YOUR TITLE

— Project Report —
Advanced Bayesian Data Analysis

Eldaleona Odole, Leonor Cunha, and Anarghya Murthy

March 7, 2025

TU Dortmund University

Contents

1 Introduction

Outline

did I miss anything? In this report we aim to investigate the relationship between urbanization and partisanship voting outcomes. To put it more concretely we want to answer the question; "How does urbanization of a particular US House distrcit affect the resulting party that is elected?". First we provide background our motivation and important background information about the US election system. We then discuss briefly our modeling approach before looking at other literature trying to understand the relationship between urbanization and voting outcomes. Next we discuss our dataset, which is a combination of four different datasets. We then discuss the set up of our four models in detail, followed by a discussion of the priors and our reasoning behind them. We then go into the results of our models and their convergence diagnostics. Finally we use both graphical and numerical methods to compare our models before concluding with limitations, improvements and reflections.

Background

The association between a voters demographics (gender, age, education etc.) and their propensity to vote for either a democratic or republican candidates is a topic of extensive study. The large political polling organization such as Gallup and Pew Research tend to analyze their polls by breaking respondants into smaller demographic groups cite. However, less is known about the relationship between voting outcomes and the voter's environment. Our modeling would like to investigate the conventional wisdom that says cities tend to be more progressive.

To answer our question about the nature of the relationship between urbanizaiton and partisan voting outcomes, we choose to investigate the 2022 House Election. Every two years the United States elects 435 officials to the House of Representatives. Each state is allocated one of the the 435 House seats with the rest being allocated roughly proportional to the share of total population living in each state cite.

The 2022 House Election takes places during a non-presidential year, and is the most recent election following the 2020 Census and redistricting, allowing us to use the most recently available district maps, and demographic data. Within this report we combine demographic data and urbanization data into a logistic regression model to predict the district voting outcomes of the 2022 House Election.

did I miss something about the models? From this data we use each of the districts as a observation for our logistic regression model. We then choose to model the effect of urbanization hierarchically. Since urbanization is tied to geography we assume a hierarchy related to the location of the districts at the state and region levels of the United States. We then built four different models to test different aspects of the hierarchical structure. In the logistic regression models we then combine urbanizaiton with various demographic covariates. By using a bayesian approach to this multilevel model, we were able to gain a more precise understanding of the relationship between urbanizaiton, geography and

voting outcomes. All of our models were fit using the `brms` R package cite.

Comparative Literature

Inherent in our analysis is the assumption that repuclican and democratic voters are not evenly distributed, rather we assume that the distribution of voters is informative for our analysis. Much of the current research related to the relationship between urbanization and partisan voting focuses on the practice of gerrymandering. Gerrymanding is practice of redrawing the voting district boundaries in favor of a particular political party. We ignore the impact of gerrymandering in our analysis, since the method of determining if a distict has been gerrymandered is rather unclear and out of the scope of this project.

Since we are using a custom curated dataset, there exists no directly comparablere-search, however we did find interesting ideas related to urbanization and partisianship. One example of this would be the idea of the inefficent distribution of Democrats in cities, as explored by CITE. In their analysis they discuss the phenomena of spatial inefficency wherein higher concentrations of democratic voters in urban areas leads to fewer districts voting for democratic candidates that would be expected. Rather than trying to predict the partisan outcome of a particular district, their analysis focuses on measuring partisan spatial efficency in order to understand the impact of gerrymandering CITE.

Although the authors in cite title look at election outcomes through the lense of spatial efficency and we are trying to look at election results of districts with respect to their urbanizaition, we can take the following insights from their analysis. First the distribution of Democrats and urbanization are inexplicably linked, they found that all over the country Democrats tend to be concentrated in urban areas which indicate that it will likely be a good predictor. We would like to investigate to what extent that is true. They also note that the effect of spatial efficency is highly dependent on state and the branch of goverment. When considering that urbanization can be considered a proxy for spatial inefficency, this further supports our decision to model urbanindex hierachically.

2 Dataset

Our dataset was made by combining four independent datasets related to the 2022 House election. The first dataset is the publically available urbanization dataset published by fivethirty eight from which we incorporate the variables urbanization index and (urban) grouping into our final dataset Holly Fuong, 2022. From the description of the dataset: "The urbanization index is calculated as the natural logarithm of the average number of people living within a five-mile radius of every census tract in a given district, based on a weighted average of the population of each census tract. The population of a census tract is according to 2020 census data. This provides a numerical value for how urban or rural a district is. " Holly Fuong, 2022. The urbanization dataset was put together by FiveThirtyEight as part of their analysis *The Republican Path To a House Majority Goes Through The Suburbs* which gave election predictions leading up to the 2022 U.S. Congressional Eleciton Skelly, n.d. The other variable we included in our curated dataset

was urban grouping, which is a collapsing of the urban index into catagories ranging from urban to rural. We however did not end up including this in our model.

The second dataset used in our analysis the Election Results Dataset from FiveThirtyEight Mehta, n.d. It is a continuously updated repository of United States Govenor, Congressional and Presidential elections. maybe talk about the funny shit with alaska and preference voting (?) DONE As this dataset includes all elections going back to 1998, we only used a subset of the data relevant to the 2022 House Election. Since each state sets its own elections rules this led to additions quirks in the dataset. For example as of 2020 Alaska uses ranked choice voting, and each stage of the rank choice vote is present in this dataset, which made the ensure the proper filtration of the dataset that much more important CITE - press release of alskan gov. From this dataset we used the party, state, and winning party variables, it also included the variable incumbent party, which we considered also using, but the data from incumbent party is partially incomplete and would likely have very high correlation with the our target variable winning party, leading to a model that relies most heavily on incumbent party. Since we are more interested in the relationship between our model and our choose covariates we therefore decided not to include it.

As a small aside, although technically a multiparty system, the U.S. is often called a two party system due to the domination of the major political parties the Democrats and Republicans (cite). These parties dominate particularly on the fedral level because political candidates are required to get a plurality of votes rather than a majority of votes which the two largest parties often reach. This is further reinforced as would be third-party voters, often vote for one of the two major parties so ensure their voice is heard, rather than using thier vote on a candidate unlikely to reach a plurality of votes (cite). Within our dataset, there are no districts represented by a third-party candidate and as such we will refer to the U.S. as being a two party system. Therfore we modeled winning party as a binary variable, with democrats encoded as (1) and republicans encoded as (0). It is also important to note that the Democratic Party is considered the progressive party and the Republican Party the conservative party, this is worth saying as these terms are at times used interchangably throughout this report.

The third data used in our analysis is a subset of the 2022 American Commuity Survey Data. The American Community Survey is a yearly survey collecting information about the occupations, education attainment, income and other demographic information carried out by the United States Census Bureau. The United States Census Bureau provides an online tool to access its extensive survey database, which can then be filtered and refined for further analysis. For our analysis we included the following variables for each House district in our curated dataset; total population, percentage women, median household income, mean household income, percentage retirees, percentage bachelors degree holders above the age of 25 years old and unemployment rate among those above the age of sixteen.

