Covariate-level Topic Analysis

Patrick Schulze, Simon Wiegrebe

Contents

1 Covariate-level Topic Analysis

After this analysis of topics at a global level, in particular of their labelling and proportions, we now proceed to analyze metadata information (i.e., document-level covariates) and its relation to 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). On possibility to study the relationship between topic proportions and prevalence covariates is to perform a regression of the estimated topic proportions on the latter. However, in contrast to a standard regression setting in this case the dependent variable was estimated itself in a first step. In particular, we have access to the posterior distribution of the topic proportions and can therefore account for the uncertainty of the dependent variable. This can be achieved by employing a sampling procedure known as the method of composition in the social sciences (Tanner 2012, 52).

1.1 Method of Composition

Let $\theta_{(k)} := (\theta_{1,k}, \ldots, \theta_{D,k}) \in [0,1]^D$ denote the proportions of the k-th topic for all D documents. As stated, 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 $\theta_{(k)}$, $q(\theta_{(k)}|X,Y,W)$. 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, repeatedly sampling $\theta^*_{(k)}$ from the approximate posterior distribution, performing a regression for each sampled $\theta^*_{(k)}$ on $\tilde{X}$, and then sampling from the estimated distribution of regression coefficients, provides an i.i.d. sample from the marginal posterior distribution of regression coefficients.

Sampling $\theta^*_{(k)}$ is achieved by first sampling the unnormalized topic proportions η^* from the approximate posterior, $q(\eta)$, applying the softmax $\theta^* = \text{softmax}(\eta*)$ (element-wise, for each of the D elements), and selecting the k-th column of θ^*. Precisely, $q(\eta) = \prod_d q(\eta_d)$ is a normal distribution, which emerges from the laplace variational inference scheme (for details see Roberts, Stewart, and Airoldi (2016), pp. 992-993). For clarity, we denote the approximate posterior as $q(\theta_{(k)}|X,Y,W)$ in order to emphasize that the parameters of this distribution are learned from the observed data, i.e. covariates and words. Furthermore, 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 method of composition can now be described by repeating the following process m times:

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

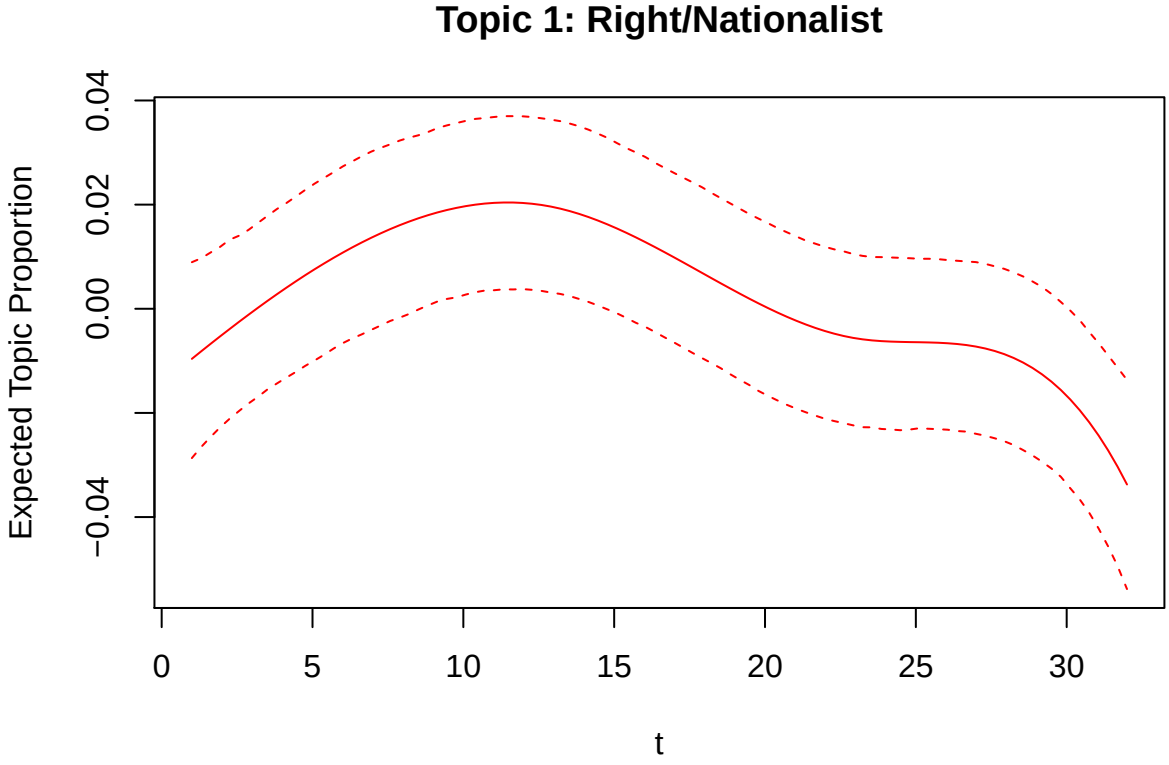
Then, $\xi^*_1, \ldots, \xi^*_m$ is an i.i.d. sample from the marginal posterior

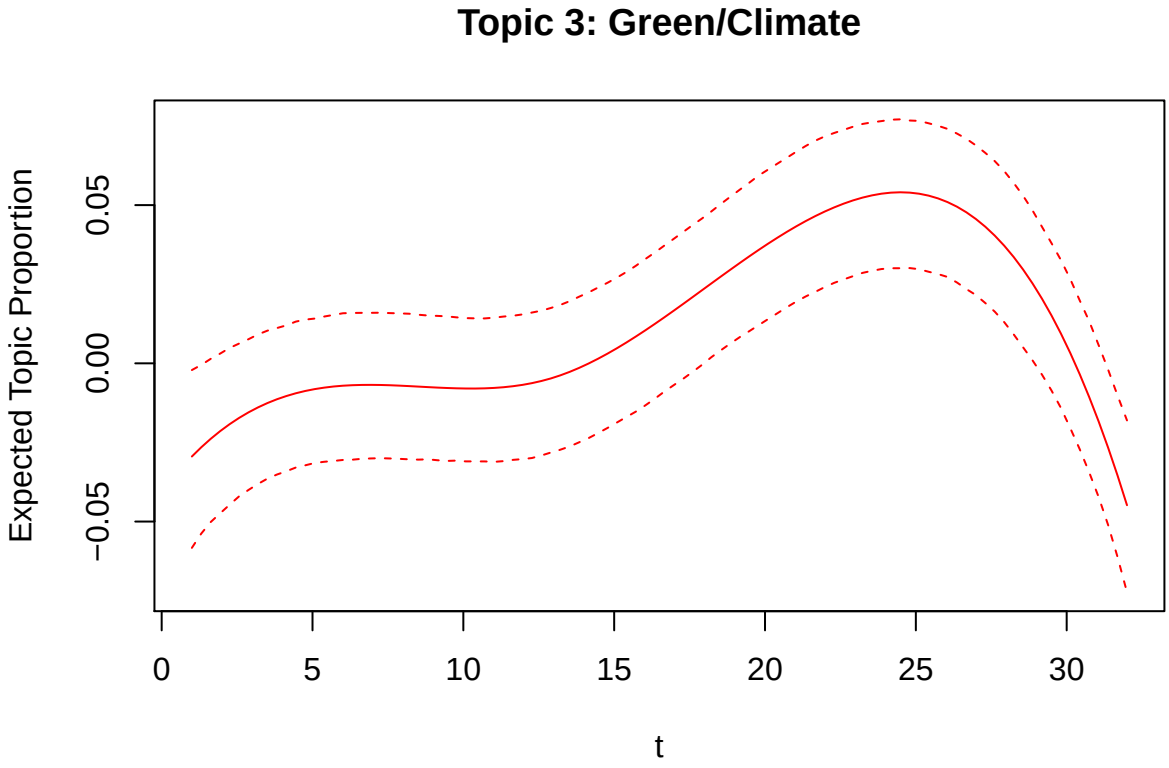
$$q(\xi|\Gamma, \Sigma, X) := \int_{\theta_{(k)}} q(\xi|\theta_{(k)}, \tilde{X}) q(\theta_{(k)}|X, Y, W) \mathrm{d}\theta_{(k)} = \int_{\theta_{(k)}} q(\xi, \theta_{(k)}|X, Y, W) \mathrm{d}\theta_{(k)},$$

where $q(\xi, \theta_{(k)}|X, Y, W) := q(\xi|\theta_{(k)}, \tilde{X}) q(\theta_{(k)}|X, Y, W)$. Thus, by integrating over $\theta_{(k)}$, this approach allows incorporating information contained in the posterior distribution of $\theta_{(k)}$ when determining ξ.

