Ludwig-Maximilians-Universität München

Statistical Consulting Project

E-Complaints of Citizens and Responsiveness Strategies of Political Candidates and Parties: Investigating Large-Scale Unstructured Text

Comparison of homepages of national and local political entities

Simon Wiegrebe, Patrick Schulze

supervised by

Prof. Dr. Christian Heumann and Prof. Dr. Paul W. Thurner

June 22, 2020

Contents

1 Covariate-level Topic Analysis

We now proceed to analyze the relationship between metadata information (i.e., document-level covariates) and topic proportions. We specify topical prevalence as

$$\mu_{d,k} = x_d^T \gamma_k = \text{party}_{d,k} + \text{state}_{d,k} + f_k(\text{t}_d) + g_k(\text{struct}_d), \tag{1.1}$$

for all documents $d = 1, \ldots, D$, and all topics $k = 1, \ldots, K$, where

$$g_k(\text{struct}_d) = g_k^{(1)}(\text{GDP}_d) + g_k^{(2)}(\text{unemployment}_d) + g_k^{(3)}(\text{immigrants}_d) + g_k^{(4)}(\text{votes}_d).$$

That is, the political party and federal state of the respective parliamentarian associated with a document are specified as simple categorial dummy effects, while date and electoral-district structural covariates (GDP per capita, unemployment rate, percentage of immigrants, and the 2017 vote share) are modeled as additive smooth functions.

Note that approximate inference implies replacing $\mu_{d,k}$ with $\lambda_{d,k}$, i.e. with the mean of the approximate gaussian posterior $q(\eta_{d,k})$. The estimates of $\Gamma = [\gamma_1, \ldots, \gamma_K]$ are updated in a bayesian linear regression during each iteration of the EM algorithm in the M-step; for details see Roberts et al. (2013), p. 993.

While topical prevalence has an effect on the estimated topic proportions, the exact specification of topical prevalence is not a decisive factor. Both estimated topic proportions as well as heldout likelihood are in general only marginally affected by the concrete choice of the functional form. However, completely removing topical prevalence, in which case the model reduces to a CTM, does result in different topic proportions, as we show in section XXX. Since evaluation metrics such as heldout likelihood are mostly unaffected by the exact choice of topical prevalence and because the computational cost of fitting an stm is rather high, automatic model selection methods w.r.t. topical prevalence are not available. A reasonable specification of topical prevalence therefore relies on the domain knowledge of the researcher.

There exist different approaches to study the relationship between topic proportions and prevalence covariates. One possibility is to directly assess the MAP estimates $\hat{\Gamma}$ and $\hat{\Sigma}$ generated by the stm. Since the document-level topic proportions θ_d follow a logistic normal distribution (with mean μ_d and covariance matrix Σ), interpretation of the results can be difficult, since the logistic normal distribution is not very accessible. Nonetheless, we can visualize the relationship between a topic and a prevalence covariate, fixing other covariates at their median (or majority vote, for categorial variables).

Alternatively, the estimated topic proportions can be used as the dependent variable of a new regression on prevalence covariates. However, in contrast to a standard regression setting, in this case the dependent variable has been estimated itself, before the regression is performed. Instead of simply using the MAP estimates of θ_d as the dependent variable,

having access to the posterior distribution of the topic proportions, we can take 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; see Tanner (2012), p.52. This procedure is implemented in the *stm* package through its function *estimateEffect*.

In the following, we will first introduce the method of composition. We will discuss its implementation in the *stm* package and provide alternative regression approaches based on the method of composition. Subsequently, we will evaluate the relationship between prevalence covariates and topic proportions by directly assessing the MAP estimates $\hat{\Gamma}$ and $\hat{\Sigma}$, as outlined above, and compare the results of both approaches.

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 naÃ¯ve 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 icients, 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 lastly 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 et al. (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 $\xi^* \sim q(\xi|\theta^*_{(k)}, \tilde{X})$.

It then holds that $\xi_1^*, \ldots, \xi_m^*$ 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.