From these covariates we hoped to capture education, income, and, demographic make-up of the districts because we thought they might be influential in determining partisan voting outcomes. However not all of these covariates were included in our final

models. As previously explained, the districts are drawn in such a way that the total population of each district should be approximately the same cite, so while this does help support our exchangability assumption, we determined that total population itself would be an unsuitable covariate. Although we intially thought that percentage women may also be a poor predictor-since women should be evenly distributed throughout the United States-however a Pew Research Survey found that women tend to lean more democrat and have higher turnout, which is why we then included it in our largest model. cite. We also initially thought that median and mean income could be combined to as a measure of inequality, but found a study saying that inequality is not a good voting indicator on its own cite. For that reason we choose to use only the median household income as it would be less skewed. Similarly unemployment is not a good indicator on its own, rather when unemployment is high the incumbent is more likely to loose regardless of partycite. We also decided to use percentage retirees instead of median age, as they were highly correlated and median age also includes a part of the population that cannot vote. Why is pct retirees included in every single one of our models?

The fourth dataset was the region dataset, which was put together manually by us following the four statistical region designations of United States Census Bureau. Cite: First link on wiki list of regions of us

Data Cleaning

We then merged these four datasets to created our own curated dataset. We did this by merging the different datasets on shared variables. As previously said, each observation represents a particular house district, so for the first three datasets, we simply merged them based on their state and district number. To include the regions we simply used the state variable for each district.

In the election results dataset, we encountered one instance of missing data. For all of the districts in Louisiana, the winning party was not recorded. There was however the incumbant party recorded in the Election Results dataset and by cross referencing this with public record we found for the 2022 House Election only candidates from the incumbant party remained in each of the districts. Therefore as is reflected in our code base we used the data from the incumbant party in place of winning party for the state of Louisiana. For all other states and disticts we did not encounter this problem.

In terms of scaling we wanted all the variables to be on roughly the same scale to aid in convergence times. In order to do that we roughly scaled median income and total population by dividing total population by one million and dividing median income by one hundred thousand. This brought each of these to roughly the same scale as the other variables. Similarly we decided to scale all of the percentage variables to be on the scale zero to one hundred rather than zero to one to make their coefficents more interpretable.

3 Models

The Winning party in each congressional district race ($y_{i,j,k}$ for district i, state j, region k) can be modeled as the outcome of Bernoulli trial, since this is a binary variable:

$$y_{i,j,k} \sim Ber\left(\pi_{j,k} = logit^{-1}(\theta_{j,k})\right) \tag{1}$$

with probability of a Democrat win $\pi_{j,k}$ modeled as the inverse logit transform of $\theta_{j,k}$, a linear combination of our covariates. The inverse logit function converts real numbers into quantities between 0 and 1, and is therefore a standard way to model probabilities cite.

We tested four different models for $\theta_{j,k}$, which include different covariates in addition to our variable of interest (Urban index) and incorporate our data's hierarchical structure in different ways. Therefore, all four are Multilevel Bayesian (Logistic) Models, which require particular assumptions: CITE first, that a logistic regression accurately represents the relationship between the log-odds of a Democrat win and the explanatory variables, that is, $\theta_{j,k}$ and our covariates are linearly related; second, exchangeability, meaning that each district is exchangeable within each state and each state is exchangeable within each region; and third, that the value of urban index (and other covariates) in a district has a different effect depending on the state/region it belongs to.

The logistic relationship is common choice for modeling binary outcomes. It allows us to model the probability of a Democrat win by a linear predictor which can take any real value, while still having an interpretation for the coefficient estimates (in terms of change of log-odds).(cite).

We assume exchangeability because we assume that the districts are drawn in such a way that they are competative for both parties. Meaning that although the districts may have different characteristics, certain mechanisms can be best captured when thinking of districts as exchangeable parts of a hierarchical model.

One example would be for complicated historical reasons certain regions of the United States are more similar to eachother than others. For example the Southern United states tends to be more religious and religous people tend to vote more conservatively, think about this then as a prior telling us about the mix of democratic and republican districts within a particular state. Although the religiousity may increase the number of potential republican districts in each state, whether one district votes repuclican does not influence the decision of another since each district outcome is determined by thousands of individual votes. Since each district is part of a state and a region, we can translate the complex geographically determined mechanisms into our model by modeling some of parameters hierarchically.

Model 1 (state level)

Our first model includes only our variable of interest, urban index, plus the percentage of retirees as covariates to explain θ, plus an intercept. Murthy didn't you find something in initial testing? please write about that Urban index was modeled hierarchically, with the coefficient varying by state, with a prior dependent on common parameters β_{urb} and

σ_{urb}, which in turn have (hyper-)priors of their own. The intercept is assumed to be non-variant for all districts, as well as the slope of percentage of retirees.

Equation 2 describes our model conceptually.

$$\theta_j = \beta_0 + \beta_{urb,j}^{uncent} \cdot \text{Urban\_Index} + \beta_{ret} \cdot \text{Pct\_Retirees} \tag{2}$$

To better understand what happens in the backend when we want to fit this model with BRMS, it is helpful to rewrite the equation 3 in terms of 'global' and 'hierarchical' effects. The previously considered coefficient is then decomposed into these effects, i.e. $\beta_{urb,j}^{uncent} = \beta_{urb} + \beta_{urb,j}$ with $\beta_{urb,j}$ centered around zero, which does not alter the meaning of the model. cite brms book

$$\begin{aligned}\theta_j = &\beta_0 + \beta_{urb} \cdot \text{Urban\_Index} + \beta_{urb,j} \cdot \text{Urban\_Index} \\ &+ \beta_{ret} \cdot \text{Pct\_Retirees}\end{aligned} \tag{3}$$

Although we assume that there is indeed state-level clustering in the district election outcomes, we have 50 states, and some of them include only one or two districts. This can make the hierarchical estimates unreliable.

Model 2 (region level)

To overcome this problem, we fit another model, with only one difference from the previous one: the hierarchy is at the region level, rather than state. This means the coefficients of urban index vary by region now, with a common mean and variance which are parameters to be estimated themselves. Equation 4 describes this model, in its specification with separate global and hierarchical effects for urban index.

$$\begin{aligned}\theta_k = &\beta_0 + \beta_{urb} \cdot \text{Urban\_Index} + \beta_{urb,k} \cdot \text{Urban\_Index} \\ &+ \beta_{ret} \cdot \text{Pct\_Retirees}\end{aligned} \tag{4}$$

Model 3 (nested)

In this model we include the entire geographical hierarchy: a *nested hierarchy* of districts within states within regions. Here the assumption is that the effect of urbanindex ($\beta_{urb,j:k}$) depends on state j and region k through a prior with mean parameter $\beta_{urb,k}$, which in turn varies by region and depends on hyper-mean β_{urb} (which has its own prior, with hyper-hyper-parameters). Equation 5 specifies the model, with the centered around zero formulation.

$$\begin{aligned}\theta_{j,k} = &\beta_0 + \beta_{urb} \cdot \text{Urban\_Index} + \beta_{urb,k} \cdot \text{Urban\_Index} \\ &+ \beta_{urb,j:k} \cdot \text{Urban\_Index} + \beta_{ret} \cdot \text{Pct\_Retirees}\end{aligned} \tag{5}$$

Model 4 (big model)

This is our most extensive model. Here we used urban index and 4 additional covariates plus an intercept to explain $\theta_{j,k}$. It can be seen as an extension of Model 1, as urban index

is modeled hierarchically by state. The region level hierarchy is instead included only in the effect of percentage of bachelors degrees. Median income effect is also considered to vary by state, and the intercept and the slopes of percentage of women and percentage of retirees were modeled non-hierarchically.

