lift: agrarministerin, angstunternehm, aufklaerungsinteress, autoboss, baulueckenkatast
##       Score: schmunzel, scheu, verkehrsminist, perli, pkwmaut
## Topic 7 Top Words:
##       Highest Prob: wichtig, europa, gemeinsam, brauch, europaeisch
##       FREX: integration, partnerschaft, fried, partn, karamba
##       Lift: bahrenfeld, bamako, entrepreneur, erbfeind, friedensmacht
##       Score: integrationsbeauftragt, europa, antisemitismus, transatlant, conduct
## Topic 8 Top Words:
##       Highest Prob: kris, wichtig, brauch, unternehm, massnahm
##       FREX: coronakris, corona, virus, pandemi, coronavirus
##       Lift: 600milliardenfond, abstandhalt, alltagsmask, antikoerp, antikoerpert
##       Score: corona, coronakris, pandemi, coronavirus, virus
## Topic 9 Top Words:
##       Highest Prob: krieg, link, frag, regier, europaeisch
##       FREX: milita, voelkerrechtswidr, aufruest, geheimdien, libysch
##       Lift: abho, airbas, antimilitarist, aufklaerungsdat, aufruestet
##       Score: voelkerrechtswidr, libysch, milita, voelkerrecht, zdebel
## Topic 10 Top Words:
##       Highest Prob: herzlich, glueckwunsch, dank, freu, stark
##       FREX: achim, parteitag, delegiert, gmuend, glueckwunsch
##       Lift: abschlussfoto, borby, dt.israel, ernstwilhelm, hessennord
##       Score: backnang, gmuend, herzlich, glueckwunsch, achim
## Topic 11 Top Words:
##       Highest Prob: berlin, besuch, gespraech, jung, thema
##       FREX: buongiorno, fdpbundestagsabgeordnet, duesseldorf, weiterles, freihold
##       Lift: aero, aign, alois.karl, andreas.scheu, andreas_mattfeldt
##       Score: buongiorno, fdpbundestagsabgeordnet, storjohann, rimkus, freihold
## Topic 12 Top Words:
##       Highest Prob: frau, gruen, sozial, kind, dank
##       FREX: mention, reach, bielefeld, automatically, retweet
```

```
##       Lift: barrientos, trainingsplaetz, automatically, unfollowed, aktivenkonferenz
##       Score: mention, unfollowed, automatically, reach, checked
## Topic 13 Top Words:
##       Highest Prob: dank, schoen, freu, berlin, abend
##       FREX: leipzig, nachh, heut, hall, wunderscho
##       Lift: bergenenkheim, mainzbing, sommergrill, altlandsberg, anwohnerinn
##       Score: dank, magdeburg, schoen, freu, abend
## Topic 14 Top Words:
##       Highest Prob: partei, demokrat, klar, link, dank
##       FREX: thuering, hoeck, faschist, kemmerich, ramelow
##       Lift: atrium, epost, kernbereich, kommissionschef, maduroregim
##       Score: kemmerich, faschist, hoeck, ramelow, thuering
## Topic 15 Top Words:
##       Highest Prob: kind, pfleg, wichtig, brauch, versorg
##       FREX: neuwied, organsp, pflegebeduerft, patient, widerspruchsloes
##       Lift: altenkirch, gesundheitsberuf, ahrweil, alltagsheldinn, anglizism
##       Score: neuwied, windhag, patient, altenkirch, nnen
```

A key task of topic analysis is to actually ascribe a meaning to the topics identified, i.e., labelling them. While this is clearly where human judgment should and does come into play, we attempt to conduct the labelling in a more stratetic (and thus less subjective) manner, following a 3-step procedure. This procedure is exemplified using topic 1.

First, we consider the *words* contained in the topic, for instance by simply inspecting the top words (see output above). For a better visualization, we use a word cloud. As shown below, for a given topic (i.e., conditional upon a specific topic being chosen), it shows words weighted by their frequency. For instance, by judging at first sight topic 1 appears to be about right-wing nationalist issues, particularly immigration.

Second, to get a more thorough insight into the topic, we take a look into actual *documents*, specifically into those showing the highest proportion for topic 1.

For instance, the most representative document for topic 1, with a proportion of 99.02% is the one by MP Hess, Martin, a member of the AfD party from Baden-Württemberg, from 2018-06 which starts with:

[1] "Ehem. Verfassungsrichter bestätigt AfD-Forderung: Zurückweisung illegaler Migranten dringend geboten. Gegenwärtige Politik widerspricht dem Verstand und auch der Verfassung. Wir müssen zurück zu Recht & Ordnung, wie die #AfD seit fast 3 Jahren fordert!"

The second most representative document, still for topic 1, has a proportion of 98.71%. Its author is the same as for the first document, Hess, Martin, but the date now is 2018-03. The document starts with:

[1] "Offenbar handelt das #BAMF nicht im Interesse der Inneren Sicherheit. Die skandalöse Vorgehensweise

dieser Behörde muss lückenlos aufgearbeitet werden. Es darf nicht sein, dass die Asyllobby über Unterbehörden Einfluss auf staatliche Entscheidungen nimmt!"