1.1.1 Implementation in the *stm* package

The R package *stm* implements a simple OLS regression through its *estimateEffect* function. However, this approach ignores that the sampled topic proportions are restricted to $(0, 1)$. As expected, using this framework we frequently observe predicted proportions outside of $(0, 1)$. Moreover, credible intervals are non-informative, due to violated model assumptions.





1.1.2 Alternative implementation

We can attempt to improve the approach employed within the *stm* package by replacing the OLS regression with a regression model that assumes a dependent variable in $(0, 1)$. However, note that since topic proportions are modeled separately, regardless of the specific model implied, distributional assumptions about $\theta_{(k)}$ will be violated. This is due to the fact that the the distribution of a subvector - and thus particularly of a single element - of θ_d is not of a simple form, when θ_d follows a logistic normal distribution (Atchison and Shen 1980).

A distribution that can be used to approximate a logistic normal distribution is the dirichlet distribution (Atchison and Shen 1980). In case of the dirichlet distribution the univariate marginal distributions are beta. One possibility is thus to perform a separate beta regression for each topic proportion on $\tilde{X}$.

An alternative is to employ a quasibinomial generalized linear model (GLM). The topic proportions can be rescaled and discretized, such that each rescaled topic proportion can be interpreted as the "number of successes" for the respective topic. To match the underlying logistic normal distribution more closely, we furthermore allow for a flexible variance specification using a quasibinomial GLM.

Note that $q(\xi|\theta_{(k)}, \tilde{X})$ is a normal distribution for both the beta regression (Ferrari and Cribari-Neto 2004, 17) and the quasibinomial GLM. Furthermore, in both cases we use a logit-link.

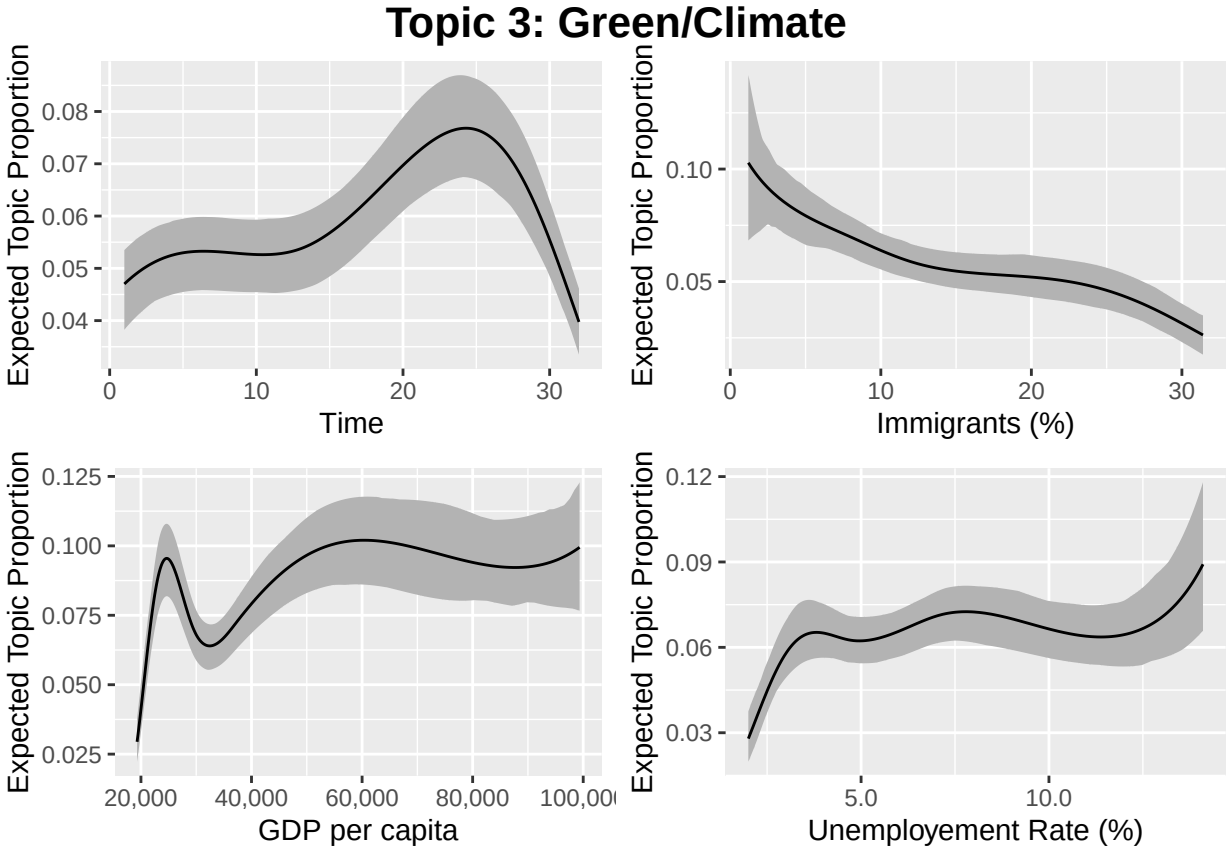
1.1.3 Visualization

We now apply the method of composition, based on either a beta regression or a quasibinomial GLM, in order to visualize covariate effects. Setting the number of simulations to 100, we 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