Equation 6 describes this model, in the brms adapted specification.

$$\begin{aligned}\theta_{j,k} =& \beta_0 + \beta_{women} \cdot \text{Pct\_Women} \\ &+ \beta_{urbindex} \cdot \text{Urban\_Index} + \beta_{urbindex,j} \cdot \text{Urban\_Index} \\ &+ \beta_{bsc} \cdot \text{Pct\_Bachelor's} + \beta_{bsc,k} \cdot \text{Pct\_Bach.} + \beta_{inc} \cdot \text{Median\_Income} \\ &+ \beta_{inc,k} \cdot \text{Median\_Income} + \beta_{ret} \cdot \text{Pct\_Retirees}\end{aligned} \tag{6}$$

4 Priors

Priors represent our initial beliefs about our model parameters' distributions. In each of our models, this means a prior for the intercept, one for the slope of each covariate that is modeled non-hierarchically (e.g., β_{ret} in all four models) and, in the case of covariates with global and hierarchical effects, priors for the hyper-parameters as well.

Table 1 lists our selected priors for each model, by corresponding covariate.

	Model 1	Model 2	Model 3	Model 4
Intercept	$\beta_0 \sim N(0, 10)$	$\beta_0 \sim N(0, 10)$	$\beta_0 \sim N(0, 10)$	$\beta_0 \sim N(0, 10)$
Urban Index	$\beta_{urb} \sim N(0, 1)$ $\beta_{urb,j} \sim N(0, \sigma_{urb})$, $\sigma_{urb} \sim Halfcauchy(10)$	$\beta_{urb} \sim N(0, 1)$ $\beta_{urb,k} \sim N(0, \sigma_{urb})$, $\sigma_{urb} \sim Halfcauchy(10)$	$\beta_{urb} \sim N(0, 1)$ $\beta_{urb,k} \sim N(0, \sigma_{urb,1})$, $\sigma_{urb,1} \sim Halfcauchy(10)$ $\beta_{urb,j:k} \sim N(0, \sigma_{urb,2})$, $\sigma_{urb,2} \sim Halfcauchy(10)$	$\beta_{urb} \sim N(0, 1)$ $\beta_{urb,j} \sim N(0, \sigma_{urb})$, $\sigma_{urb} \sim Halfcauchy(10)$
Pct.retirees	$\beta_{ret} \sim t(1, -2, 1)$	$\beta_{ret} \sim t(1, -2, 1)$	$\beta_{ret} \sim t(1, -2, 1)$	$\beta_{ret} \sim t(1, -2, 1)$
pct.women				$\beta_{women} \sim N(0, 1)$
pct bsc				$\beta_{bsc} \sim t(1, 0, 1)$ $\beta_{bsc,k} \sim N(0, \sigma_{bsc})$, $\sigma_{bsc} \sim Halfnormal(0, 1)$
median income				$\beta_{inc} \sim N(0, 1)$ $\beta_{inc,j} \sim N(0, \sigma_{inc})$, $\sigma_{inc} \sim Halfnormal(0, 1)$

Table 1: Priors defined for each of our four models, for each parameter; parameters and distributions are listed by their corresponding covariate.

We define such distributions before seeing the data, based on existing literature and our own intuition about the effects of our covariates on the probability of a democrat win. We assumed the same priors for the same terms included in different models (intercept, percentage of retirees, urban index for the same levels).

For the intercept β_0 we set a Normal prior centered at zero with a large standard deviation. This represents a weakly informative prior, as we had no strong beliefs about the intercept value, nor does it have any straightforward interpretation in our model: it theoretically represents the (logit of the) probability of a Democrat win in a district with no urbanization at all, a median income of zero dollars, and 0% of women, retirees and citizens with a bachelor's degree in the population; such a district is obviously nonexistent.

For the population-level component of the urbanindex slope we opted for a standard normal prior in all models. We chose not to make assumptions on the sign of the effect of this variable, as it is this variable that we are interested in studying, although we are assuming that its absolute value will be below 1.96 with 95% certainty.

All group-level (zero-centered) priors are Normal, by brms specification is there a reason???

For the standard deviation of the hierarchical effects we opted for a relatively weakly informative prior, a $(half)Cauchy(0, 10)$. We do not want to place strong constraints on the effect of our variable of interest, hence we 'allow the estimates to fluctuate'.

As found in the literature older votes tend to vote more conservatively CITE, so it makes sense that a higher he Percentage of retirees in each district negatively correlates with the probability of a Democrat winning, but we do not know how strong this effect ought to be. Therefore, for the prior on β_{ret} we chose a distribution centered around a negative number, and with relatively heavy tails, reflecting our uncertainty, for all models.

In Model 4, Pct.Women is not modeled hierarchically. Although women tend to vote more democratic and have higher voter turnout CITE, the percentage of women is roughly the same in every district, so we do not expect this covariate to have a strong effect on the probability of either party winning, i.e, we expect β_{women} to be close to zero. So, we set a prior for this slope which is centered around zero and has little variability: a standard normal prior.

The effects of percentage of bachelor degrees and median income are parameterized in 2 hierarchical levels: an average slope across all districts, and a varying slope by group (State or Region), $\beta_{covariate,j}$ or $\beta_{covariate,k}$, which follows a Normal distribution centered at zero with standard deviation modeled at group level (by a hyperprior). This was done to capture the idea of different costs of living and importance of education respectively, as what is considered rich, poor, or well educated for an individual is highly dependent on the environment. what do u think?

As we expected the population-level effects of both Median Income and Pct Bscs to be positive in some cases and negative in others, we picked symmetric priors for both β_{bsc} and β_{inc}. We are, however, less sure about the on-average-null effect of the Percentage of Bachelor degrees, so for this slope parameter we opted for a prior with 'fatter tails', the standard Cauchy distribution rather than the Normal one, representing a higher degree of uncertainty.

For the hyperparameters, we chose a half standard normal prior for both the standard deviations of $\beta_{bsc,k}$ and $\beta_{inc,j}$. This is a narrow distribution, with most values falling between 0 and 1, as we expect to see weak effects for these covariates, and thus small standard deviations (and positive, as any SD is by definition).

5 Code

choices

1. number of chains and iterations

2. brms i guess

6 Results

This section is not on the instructions but is probably the easiest way to talk about the results we got

Each of our models has different estimates for the parameters, even when they are starting from the same prior. As the estimate is given by the mean of the posterior distribution for that parameter, this reflects the different posterior distributions between models, which is only natural given the different structures each model assumes for our data.

The intercept for example, is very different between models, and is particularly large (in absolute terms) in Model 4. Although, as discussed before, the intercept does not have a practical interpretation in our model, it is interesting that this Model's estimate is so far away from the value we initially assumed it would take (between -19.6 and 19.6 with 95% certainty, with a $N(0, 10)$ prior). It can indicate ....