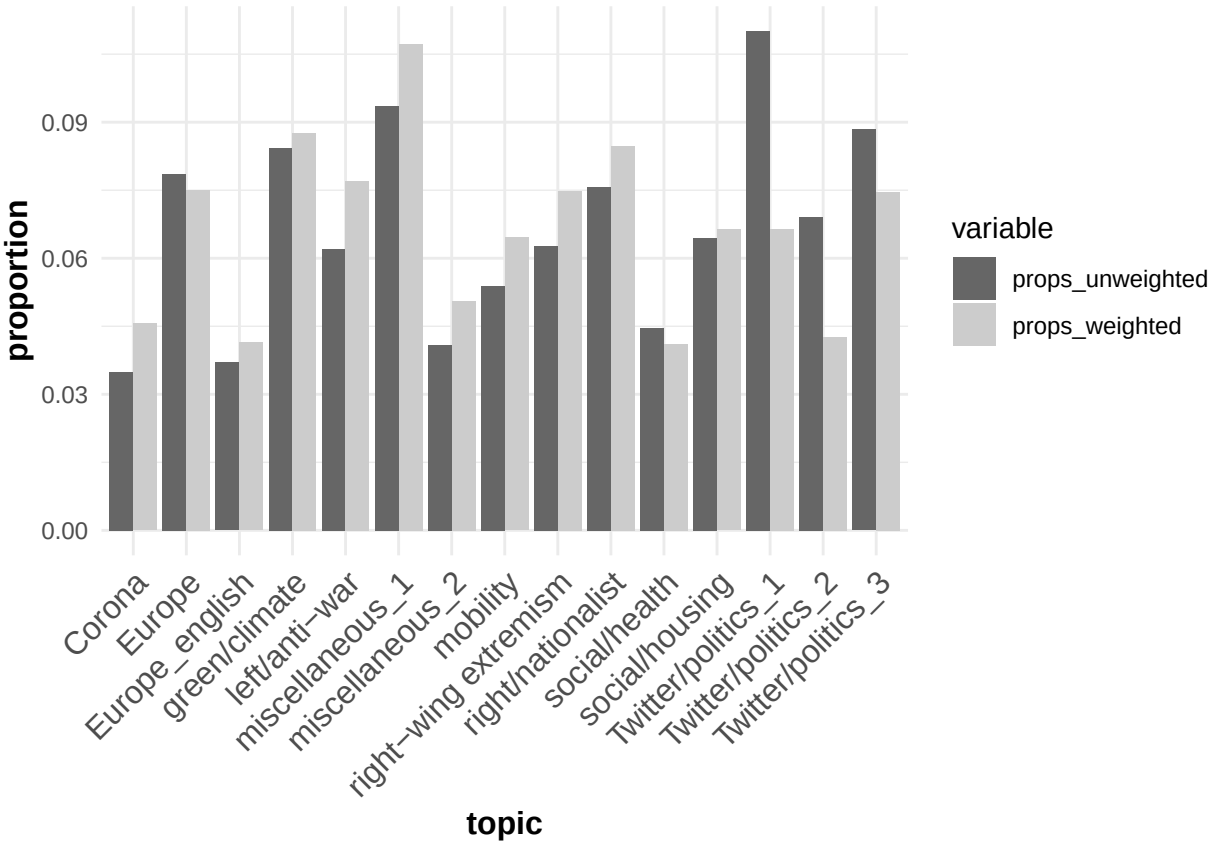
The documents confirm the first impression gained through top words and the word cloud: 1 concerns right-wing nationalist issues, in particular immigration. Thus, as a third step, we finally label the topic: in this case, as right/nationalist.

We repeat this 3-step procedure (inspecting top words and word cloud, reading through top documents, assigning a 1- or 2-word label) for all remaining topics, arriving at the following manual labels:.