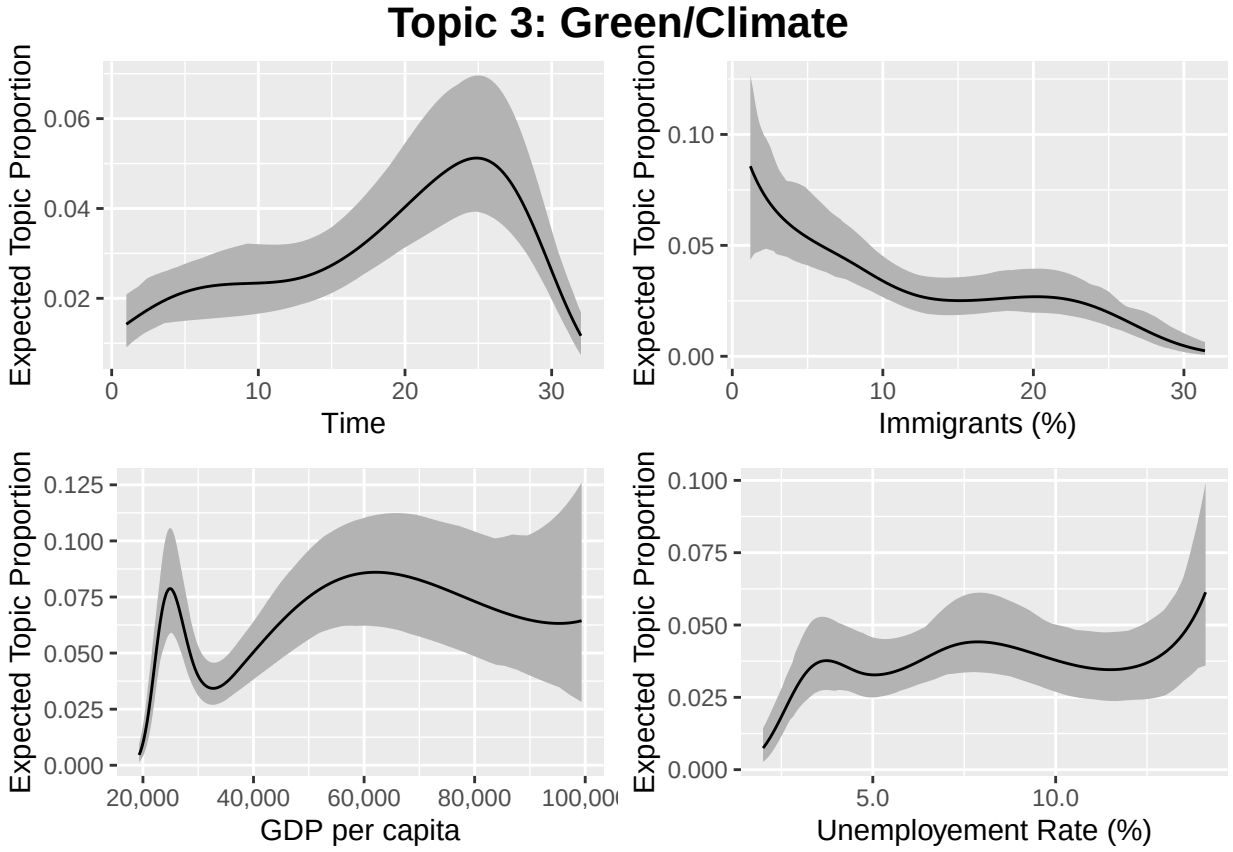
are held at their median (or, if categorial, the majority vote), in line with the methodology employed in the *stm* package.

We illustrate the relationship of covariates and 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, and the shaded area represents 95% credible intervals.

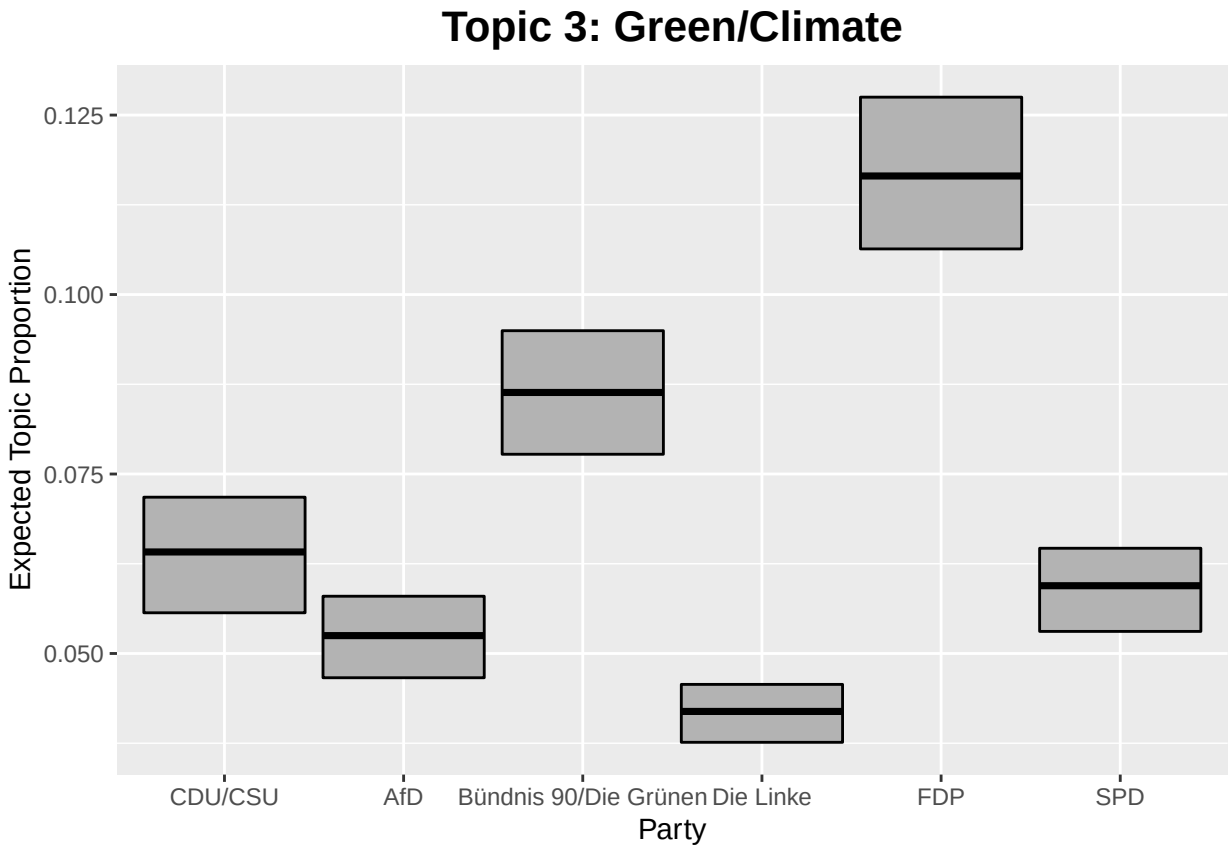
For the smooth effects of topic 3, we find that its proportion increases over time until September 2019, decreasing afterwards. Note that the absolute changes in topic proportions over time for the green/climate topic are rather small (around 4%). The effect of immigrants (as percentage of the total population) is negative across the entire domain, and rather steadily so. The impact of GDP per capita on topic 3 is rather ambiguous, although generally topic 3 is discussed more frequently if GDP is high. Finally, unemployment rate shows an overall positive effect on topic 3.