Our variable of interest, urbanindex, has a more consistent slope estimate across models, between 1.22 (Model 2) and 1.66 (Model 1). Figure 1 shows the histograms of posterior draw for the urbanindex slope parameter, by Model. The results indicate that the variable has, as expected, on average a positive effect on the log-odds of a Democrat Party win. Model 3, for example, estimates that on average a 1 unit increase in urbanindex (e.g., 10 to 11) translates into a 1.64 increase in the log-odds (or an $e^{1.64} \approx 5.155$ increase in the odds) of a Democrat being elected in the district. While the estimates are within the range we assumed with the prior for this parameter, the upper bounds of the credible intervals are already close to the boundary of that prior. A prior with more probability mass on the positive range, or a less informative prior, could push the estimates further up. not sure about this though

The standard deviation estimates for the hierarchical effect of urbanindex are all on the lower side of the prior, close to zero. This is true both for the cases when the hierarchy on urbanindex includes State or Region.

The estimated slope of pct.ret is also different between models, particularly in Model 4, where it is larger in absolute value, but always negative as we had assumed. Note that although this estimate is smaller in absolute terms than the slope of urbanindex, it does not mean that percentage of retirees has a weaker effect than urbanization in election outcome. These are different variables on a different scale: in model 3 for example, it is estimated that an increase of 1 percentage point in the percentage of retirees in a district correlates with a 0.16 decrease in the log-odds of a Democrat being elected.

The other covariates included in model 4 also have slope estimates within our expected range. Median Income has a larger (absolute) estimated slope coefficient and standard deviation, but this is likely due to the scale of the variable: 1.43 is the estimated average decrease in log-odds of a Democrat win when the median income of a district increases by 100,000 dollars, which is a big jump. The inclusion of these covariates has some effect on the magnitude of the slope estimate for our variable of interest (urbanindex),

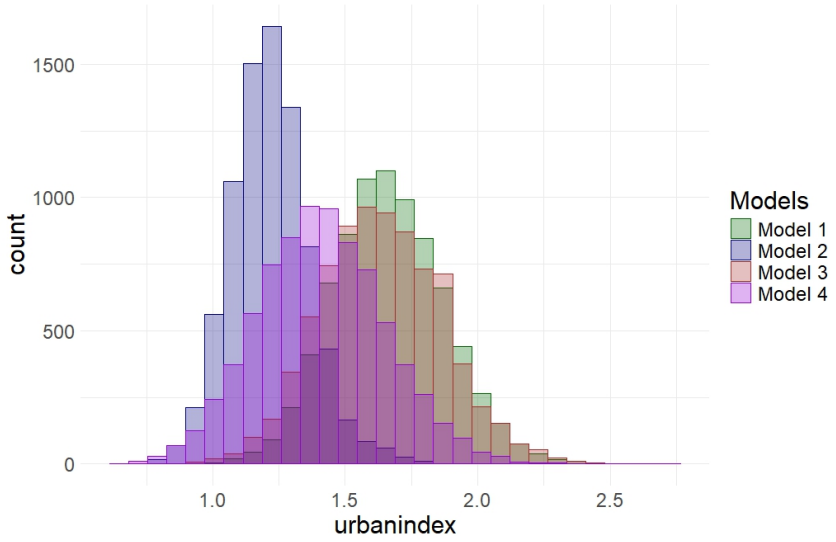


Figure 1:

but not on its sign, and not as much as including a State-specific hierarchical effect vs. Region-specific one.

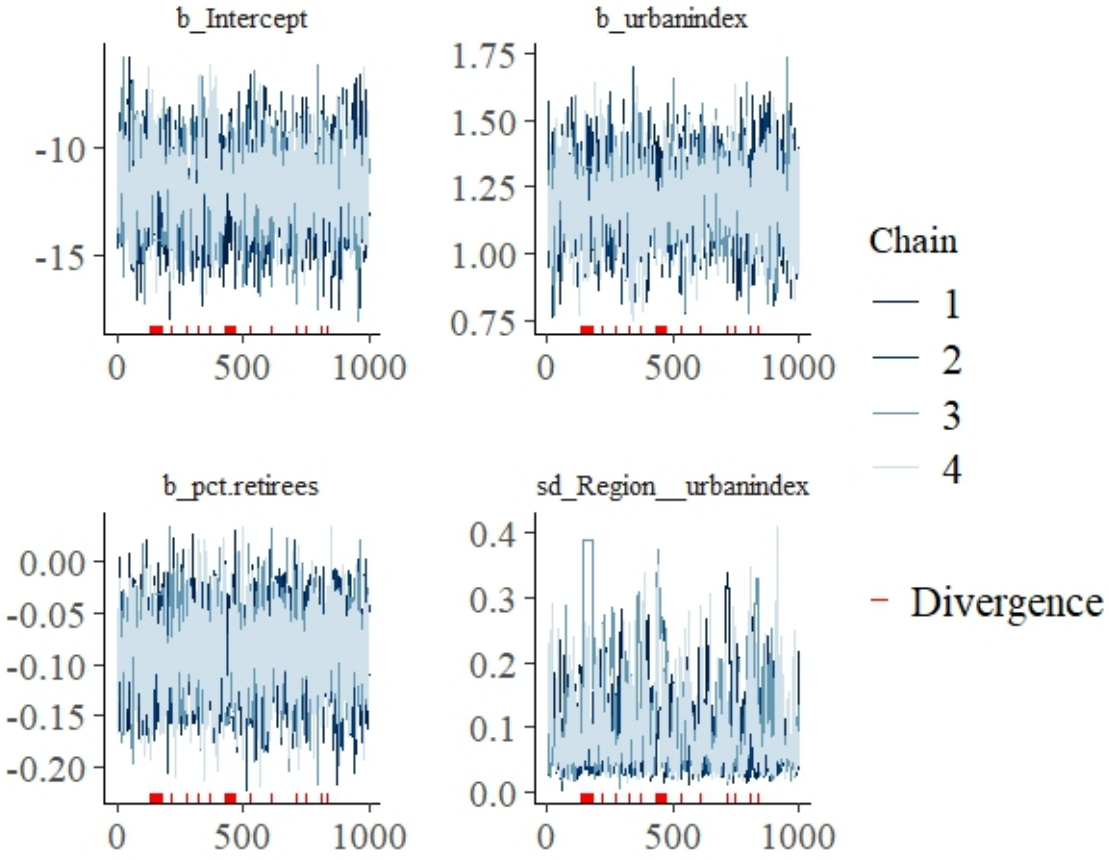
MARGINAL EFFECTS??????