Topic1	right/nationalist
Topic2	miscellaneous_1
Topic3	green/climate
Topic4	social/housing
Topic5	Europe_english
Topic6	mobility
Topic7	Europe
Topic8	Corona
Topic9	left/anti-war
Topic10	Twitter/politics_1
Topic11	Twitter/politics_2
Topic12	miscellaneous_2
Topic13	Twitter/politics_3
Topic14	right-wing extremism
Topic15	social/health

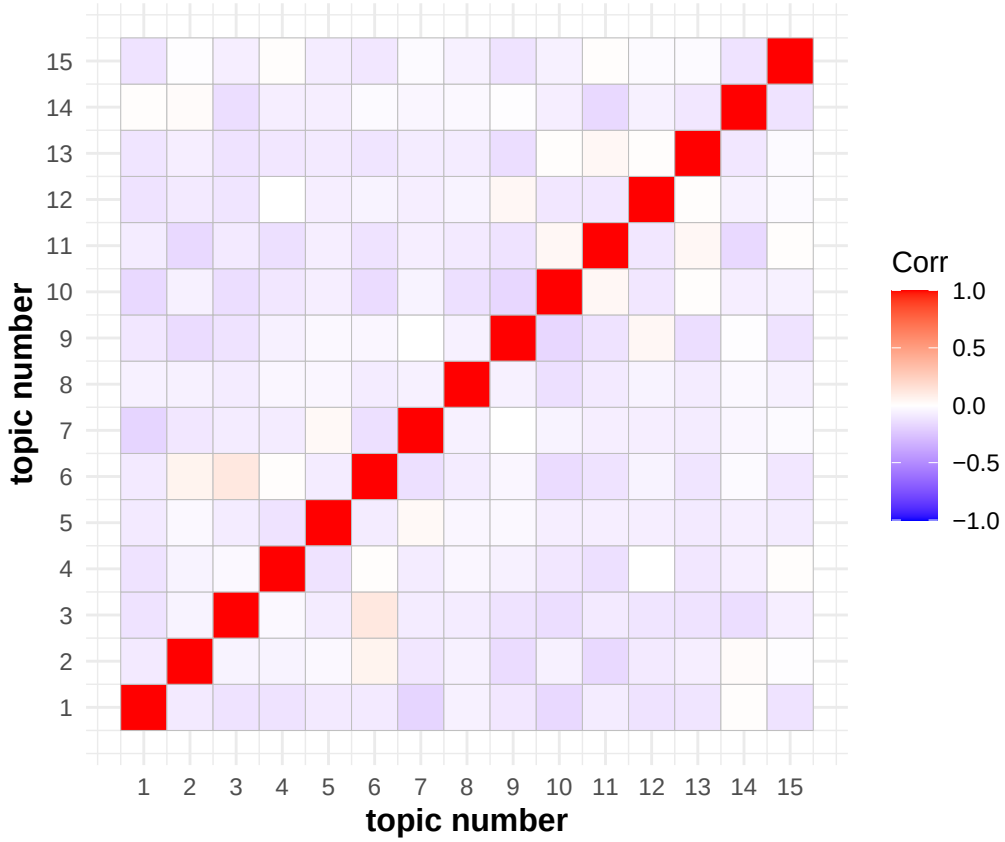
4.3 Global-level Topic Analysis

Next, we identify two ways to calculate global topic proportions: either as the simple (unweighted) average of θ_d across all documents (i.e., as the average of MP-level proportions across all MPs); or by weighting each θ_d by the number of words in the respective documents, N_d. The table below shows all topics with their respective global proportions, for both weighting methodologies. We observe that for most topics, weighted and unweighted proportions are rather similar, but there are exceptions. In particular, the topics concerned with everyday political tweets have much higher unweighted than weighted frequencies; this makes sense, however, since such "diplomatic" tweets tend to be shorter than those which actually discuss a specific content.



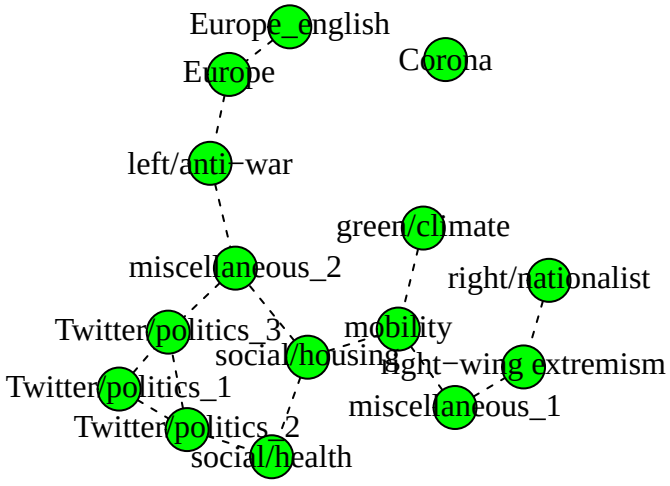
While labelling tells us which words best represent each topic - and thus, what each topic truly represents - it does not yet tell us to which extent individual topics are related to each other. In the graph below, we visualize the similarity of two topics, Topic 3 (green/climate) and Topic 6 (mobility), in terms of their vocabulary usage. As suggested by the topic labels already, there is a significant overlap in vocabulary usage.

More generally, we can evaluate the connectedness between different topics with the topic correlation matrix of the correlations between document-level topic proportions θ_d. This is visualized in the graph below.



Most topics are negatively correlated with each other, which does not come as a surprise, given the relatively low total number of topics, 15, and that topic proportions are "supplements": the higher one topic proportion, the lower the total of the others. Moreover, most topic correlations are rather weak in absolute size: the strongest negative correlation (-17.57%)is the one between topic 1 (right/nationalist) and topic 7 (right/nationalist), while the strongest positive correlation (11.53%) is the one shown before, between (green/climate) and (mobility).

We can also visualize these correlations using a network graph, where topics are connected whenever they are positively correlated. Most topics are only related to two other topics, while none are related to more than three. The only "isolated" topic is topic 8, Corona, which makes sense since it only entered the public sphere in early 2020, i.e., during the last months of our data collection period. In general, the relationships between the topics, as depicted below, are very much in line with their labelling.