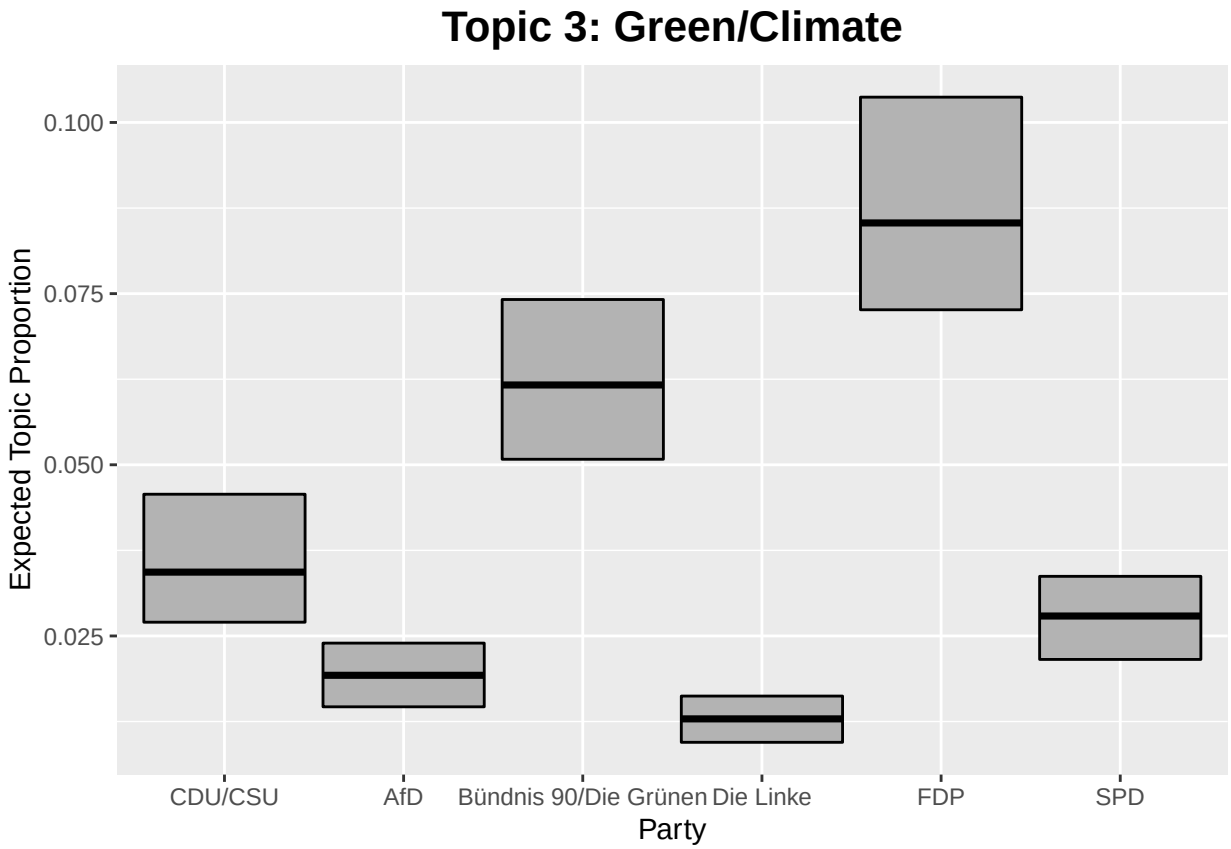
We can compare these results with the predictions using a quasibinomial GLM instead of a beta regression:



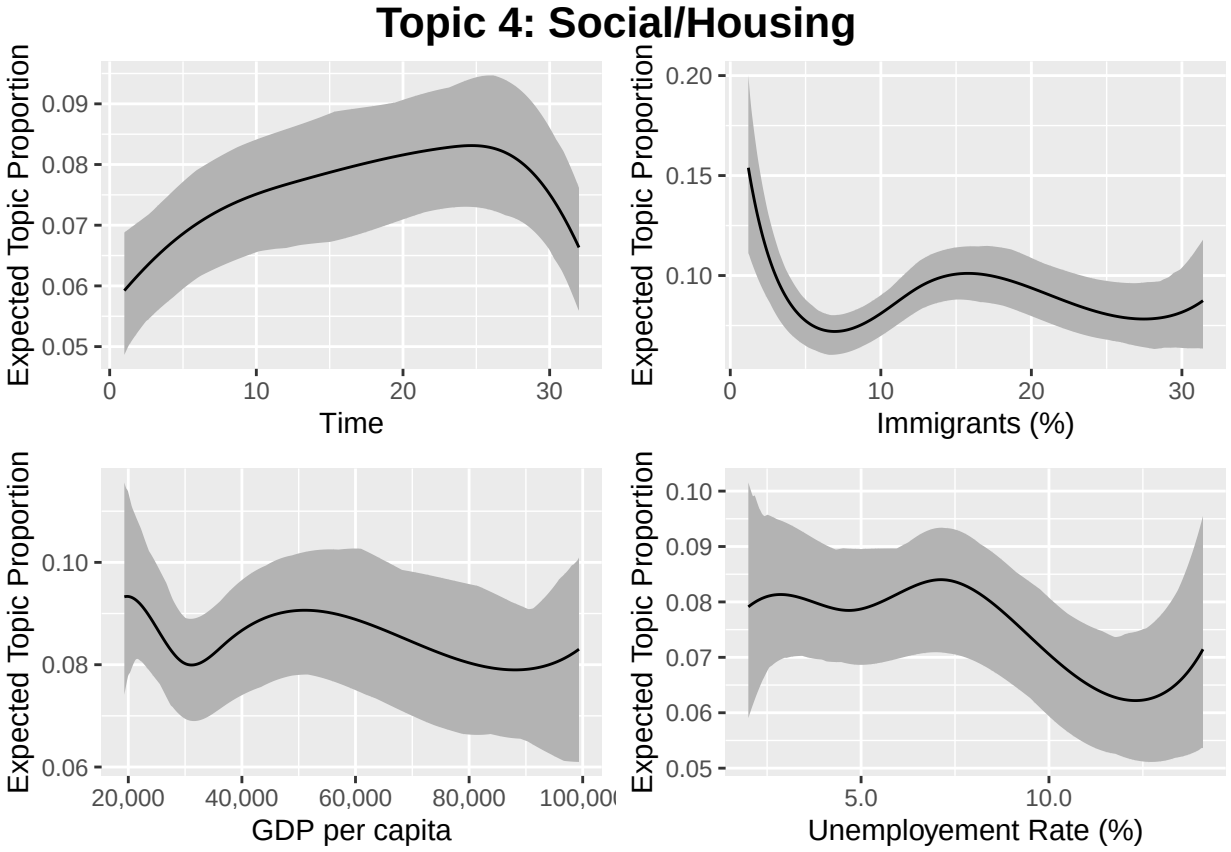
Regarding the effect of categorical variables on topic green/climate, we consider the political party, arguably the most decisive covariate. As was to be expected, we find high topic prevalence for the green party, yet the liberal party is, somewhat surprisingly, the party with the highest prevalence.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