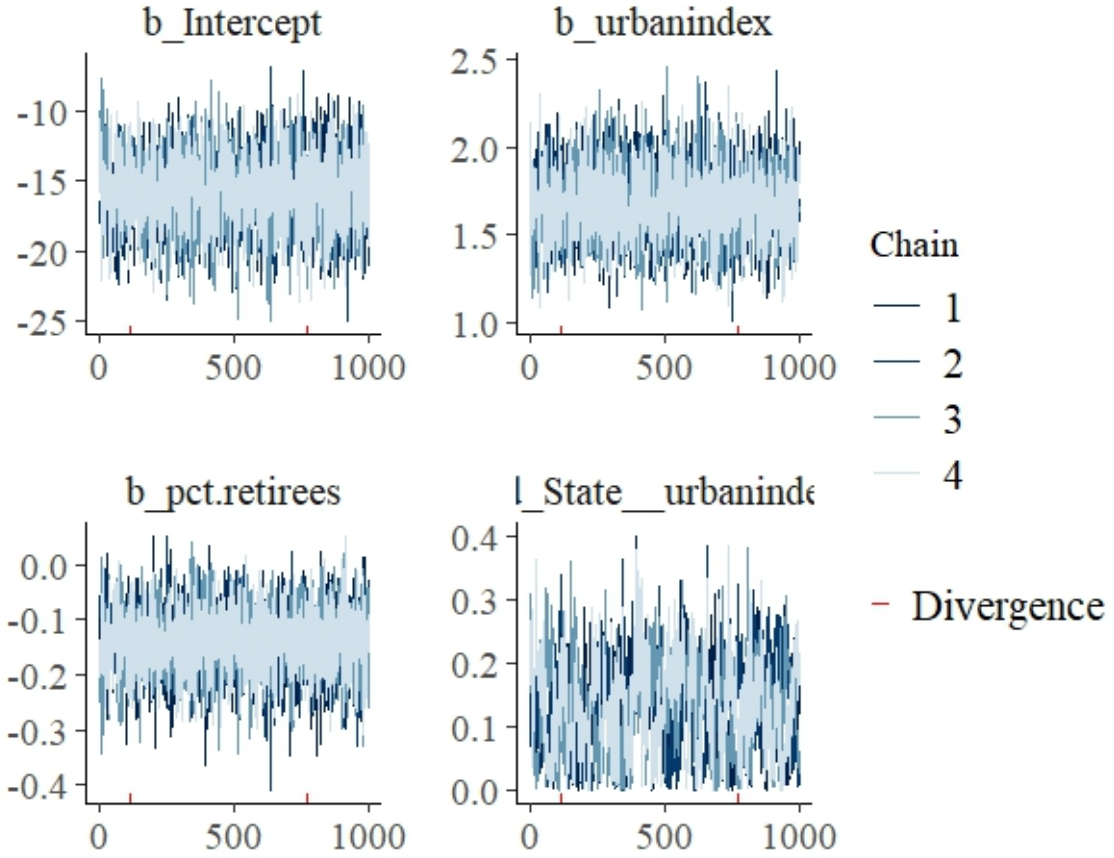
7 Convergence Diagnostics

One of the fundamentals of Bayesian analysis is its reliance on MCMC sampling. This ensures we have access to both the posterior samples and (in our case) the posterior regression coefficients themselves. All our data analysis was done using BRMS, which runs on STAN, which itself uses the Hamiltonian Monte Carlo algorithm for the posterior generation.

"Convergence " in layman's terms can be described as, 'Do the posterior draws get closer and closer to a specific value?'.

HMC Convergence diagnostics in itself can be a rather extensive topic, so for this project we mainly consider graphical diagnostics, namely: the MCMC trace plots as provided by BRMS.


b_Intercept
-10
-15
0
500
1000
b_urbanindex
1.75
1.50
1.25
1.00
0.75
0
500
1000
Chain
1
2
3
4
b_pct.retirees
0.00
-0.05
-0.10
-0.15
-0.20
0
500
1000
sd_Region__urbanindex
0.4
0.3
0.2
0.1
0.0
0
500
1000
Divergence



The trace plots for model 2 (top) and most coefficients of model 3 (bottom) are plotted above. We see divergent transitions in both cases, but there are two key differences:

Firstly, the number of divergent transitions: There are many more of these for Model 3 compared to Model 2. Secondly, the place of occurrence: Model 3's divergent transitions occur throughout the chains, while those in Model 2 seem to be concentrated around the 1000th iteration, while noticeably decreasing towards the 2000th iteration. This suggests that increasing iterations further might help with Model 2, but is unlikely to help with Model 3.

The trace plots of the other models and coefficients (in Appendix) either resemble those of Model 2, or in the case of Model 1, have no divergent transitions whatsoever.

So our key takeaway is that the HMC chains behave strangely for Model 3, and reasonably for the rest provided we increase the number of iterations from the default 1000 to 2000. Luckily for us, there is one glaring difference between Model 3 and the rest which partially explains this behaviour: Namely, it is the only model which takes into account both a regional hierarchy and a statewise hierarchy for the same coefficient, with the former being nested in the latter. The other models have at the most a single hierarchy per coefficient.

While this in itself is not enough to conclude that the specifications of the other models are correct, it is certainly a good preliminary finding (that the presence of an explicit nested hierarchy causes HMC chain convergence issues).

8 Model Comparison

Our four models were built based on somewhat different assumptions about the structure of our data, and all produced slightly different results. We need to know which of these is *better*, that is, which results are more trustworthy and allow us to answer our original research question. To this end, we measured and compared our models' predictive performance first by looking at absolute in-sample predictive performance, then at relative and finally at the Leave-One-Out statistics to compare out-of-sample Predictive Performance.

Absolute predictive performance metrics directly tell us directly how well the model performs, without looking at other models. To measure absolute predictive performance we used the Root Mean of Squared Error (RMSE). The RMSE for the s-th posterior draw is obtained from the predictive errors, that is, observed outcome y_n minus posterior draw $\hat{y}_n^{(s)}$, squaring those errors and taking the root of their average over all observations, as explained in Equation 7. Since it is computed for each draw, as opposed to a single point estimate, it takes into account the posterior uncertainty, making it a fully Bayesian measure citation needed?.

$$RMSE^{(s)} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(y_n - \hat{y}_n^{(s)}\right)^2} \tag{7}$$

The RMSE measure works in a similar way to the R squared statistic that is commonly used to assess the fit of linear models, by evaluating differences between observations and model predictions, but RMSE retains the scale of the response variable. Usually this would mean it has a direct interpretation in the context of the problem (e.g., by how many units is our prediction off), but here we are dealing with a binary response variable, so what the RMSE actually represents is the average distance between the predicted value and 0 or 1, not the true probability of a Democrat win. A model with higher RMSE in this context is not necessarily better at estimating this true value, just on average estimates probability values that are larger for districts where a Democrat won and smaller where a Republican won instead.

Figure 2 shows the histograms of the RMSE with all draws for each of our models. Model 4 estimates on average closer percentages to the actual outcome which, in a certain sense, can mean it estimates voting outcomes better. This is however unsurprising, as this model has many more covariates, and at least some have a large effect on the probability of a Democrat win. More important for our analysis is the comparison between models 1,2 and 3. Model 2 (the one with only the Region-level hierarchy) shows by far the largest average errors, which can mean this model's predictions are more often "wrong" (a small estimated probability when a Democrat has won or a large one when a Republican was elected) or that the predictions are in general closer to the center (0.5) rather than the extremes. In either case, it suggests that the Region-level hierarchy in urbanindex is not as powerful as the State-level one for estimating probabilities. Model 3 has on slightly smaller RMSE values than Model 1, but the histograms are very close, suggesting the inclusion of the Region hierarchy really does not have a large impact on posterior predictions.