4.4 Covariate-level Topic Analysis

After this analysis of topics at a global level, in particular of their labeling and proportions, we now proceed to analyze metadata information (i.e., document-level covariates) and its impact on topic proportions. As mentioned before, the covariates included are party, state (both categorical), date (smooth effect), percentage of immigrants, GDP per capita, unemployment rate, and the 2017 vote share (the last four as smooth effects, on an electoral-district level). Since the target variable $\theta_{(k)}$ is not observable and being estimated itself during the estimation of the STM, we recur to the method of composition to account for the uncertainty contained within $\theta_{(k)}$.

4.4.1 Method of Composition

Let $\theta_{(k)} \in [0,1]^D$ denote the proportions of the k-th topic for all D documents. Suppose that we want to perform a regression of these topic proportions $\theta_{(k)}$ on a subset $\tilde{X} \in \mathbb{R}^{D \times \tilde{P}}$ of prevalence covariates X. The true topic proportions are unknown, but the STM produces an estimate of the approximate posterior of $\theta_{(k)}$, $q(\theta_{(k)}|\Gamma, \Sigma, X)$, where $\Gamma := \Gamma(w, X, Y)$ and $\Sigma := \Sigma(w, X, Y)$. A naïve approach would be to regress the estimated mode of the approximate posterior distribution on $\tilde{X}$. However, this approach neglects much of the information contained in the distribution. Instead, sampling $\theta^*_{(k)}$ from the posterior distribution, performing a regression for each sampled $\theta^*_{(k)}$ on $\tilde{X}$, and then sampling from the estimated distributions of regression coefficients, provides an i.i.d. sample from the marginal posterior distribution of regression coefficients. This procedure is known as the method of composition in the social sciences (Tanner 2012, 52).

Formally, let ξ denote the regression coefficients from a regression of $\theta_{(k)}$ on $\tilde{X}$, and let $q(\xi|\theta_{(k)}, \tilde{X})$ be the approximate posterior distribution of these coefficients, i.e. given design matrix $\tilde{X}$ and response $\theta_{(k)}$.

The R package *stm* implements a simple OLS regression through its *estimateEffect* function. Using this framework we frequently observe predicted proportions outside of $(0, 1)$, given that the restricted domain of $\theta_{(k)}$ is not taken into account. Moreover, credible intervals are non-informative, due to violated model assumptions. Therefore, to adequately model the topic proportions we perform a beta regression (with logit-link), since the sampled proportions are restricted to the interval $(0, 1)$. More information on why beta regression is useful in such a scenario can be found in Ferrari and Cribari-Neto (2004). In case of a beta regression, $q(\xi|\theta_{(k)}, \tilde{X})$ is a normal distribution (see e.g. Ferrari and Cribari-Neto (2004), p. 17).

The method of composition can now be described by repeating the following process m times:

1. Draw $\theta^*_{(k)} \sim q(\theta_{(k)}|\Gamma, \Sigma, X)$.
2. Draw $\xi^* \sim q(\xi|\theta^*_{(k)}, \tilde{X})$.

Then, $\xi^*_1, \ldots, \xi^*_m$ is an i.i.d. sample from the marginal posterior

$$q(\xi|\Gamma, \Sigma, X) := \int_{\theta_{(k)}} q(\xi|\theta_{(k)}, \tilde{X}) q(\theta_{(k)}|\Gamma, \Sigma, X) \mathrm{d}\theta_{(k)} = \int_{\theta_{(k)}} q(\xi, \theta_{(k)}|\Gamma, \Sigma, X) \mathrm{d}\theta_{(k)},$$

where $q(\xi, \theta_{(k)}|\Gamma, \Sigma, X) := q(\xi|\theta_{(k)}, \tilde{X}) q(\theta_{(k)}|\Gamma, \Sigma, X)$. Thus, by integrating over $\theta_{(k)}$, this approach allows incorporating uncertainty about $\theta_{(k)}$ when determining ξ.