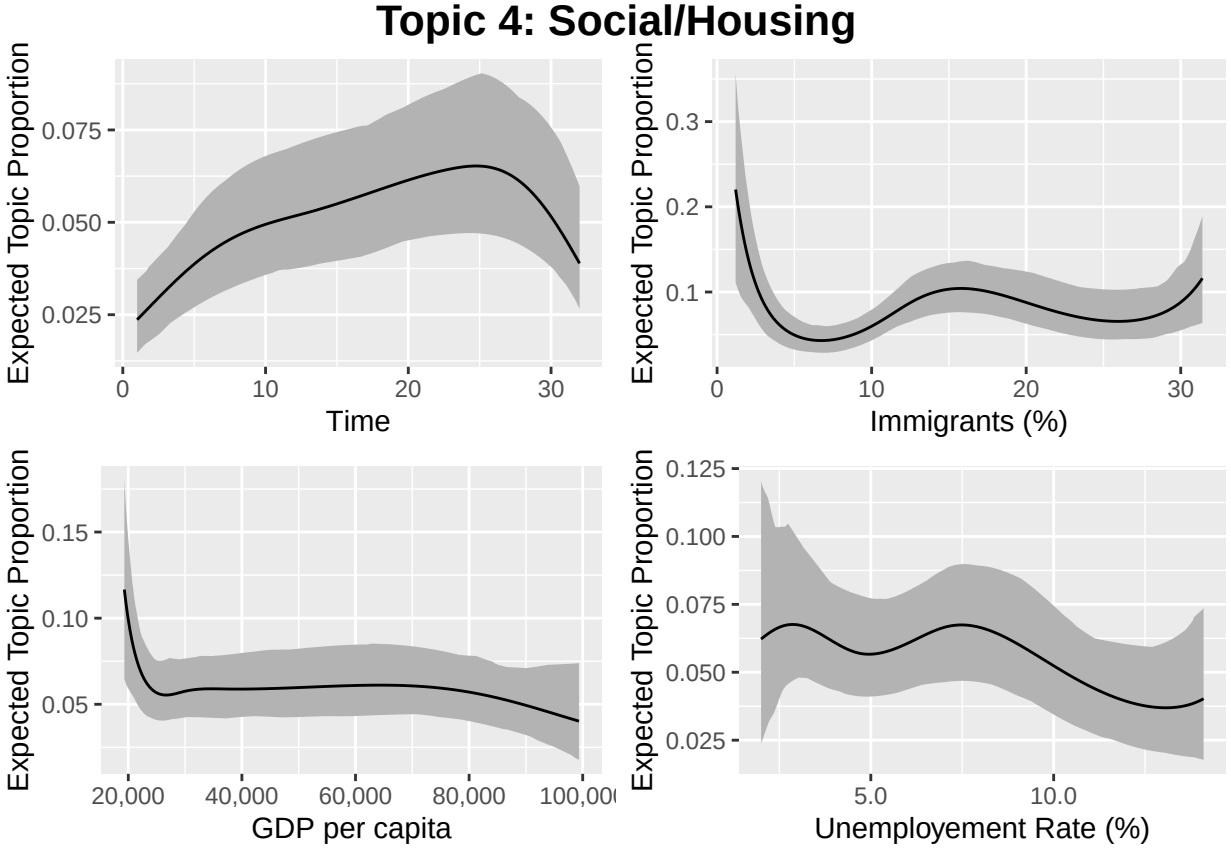


Again, we can compare this to the quasbinomial GLM:

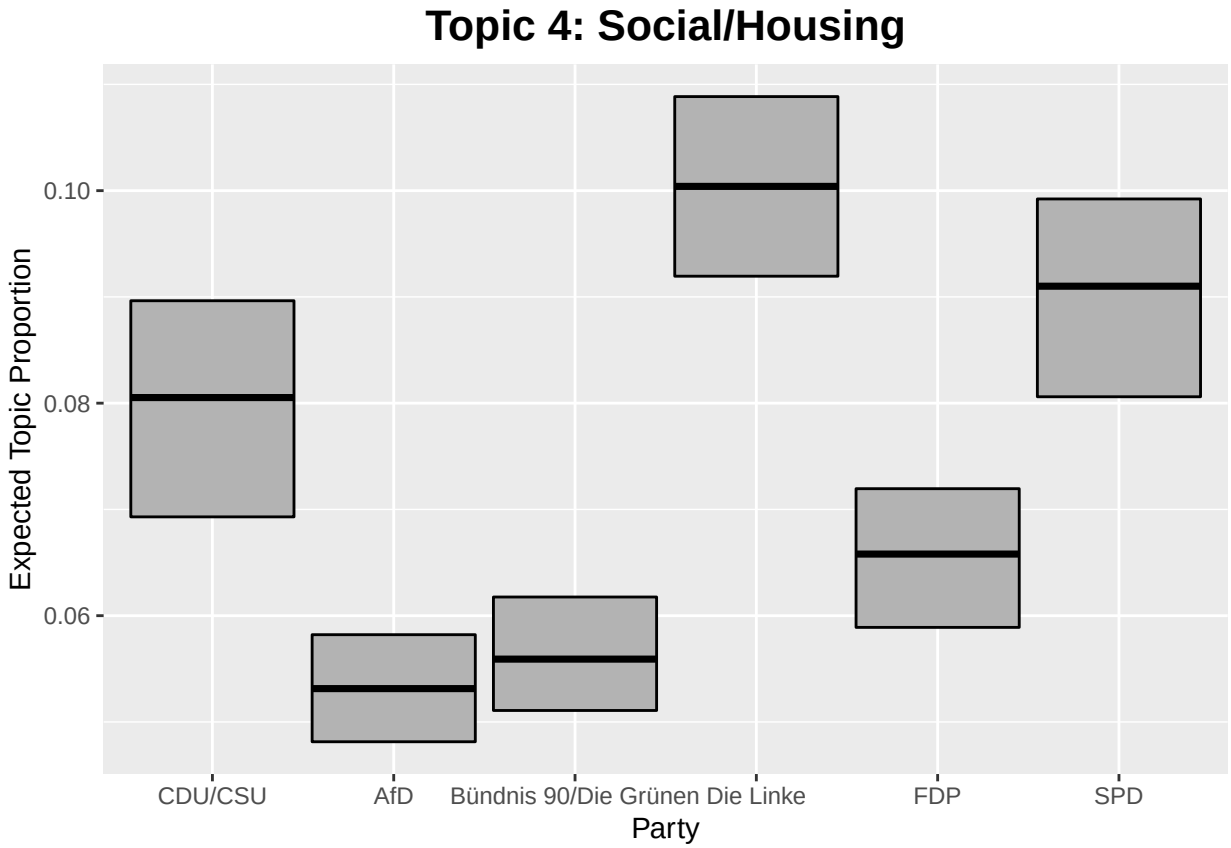


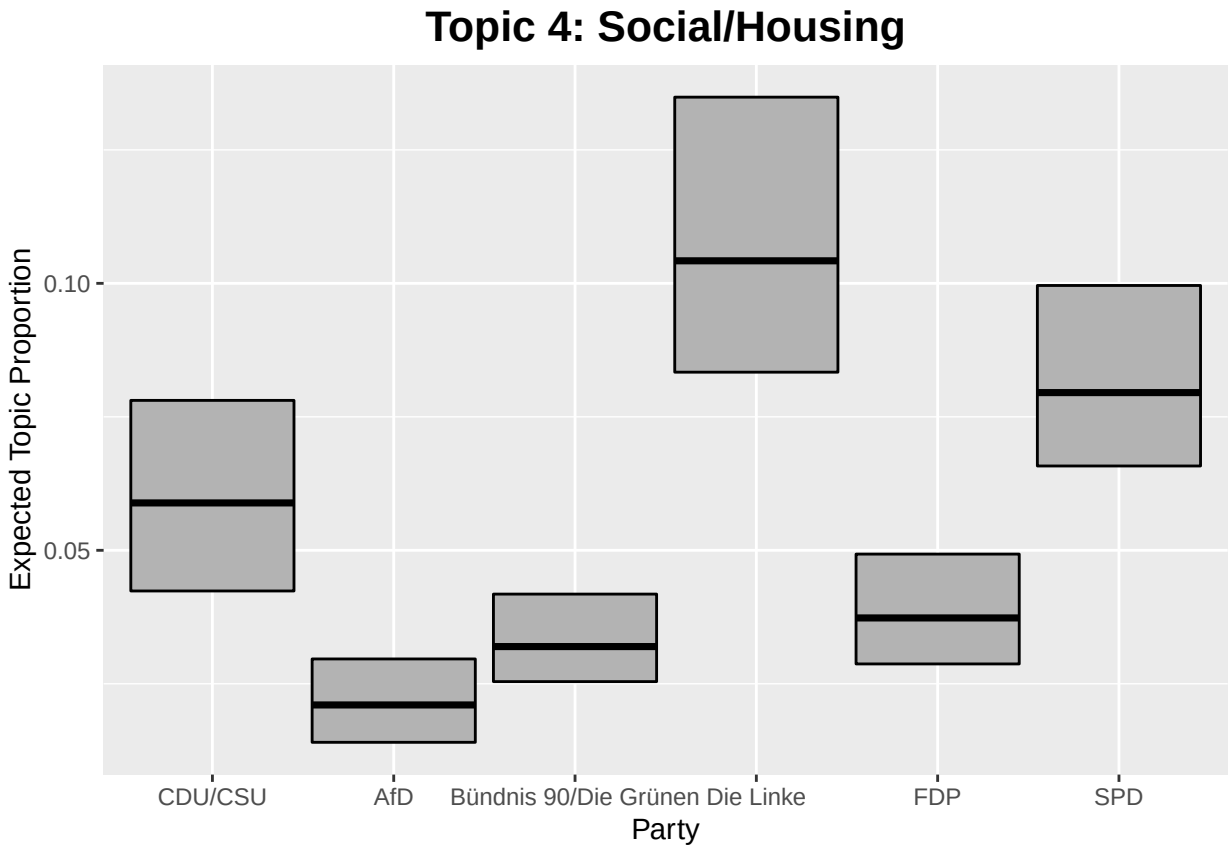
As for topic 4, social/housing, we observe that most (quasi-)continuous variables have a small effect in absolute terms: the absolute variation in topic proportion across the covariate domains merely amounts to 4%, compared to around 8% for the green/climate topic. The time effect is similar to the one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