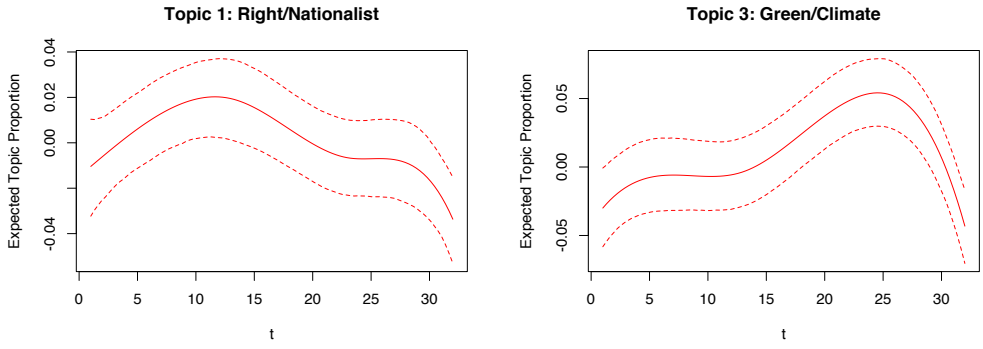

Figure 1: Estimated prevalence of topics 1 and 3 over time, generated using *estimateEffect* from the *stm* package

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, see e.g. Atchison and Shen (1980).

As shown by Atchison and Shen (1980), a distribution that can be used to approximate a logistic normal distribution is the dirichlet distribution. However, note that the dirichlet distribution assumes less interdependence among single elements than the logistic normal

distribution. 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 asymptotically normal for both the beta regression, see Ferrari and Cribari-Neto (2004), p. 17, and the quasibinomial GLM, see e.g. Fahrmeir et al. (2007), p. 285. 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. Here we only visualize the results obtained by the quasibinomial GLM; the results of the beta regression, which show similar trends, are found in the appendix. Setting the number of simulations to 100, we sample $\xi_1^*, \ldots, \xi_{100}^*$ from the marginal 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 regression with logit-link, and calculate the predicted proportions. As mentioned, when visualizing the impact of a particular covariate, all other covariates are held at their median (or majority vote, if categorial), in line with the methodology employed in the *stm* package.

We exemplarily illustrate the relationship of covariates and topic proportions for topics 3 ("green/climate") and 4 ("social/housing") for a subset of prevalence covariates. However, the linear predictor of our regressions takes the same form as in (1.1), i.e. we do not use a subset $\tilde{X}$, but the full set of prevalence covariates X, to estimate the effects. 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 observe its proportion increases over time until September 2019, corresponding to month t=25, decreasing afterwards. The absolute changes in topic proportions over time are rather small (around 4%). The percentage of immigrants within a electoral district shows a negative and rather steady relation to the climate topic. Furthermore, topic tends to be discussed more frequently is GDP is high, although the link to GDP is somewhat ambiguous. Finally, the unemployment rate is positively linked to topic 3.

Regarding the effect of categorical variables on topic "green/climate", we consider the

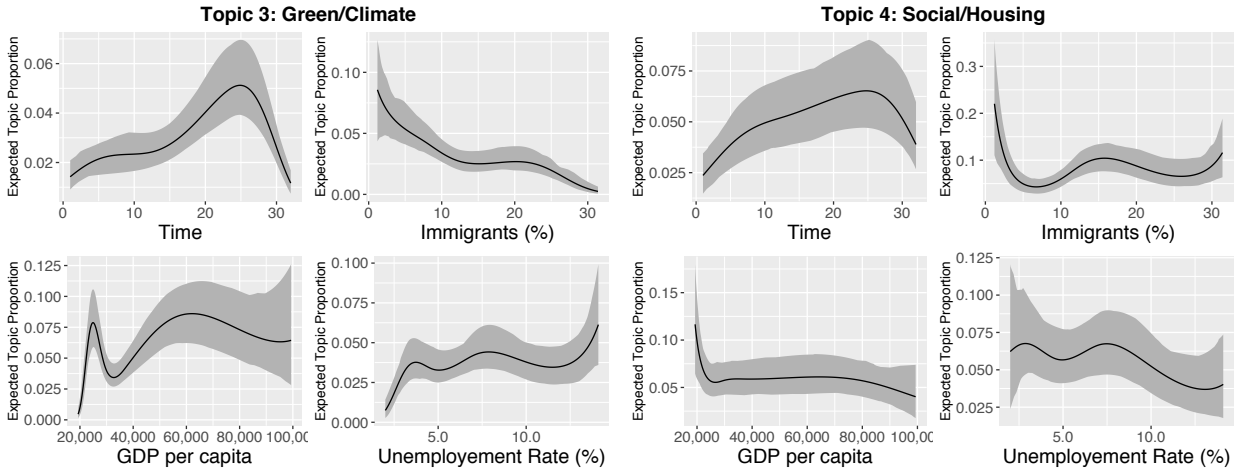

Figure 2: Mean and 95% credible intervals for smooth effects, obtained using a quasibinomial GLM.