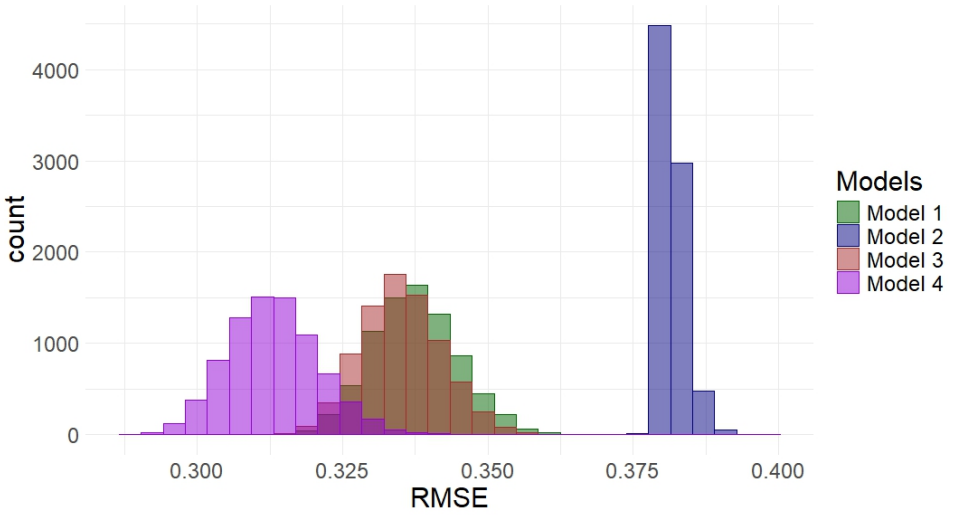


Figure 2:

Relative predictive performance measures, contrary to absolute ones, do not have an interpretation in themselves, only as a comparison between models. To assess relative predictive performance, we looked at log-likelihood scores, that is, the average of posterior draws' log-likelihoods for each observation. This is a relative predictive performance measure in the sense that it does not tell us anything about the model's predictive performance alone, we need to compare it between different models to establish which is better. So, we examine the differences in log-likelihood scores between models. The sum of these differences corresponds to the difference in Log Predictive Density (LPD) between models, where LPD for a given model is the sum of Log-Likelihood scores of all observations in the model.

Comparison	LPD_Diff	SE_LPD_Diff
Model 1 - Model 2	42.00	7.25
Model 1 - Model 3	-2.32	1.29
Model 1 - Model 4	-20.24	3.91
Model 2 - Model 3	-44.32	7.31
Model 2 - Model 4	-62.24	8.20
Model 3 - Model 4	-17.92	3.66

Table 2: text

Table 2 shows the differences in LPD between all four of our models and corresponding standard errors. Model with higher LPD are preferred, as this reflects an overall higher (less negative) log-likelihood score across all observations, meaning the model more accurately predicted each result. The difference between Models 1 and 2 is positive, meaning Model 1 has higher LPD, so it is preferred over 2 using this statistic. In fact, Model 2 is never preferred to any of the other models. Once again, Model 4 provides the best fit

and Models 1 and 3 are very close, with 3 performing slightly better. This, together with the comparatively poorer fit of Model 2 points towards Region-level hierarchy possibly not being the correct choice for modeling urbanindex.

Both RMSE and LL scores are in-sample predictive performance metrics. In-sample predictive performance measures evaluate only model predictions for the same observations which were used to fit the model in the first place, therefore they only evaluate how well a model predicts the data it was trained on, which means there is a danger of overfitted models, which would not generalize well to new data, performing much better under these metrics cite brms?. In our case, the bigger model (Model 4) was indeed the preferred one using both RMSE and LL scores.

Because we are comparing models with different degrees of complexity, it is essential to check also out-of-sample predictive performance metrics. These metrics are computed by splitting the dataset into training data and test data, fitting the model on the former and computing the expected LPD (ELPD) of the observations in the latter, given the model estimates with the training set.

$$ELPD = \sum_{n=1}^{\tilde{N}} p(\hat{y}_n|y) \tag{8}$$

The way we choose to split the data into training/test sets naturally impacts the ELPD. So, we rely on cross-validation: we do multiple different splits and average over the results. Our chosen method was Leave-One-Out (LOO) cross-validation, which in theory performs as many splits as observations in the dataset, each time leaving one "out" as the test data. In practice, a different posterior is not actually computed this many times, but rather an estimate from the full model posterior using importance sampling (Pareto-Smoothed Importance Sampling in this case).

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
Model 4	0.00	0.00	-156.49	11.59	32.57	3.25	312.98	23.19
Model 3	-17.99	4.23	-174.48	12.11	33.08	3.33	348.97	24.22
Model 1	-20.35	4.61	-176.84	12.09	33.37	3.43	353.68	24.18
Model 2	-45.87	8.39	-202.36	12.25	6.19	0.58	404.71	24.50

Table 3: text

According to the LOO statistics (Table 3), Model 4 (the largest) is the preferred one, followed by Models 3, 1 and 2. The first column of the table shows the difference between each model and the best performing one (in terms of ELPD score, shown in the third column), ranked by best to worst model. This is the same ranking we had seen before, with the more complex models performing better. There is a relatively sizable difference in LOO scores between Model 4 and 3, as well as between 1 and 2, but the difference between 3 and 1 is minimal. The extra level of hierarchy (Region) in the urbanindex slope estimate improves model predictions, although not by much.

All in all, the ranking of our Models' posterior predictive ability is the same, whether we use in- or out-of-sample techniques: Model 4 is the better model for prediction, the

second best is Model 3, closely followed by 1 and lastly 2. All seem to indicate that modeling urbanindex with a Region hierarchy is not the appropriate choice if the goal is to better predict the election outcome.

9 Prior Sensitivity Analysis

another table with priors here, maybe not all because thats a lot

One of the most important parts of Bayesian Data Analysis is setting the prior distributions. The choice of priors could greatly affect the final results in a model.. (cite prior sensitivity guys) So, we conducted a prior sensitivity analysis, by refitting our model(s) using alternative priors (which also fit our model assumptions) and assessing the impact in our results.

[BRIEFLY explain new priors, graphs comparing them]

10 Limitations and Improvements

11 Changes from Presentation

Please keep track of the changes we make from the presentation

1. stuff probably for exmaple the priors
2. changed the scales for percentages to be on the 100
3. priors: intercept from N(0,0.5) to N(0,10); sd for urban index in bigger model from Gamma(2,5) to halfcauchy(10) to match all other models; percentage retirees in bigger model center from -1 to -2, to match other models
4. model order (1-2-3-4 to 3-4-2-1)