Topic 4: Social/Housing
Expected Topic Proportion
0.09
0.08
0.07
0.06
0.05
0
10
20
30
Time
Expected Topic Proportion
0.20
0.15
0.10
0
10
20
30
Immigrants (%)
Expected Topic Proportion
0.10
0.08
0.06
20,000
40,000
60,000
80,000
100,000
GDP per capita
Expected Topic Proportion
0.10
0.09
0.08
0.07
0.06
0.05
5.0
10.0
Unemployement Rate (%)

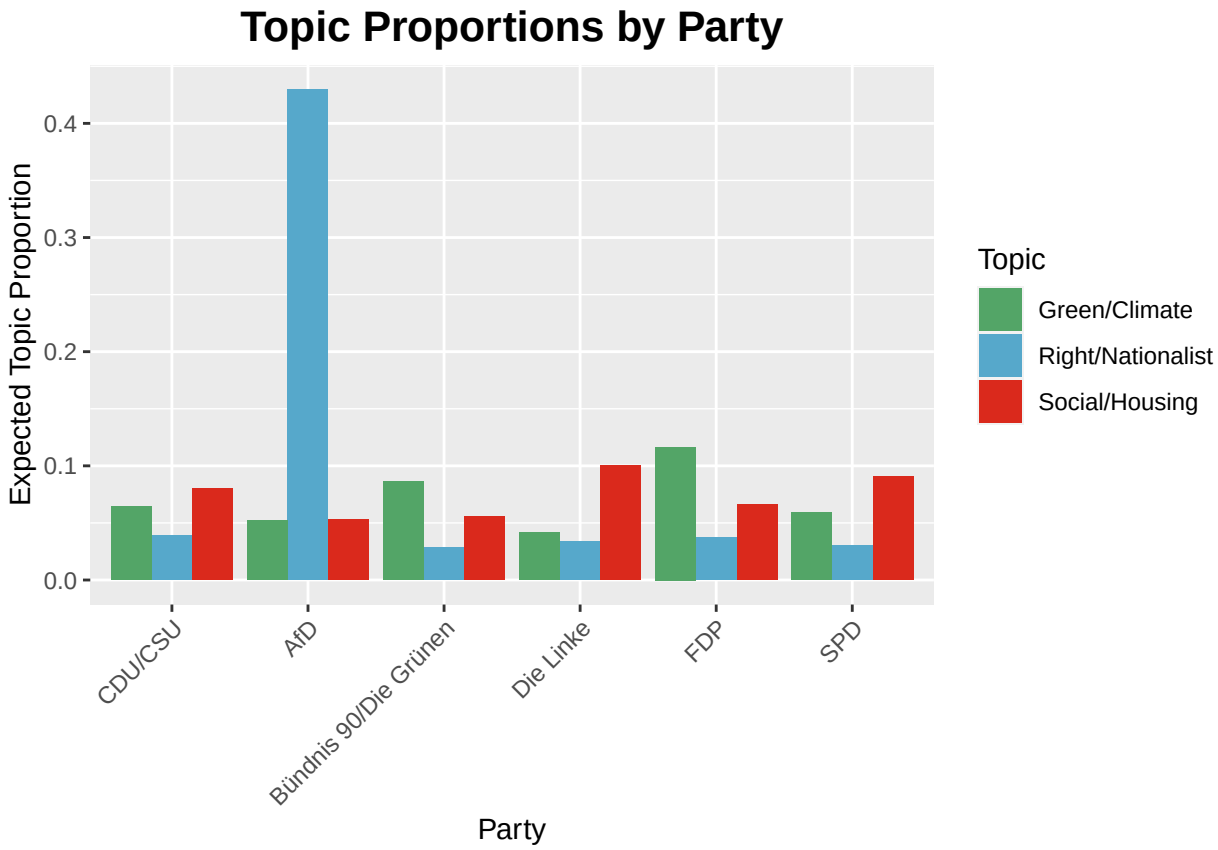

Topic 4: Social/Housing
Expected Topic Proportion
0.075
0.050
0.025
0
10
20
30
Time
Expected Topic Proportion
0.3
0.2
0.1
0
10
20
30
Immigrants (%)
Expected Topic Proportion
0.15
0.10
0.05
20,000
40,000
60,000
80,000
100,000
GDP per capita
Expected Topic Proportion
0.125
0.100
0.075
0.050
0.025
5.0
10.0
Unemployement Rate (%)

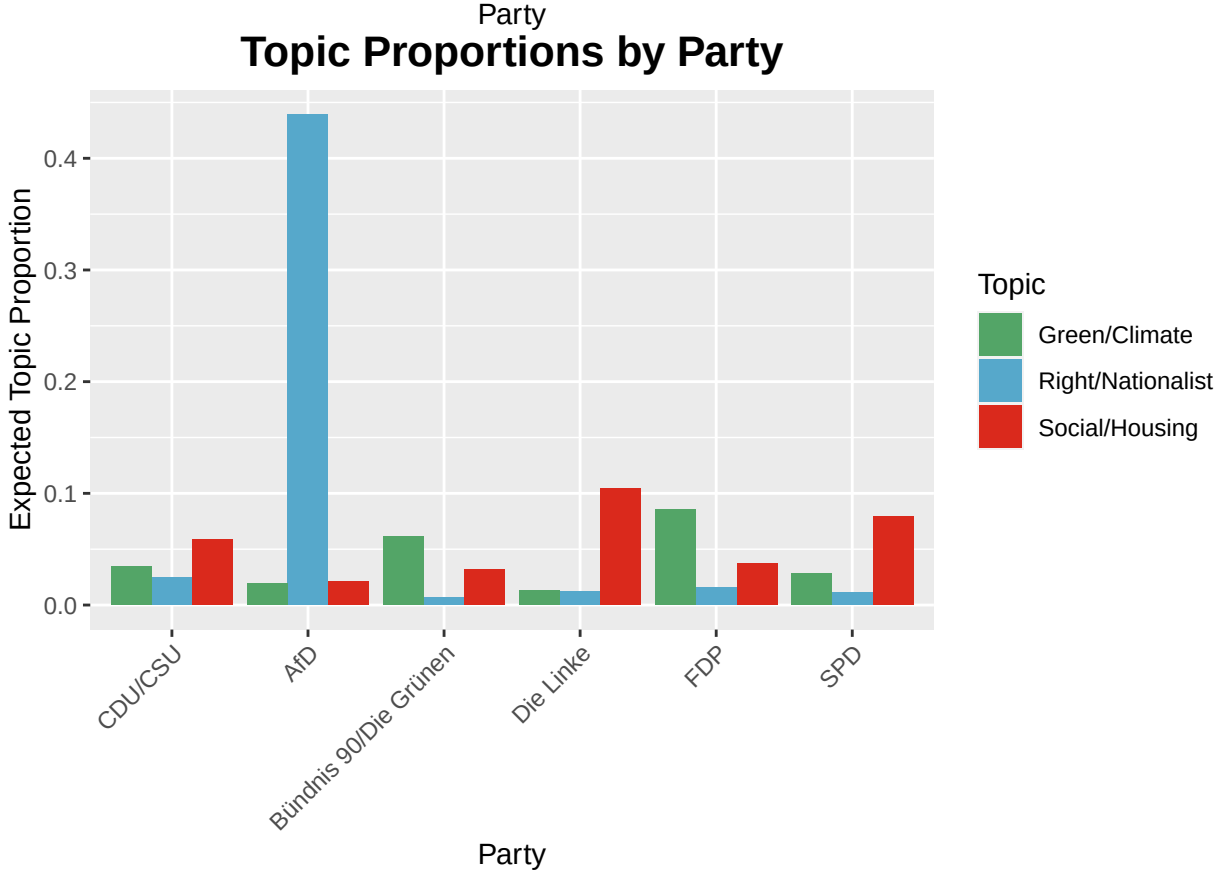
The effect of political party on the relevance assigned to the social/housing topic is very much in line with a priori expectations: the left party and social democrats have the highest topic prevalence, at around 10%, the nationalist party the lowest one at 5%. The overall effect of covariate party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