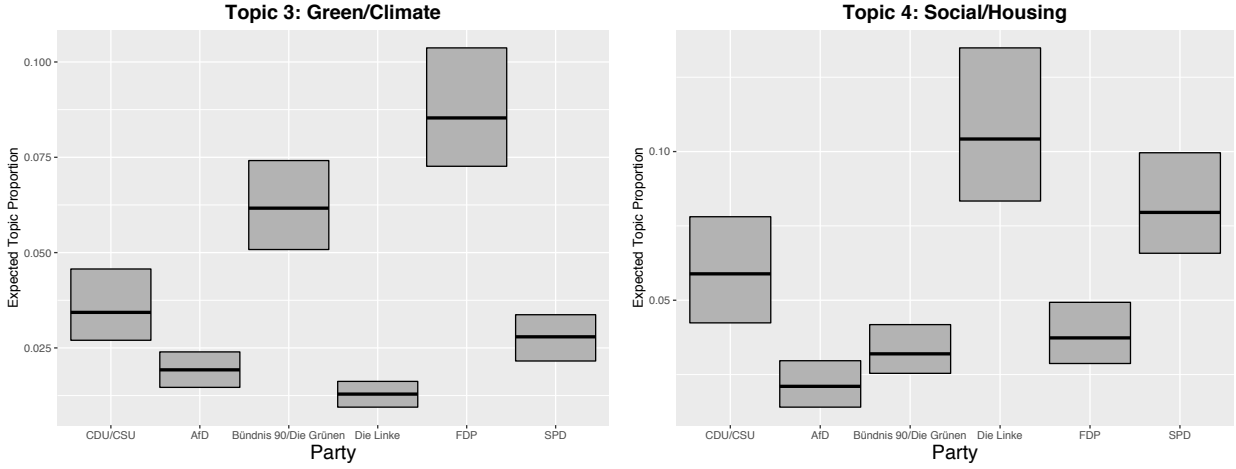

Figure 3: Mean and 95% credible intervals for different political parties, obtained using a quasibinomial GLM.

political party, arguably the most decisive covariate. As 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 for topic 4, "social/housing", we observe that most 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 8% for topic 3. The time effect is similar to the one for topic 3, particularly the decreasing topical prevalence since September 2019. For the other variables, no clear effect is discernible.

The effect of the political party on the relevance assigned to the topic "social/housing" is very much in line with a priori expectations: the left party and social democrats have the highest topical 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 topical 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/climate" and "social/housing" vary between 4% and 12% and between 5% and 10%, respectively. For topic 1, "right/nationalist", note how topical 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 below 3%.

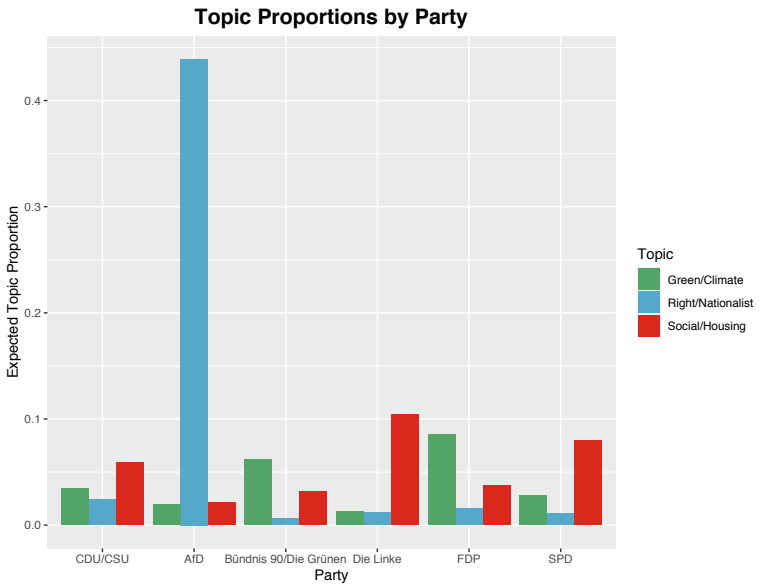

Figure 4: Topical prevalence by political party for topics 1, 2, and 3.