12 Conclusion

12.1 Reflection and Discussion

please lets call this subsection something else, this sounds so childish

Appendix

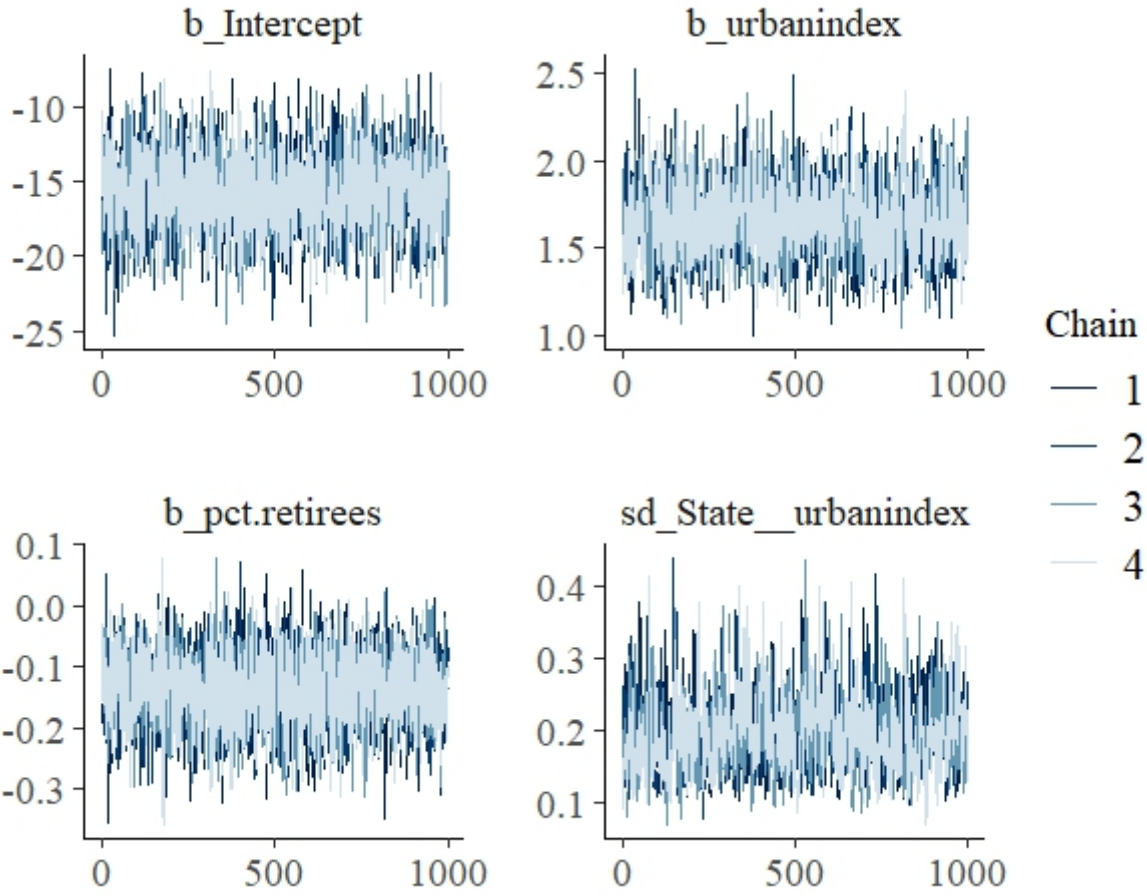
Descriptives

	Winning.party	urbanindex	pct.women	Median.Household.Income	pct.retirees	pct.bach
Winning.party	1.00	0.60	0.20	0.24	-0.30	0.29
urbanindex	0.60	1.00	0.30	0.39	-0.39	0.46
pct.women	0.20	0.30	1.00	-0.15	0.14	0.05
Median.Household.Income	0.24	0.39	-0.15	1.00	-0.10	0.76
pct.retirees	-0.30	-0.39	0.14	-0.10	1.00	-0.09
pct.bach	0.29	0.46	0.05	0.76	-0.09	1.00