Topic Proportions by Party
Expected Topic Proportion
0.0
0.1
0.2
0.3
0.4
CDU/CSU
AfD
Bündnis 90/Die Grünen
Die Linke
FDP
SPD
Party
Topic
Green/Climate
Right/Nationalist
Social/Housing


Topic Proportions by Party
Expected Topic Proportion
0.0
0.1
0.2
0.3
0.4
CDU/CSU
AfD
Bündnis 90/Die Grünen
Die Linke
FDP
SPD
Party
Topic
Green/Climate
Right/Nationalist
Social/Housing

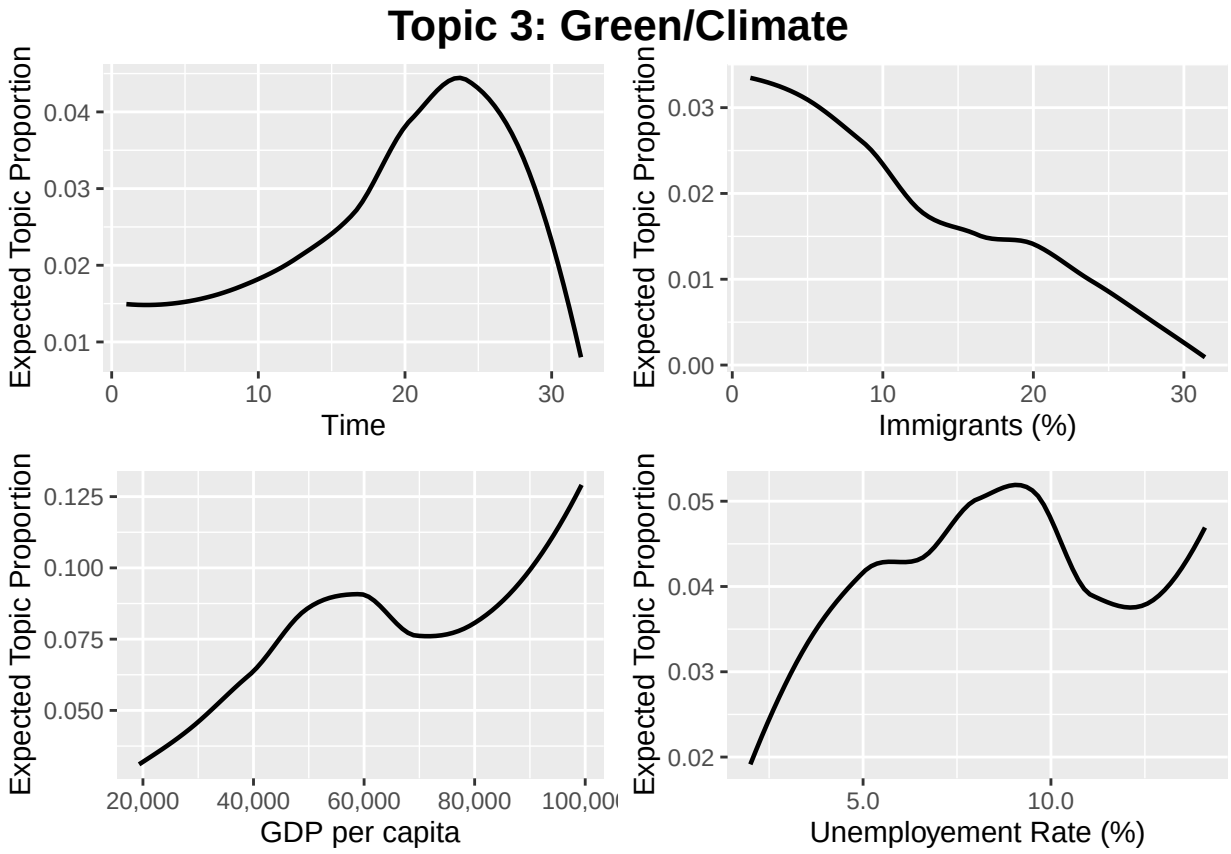
1.2 Direct assessment using $\hat{\Gamma}$ and $\hat{\Sigma}$

The *stm* being an extension to the correlated topic model (CTM), it is assumed that the topic proportions follow a logistic normal distribution, such that $\theta_d \sim \text{LogisticNormal}_{K-1}(\Gamma^T x_d^T, \Sigma)$. Within the CTM the dirichlet distribution is replaced by a logistic normal distribution in order to allow for a joint dependence among topic proportions (Blei, Lafferty, and others 2007). Therefore, as mentioned above, separately modeling topic proportions is a simplification; in particular credible intervals should be treated with caution.