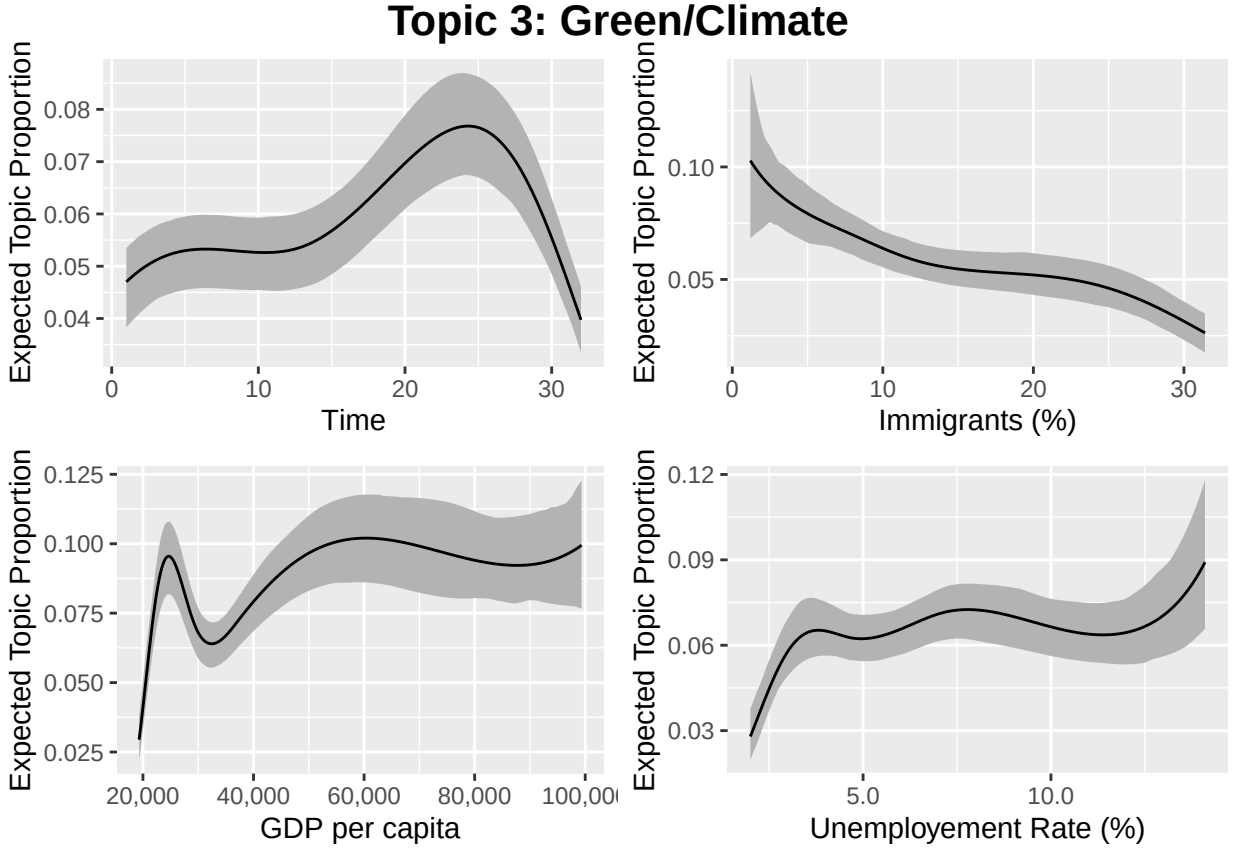
4.4.2 Visualization

We can now apply the method of composition, based on a beta regression, in order to quantify covariate effects. Setting the number of simulations to 100, we thus sample $\xi^*_1, \ldots, \xi^*_{100}$ from the approximate posterior distribution $q(\xi|\Gamma, \Sigma, X)$. In order to plot the predicted effects, we input $\tilde{X}\xi^*$ into the sigmoid function, which is the response function corresponding to a beta regression with logit-link, and calculate the predicted

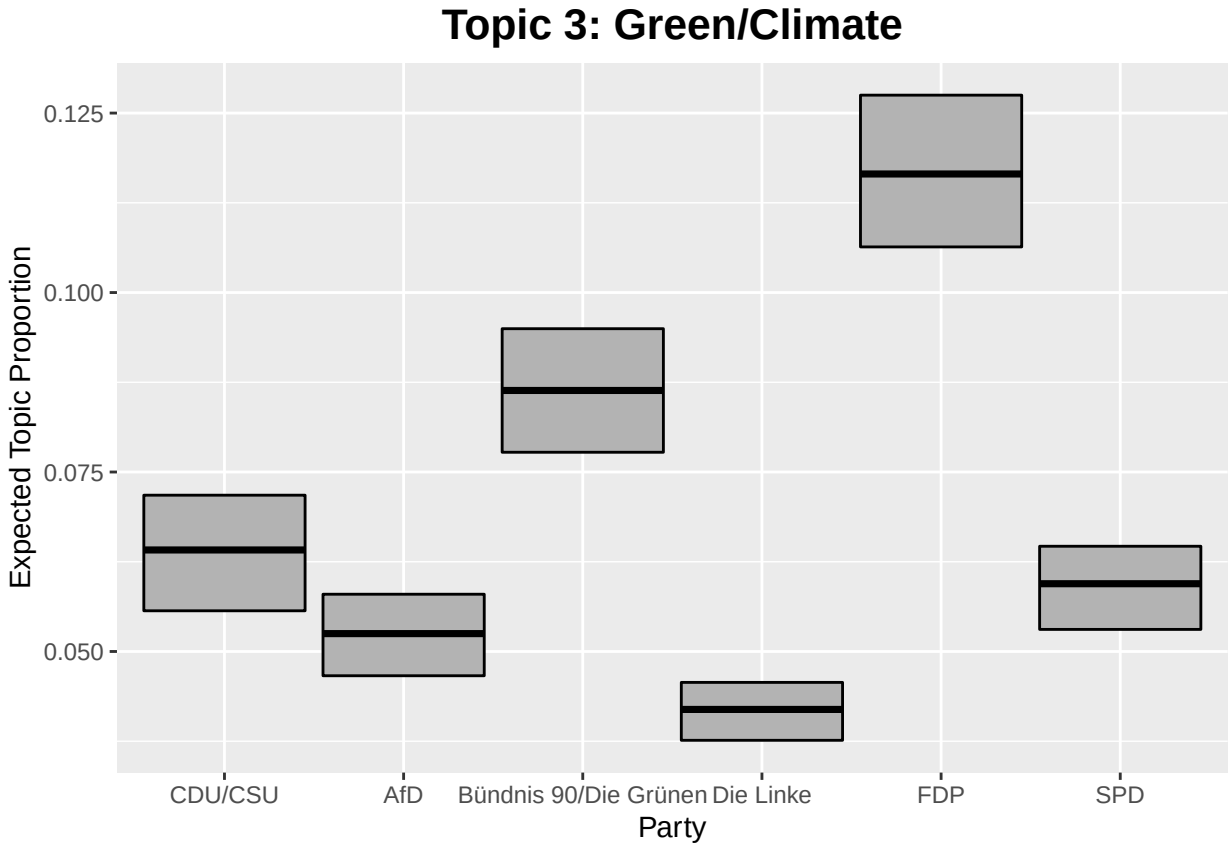
proportions. When visualizing the impact of a particular covariate, all other covariates are held at their median, in line with the methodology employed in the *stm* package.