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, Blei et al. (2007) replaced the dirichlet distribution of the LDA by a logistic normal distribution in order to allow for a joint dependence among topic proportions. 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 $\hat{\Sigma}$. Thus, for a given new observation x_d^*, we can sample θ_d^* from $\text{LogisticNormal}_{K-1}(\hat{\Gamma}^T (x_d^*)^T, \hat{\Sigma})$:

1. Draw $\eta_d^* \sim \mathcal{N}_{K-1}(\hat{\Gamma}^T (x_d^*)^T, \hat{\Sigma})$ and set $\eta_{d,K}^* = 0$.

2. For all $k = 1, \ldots, K$: $\theta^*_{d,k} = \frac{\exp\left(\eta^*_{d,k}\right)}{\exp\left(\sum_{i=1}^{K} \eta^*_{d,i}\right)}$.

3. $\theta^*_d := (\theta^*_{d,1}, \ldots \theta^*_{d,K})^T$.

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

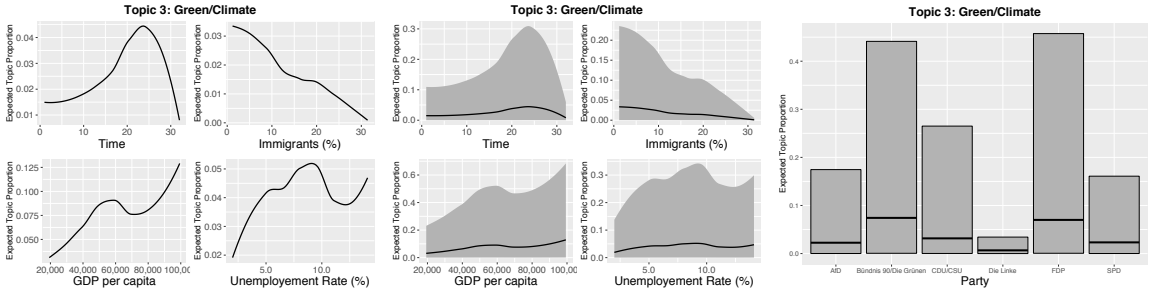

Figure 5: Smooth effects without credible intervals, smooth effects with credible intervals and effect of the political party.

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 5 provide a more realistic picture than those obtained in case of a separate modeling of topic proportions, since the usage of the logistic normal distribution of topic proportions is an implicit assumption made within the stm that there is a dependence among topics, as argued above. This ultimately produces a large variance of the univariate marginal distributions of topic proportions, as can be observed.