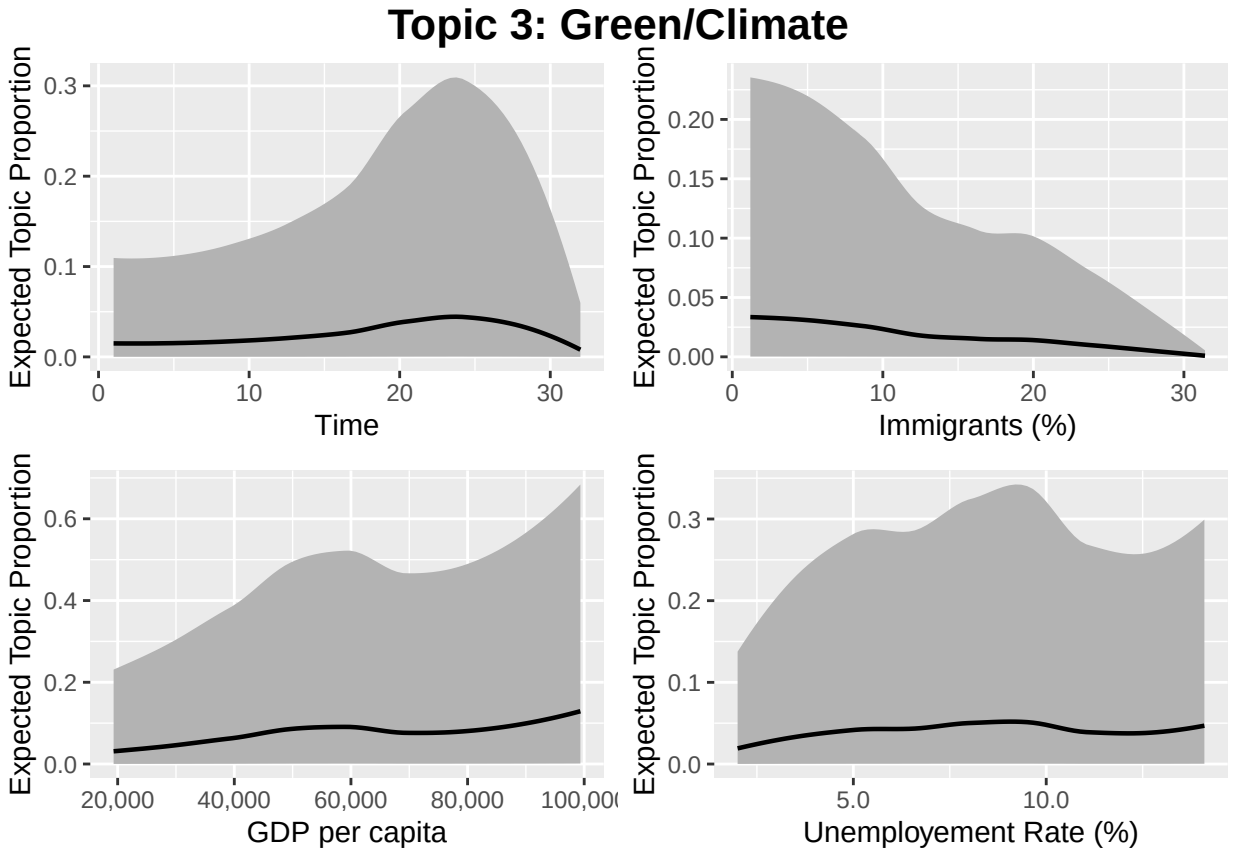
In order to examine the relation of prevalence covariates and topic proportions considering the joint dependence among the latter, we can directly use the output produced by the *stm*: Inference of the *stm* involves finding the MAP estimates $\hat{\Gamma}$ and $\hat{\Sigma}$. For a given x_d^* we can sample θ_d^* from $\text{LogisticNormal}_{K-1}(\hat{\Gamma}^T (x_d^*)^T, \hat{\Sigma})$ by performing the following steps:

1. Draw $\eta_d^* \sim \mathcal{N}_{K-1}(\hat{\Gamma}^T (x_d^*)^T, \hat{\Sigma})$.
2. For all $k = 1, \ldots, K$: $\theta_{d,k}^* = \exp(\eta_{d,k}^*) / \exp(\sum_{i=1}^{K} \eta_{d,i}^*)$.
3. $\theta_d^* = (\theta_{d,1}^*, \ldots \theta_{d,K}^*)^T$.

Here η_d^* denote the unnormalized topic proportions and $\eta_{d,K}^*$ is fixed to zero.

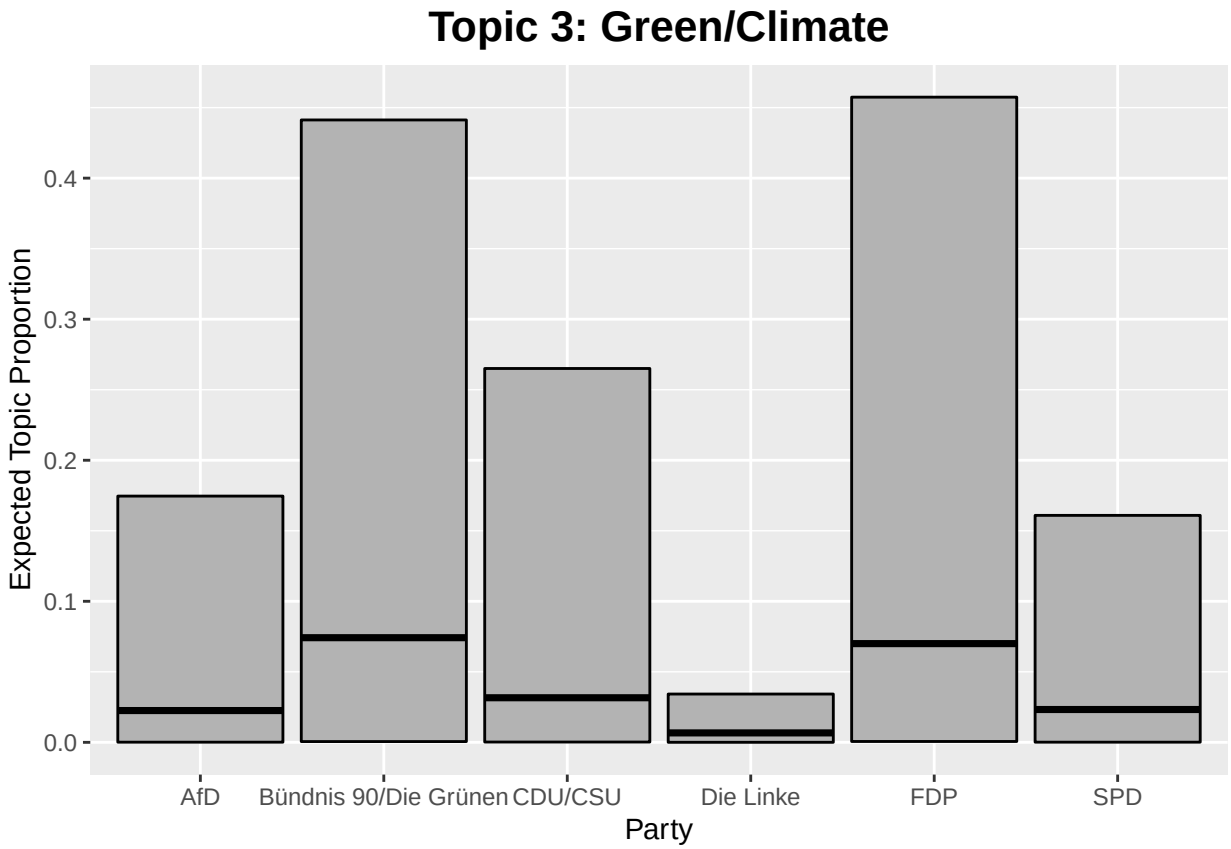


Plotting the credible intervals we observe that the spectrum of expected topic proportions is very broad:



The large fluctuations for a specific topic proportion can be ascribed to the fact that the unnormalized topic proportions are drawn from a $K-1$-dimensional *multivariate* normal distribution, before the softmax is applied. Therefore, a single normalized proportion depends heavily on the sampled unnormalized proportions of the remaining topics. While the variance of a topic-specific unnormalized proportion is independent of the remaining unnormalized proportions and c.p. constant for an increasing number of topics, the application of the softmax function induces a large increase in the variance of a topic-specific normalized proportion.

We suspect that the credible intervals in figure ... provide a more realistic picture than those obtained in case of a separate modeling of topic proportions, since the logistic normal distribution of topic proportions is an assumption made within the *stm*, in order to incorporate a covariance structure among the topics, as argued above. This ultimately produces a large variance of the univariate marginal distributions of topic proportions, as observed.



Atchison, J, and Sheng M Shen. 1980. "Logistic-Normal Distributions: Some Properties and Uses." *Biometrika* 67 (2). Oxford University Press: 261–72.

Blei, David M, John D Lafferty, and others. 2007. "A Correlated Topic Model of Science." *The Annals of Applied Statistics* 1 (1). Institute of Mathematical Statistics: 17–35.

Ferrari, Silvia, and Francisco Cribari-Neto. 2004. "Beta Regression for Modelling Rates and Proportions." *Journal of Applied Statistics* 31 (7). Taylor & Francis: 799–815.

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.

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