We discuss the impact of covariates on topic proportions for topics 3 (green/climate) and 4 (social/housing), sub-dividing the analysis into smooth effects (time, immigration, GDP, and unemployment) and categorical variables (party and state). For smooth effects, it is important to recall that their borders are inherently unstable, which is why one should refrain from (over-)interpreting them. For both continuous and categorical variables, black lines indicate the *mean*, the shaded area represents 95% credible intervals.

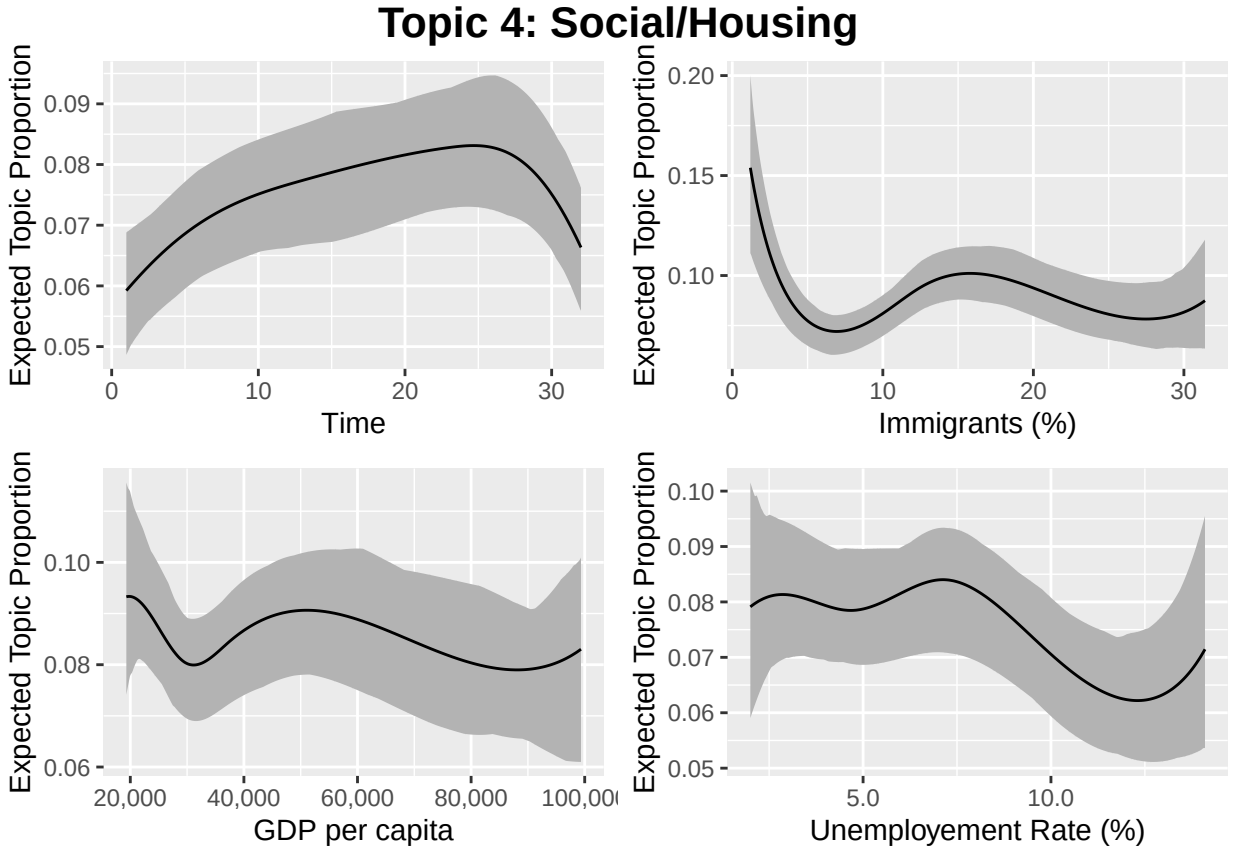
By looking at the smooth effects for topic 3 below, we find that its proportion increases over time until the 25th month, corresponding to September 2019, decreasing sharply afterwards. However this sharp decline is to be taken with caution due to the instability of splines at the borders of the covariate domain. Note that the absolute changes in topic proportions over time for the green/climate topic are rather small (around 4%). The effect of immigrants (as percentage of the total population) is negative across the entire domain, and rather steadily so. The impact of GDP per capita on topic 3 is unclear/constant, while unemployment rate show an overall positive effect.