2 Appendix

2.1 Plots of section 4.4

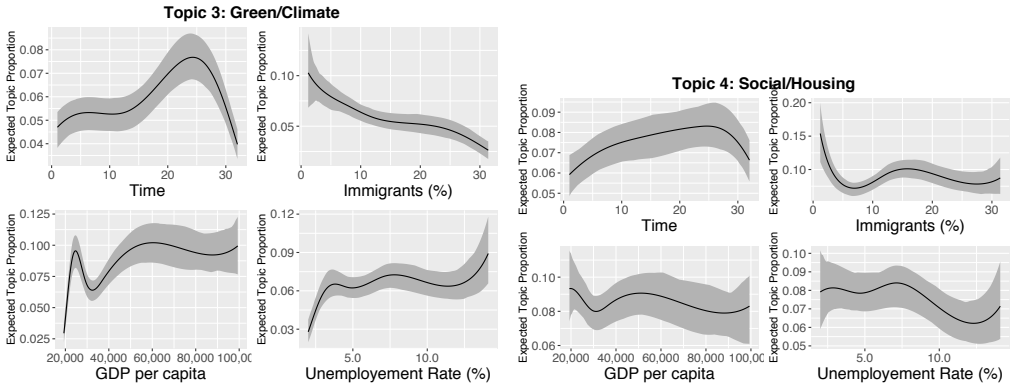

Figure 6: bla

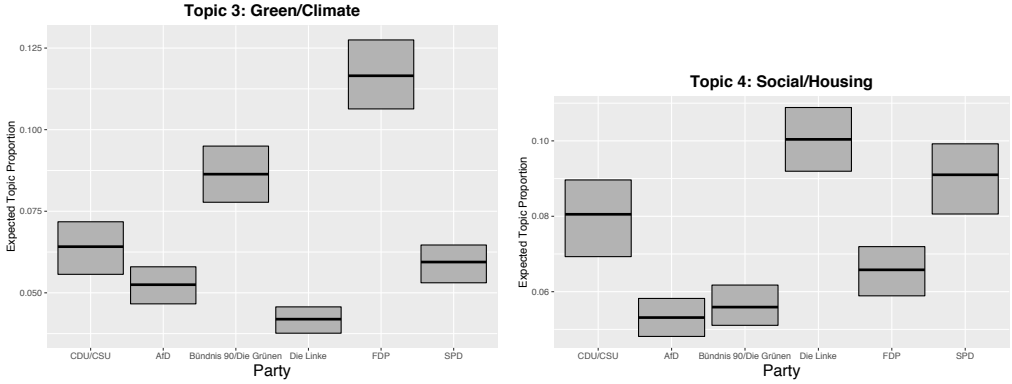

Figure 7: blabla

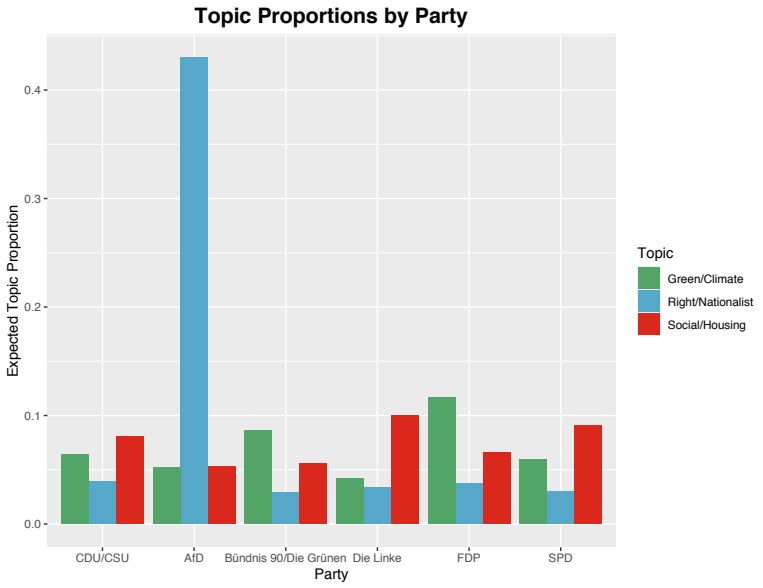

Figure 8: Topical prevalence by political party for topics 1, 2, and 3.

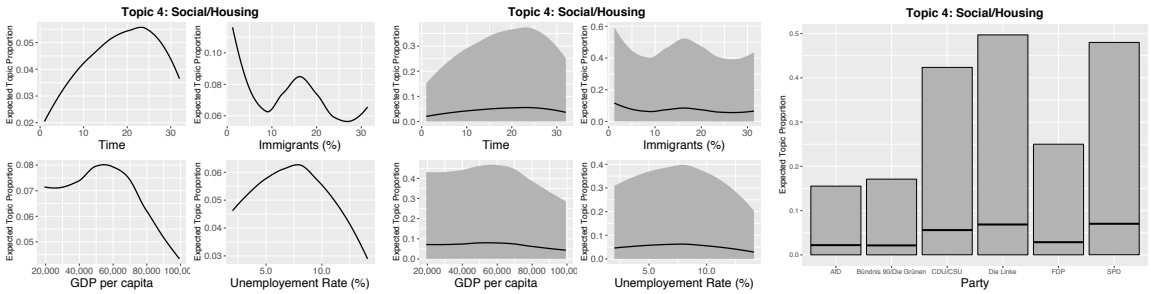

Figure 9: bla

References

J Atchison and Sheng M Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

David M Blei, John D Lafferty, et al. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression*. Springer, 2007.

Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815, 2004.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, pages 1–20. Harrahs and Harveys, Lake Tahoe, 2013.

Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003, 2016.

Martin A Tanner. *Tools for statistical inference*. Springer, 2012.