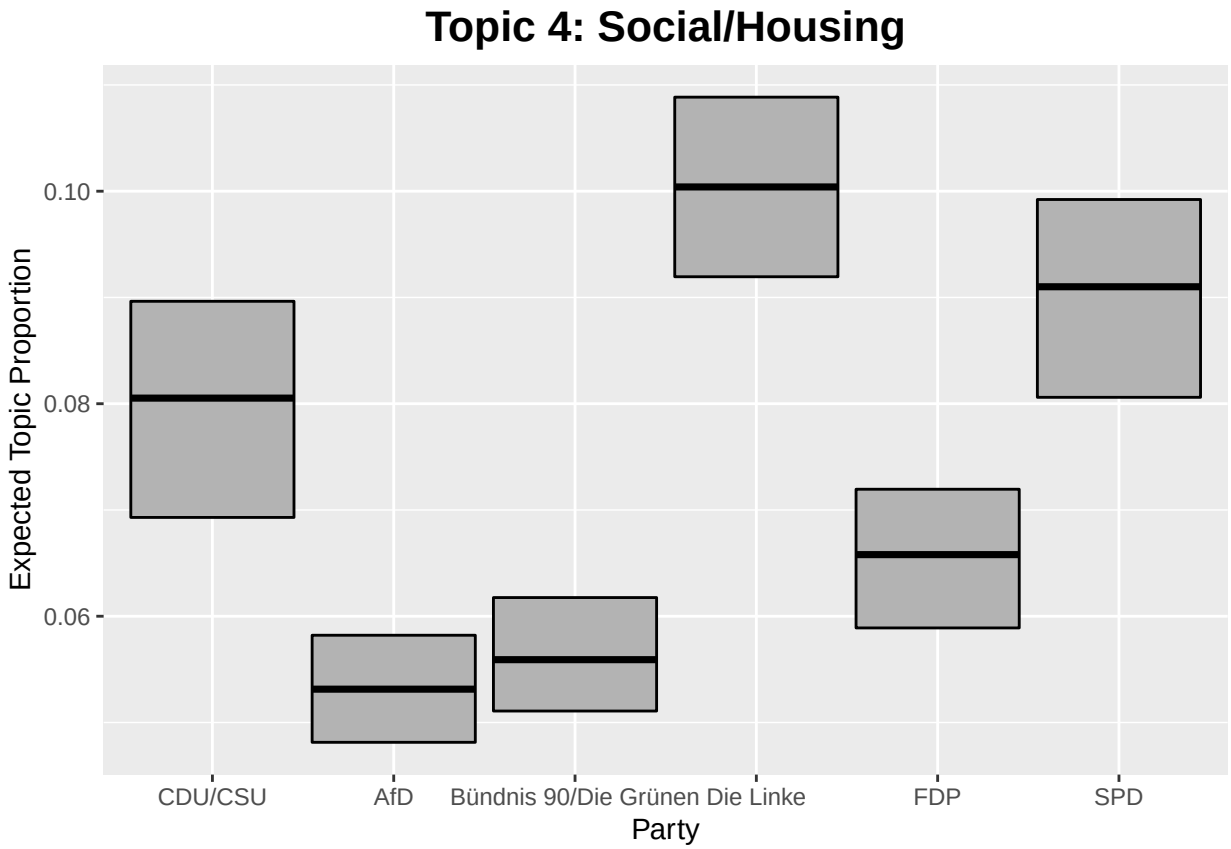
Regarding the effect of categorical variables on topic green/climate, we consider the political party, arguably the most decisive covariate. As was to be expected, we find high topic prevalence for the green party, yet the liberal party is, somewhat surprisingly, the party with the highest prevalence. Similar to the smooth effects, total variation in topic proportions across parties amounts to approximately 8%, as can be seen in the graph below.



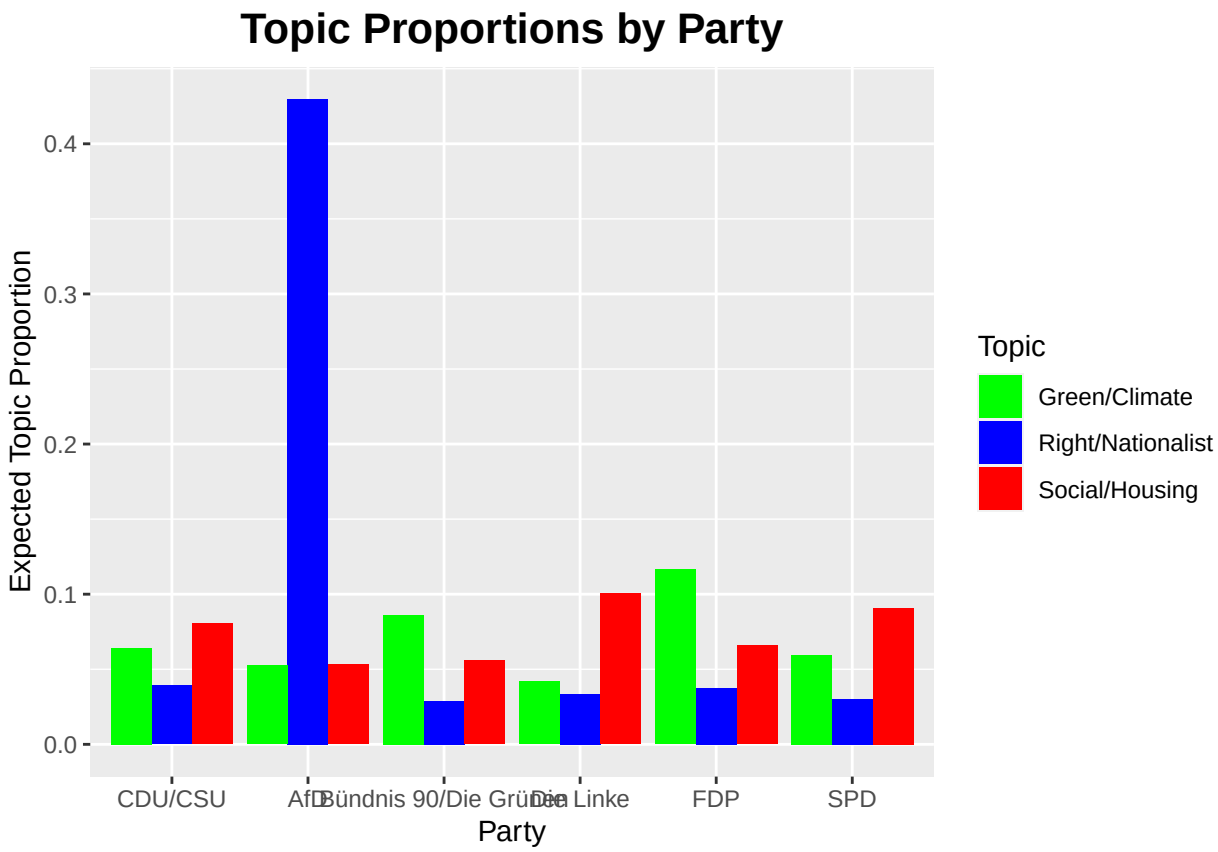
As for topic 4, social/housing, we observe that most (quasi-)continuous variables have a small effect in absolute terms: the absolute variation in topic proportion across the covariate domains merely amounts to 4%, compared to around 8% for the green/climate topic. The time effect is similar to the one for topic 3, particularly the decreasing topic prevalence since September 2019. For the other variables, no clear effect is discernible.



The effect of political party on the relevance assigned to the social/housing topic is very much in line with a priori expectations: the left party and social democrats have the highest topic prevalence, at around 10%, the nationalist party the lowest one at 5%. The overall effect of covariate party is thus similar for topics green/climate and social/housing.



Finally, the graph below shows a summary comparison of topic prevalence across all parties, for topics right/nationalist, green/climate, and social/housing. The results are generally consistent with expectations. The proportions of topics green/climateand social/housing vary between 4% and 12% and between 5% and 10%, respectively. For topic 1, right/nationalist, note how topic prevalence for the AfD party amounts to more than 40%, implying that more than 40% of the total content tweeted by AfD party members is about right-wing/nationalist issues, particulary immigration; for all other parties, topic 1 is rather marginal at 3-4%.



4.5 Train-Test-Split

- causal inference within a topic analysis setting (see Egami et al. (2018))
- "predictive power" of covariates

5 Conclusion

TBD

Bibliography

Bischof, Jonathan, and Edoardo M Airoldi. 2012. "Summarizing Topical Content with Word Frequency and Exclusivity." In *Proceedings of the 29th International Conference on Machine Learning (Icml-12)*, 201–8.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Chang, Jonathan, and Maintainer Jonathan Chang. 2010. "Package 'Lda'." Citeseer.

Egami, Naoki, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. "How to Make Causal Inferences Using Texts." *arXiv Preprint arXiv:1802.02163*.

Ferrari, Silvia, and Francisco Cribari-Neto. 2004. "Beta Regression for Modelling Rates and Proportions."

Journal of Applied Statistics 31 (7). Taylor & Francis: 799–815.

Lucas, Christopher, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23 (2). Cambridge University Press: 254–77.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111 (515). Taylor & Francis: 988–1003.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. "stm: An R Package for Structural Topic Models." *Journal of Statistical Software* 91 (2): 1–40. doi:10.18637/jss.v091.i02.

Taddy, Matt. 2012. "On Estimation and Selection for Topic Models." In *Artificial Intelligence and Statistics*, 1184–93.

Tanner, Martin A. 2012. *Tools for Statistical Inference*. Springer.