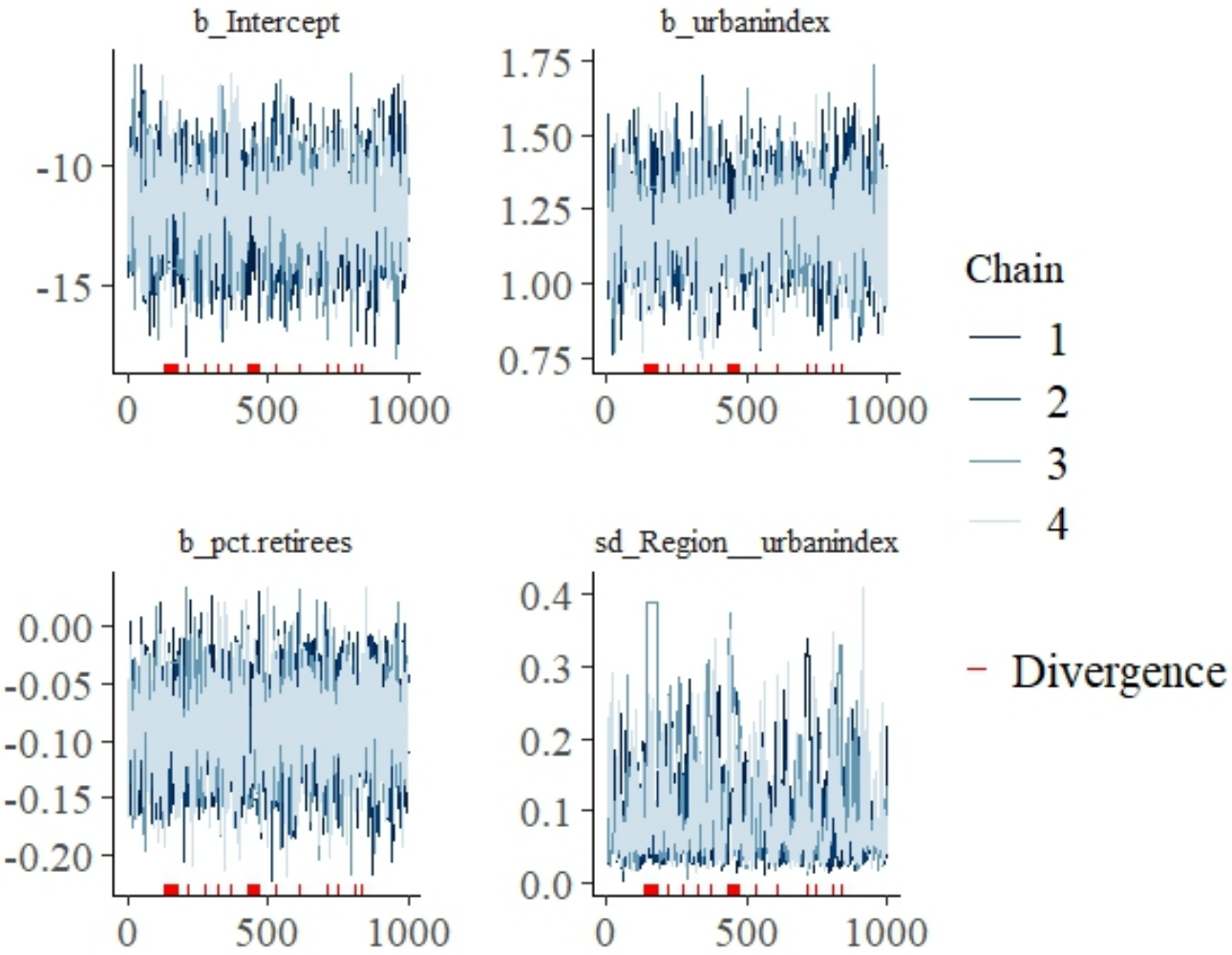
Table A1: Correlation matrix

Trace Plots

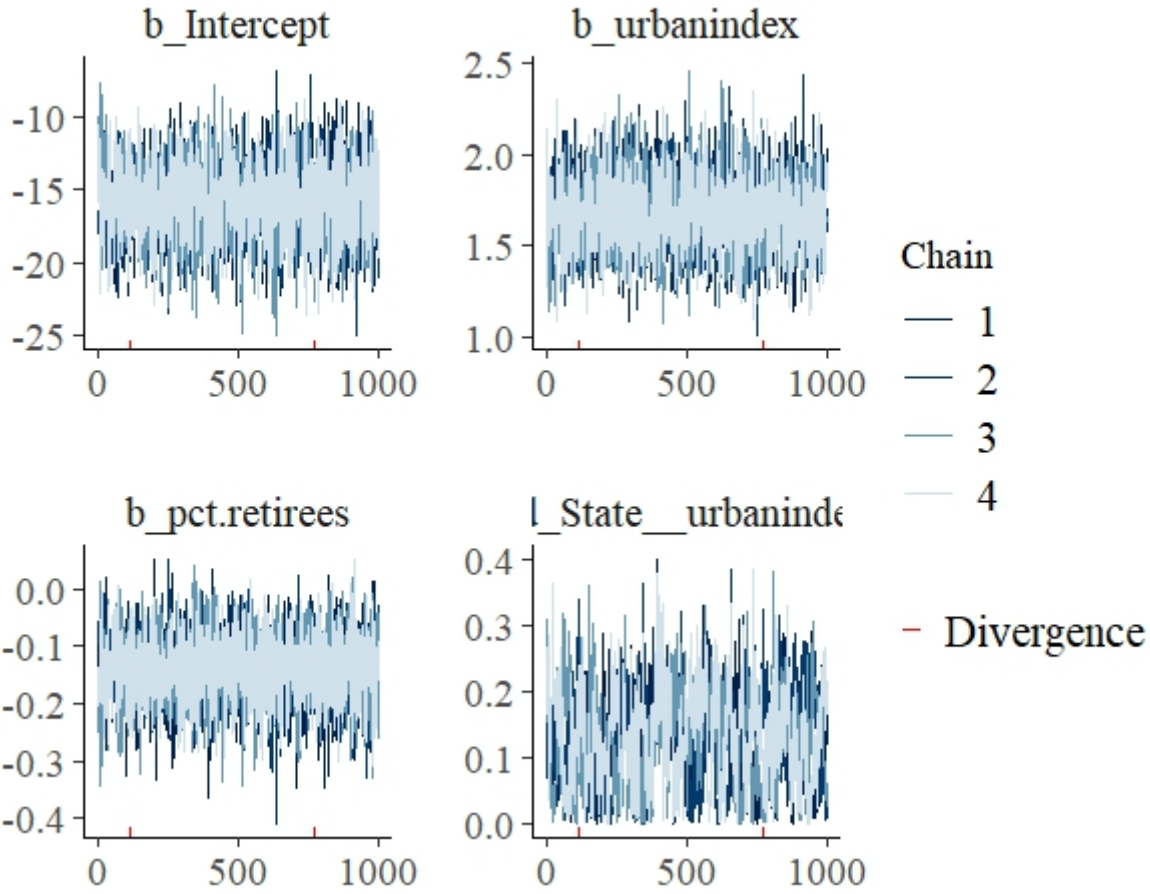
Model 1:



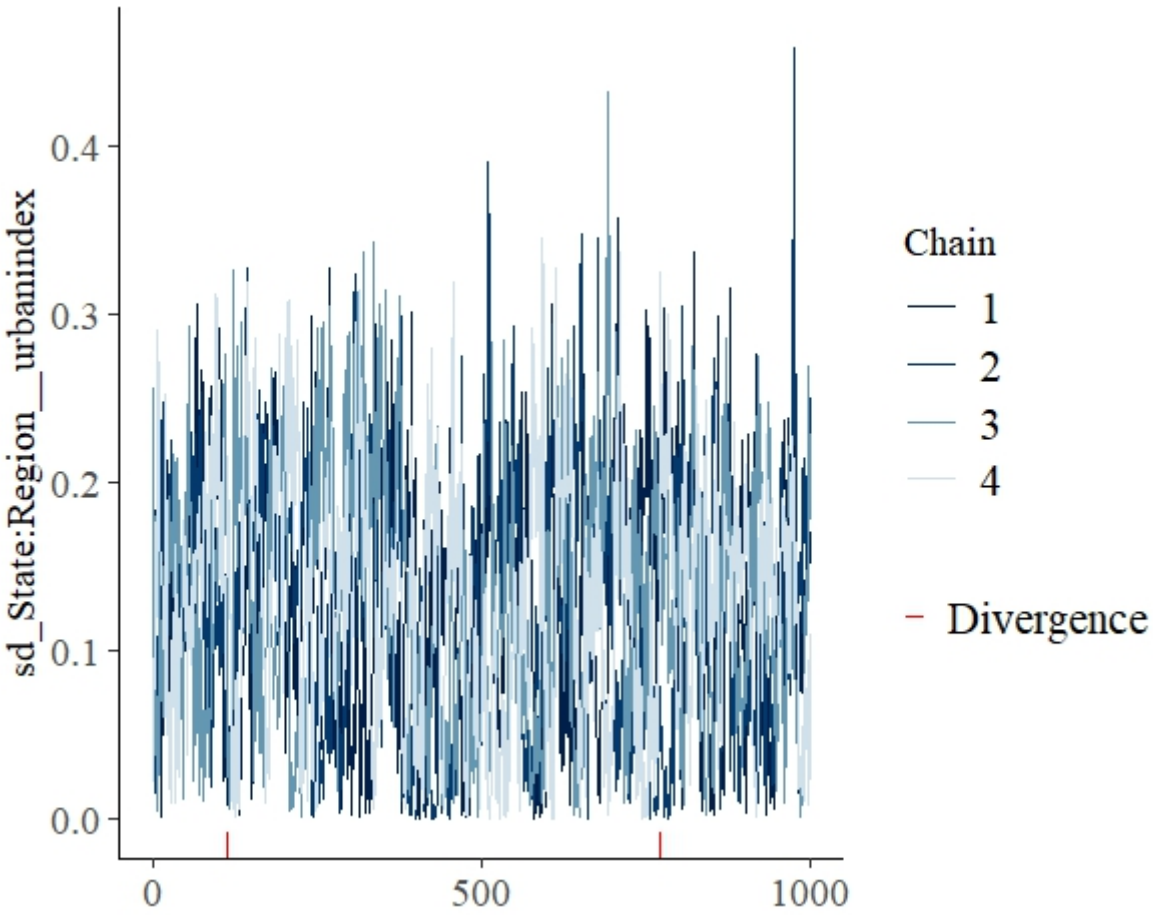
Model 2:


b_Intercept
b_urbanindex
b_pct.retirees
sd_Region__urbanindex
-10
-15
1.75
1.50
1.25
1.00
0.75
0.00
-0.05
-0.10
-0.15
-0.20
0.4
0.3
0.2
0.1
0.0
0
500
1000
Chain
1
2
3
4
Divergence

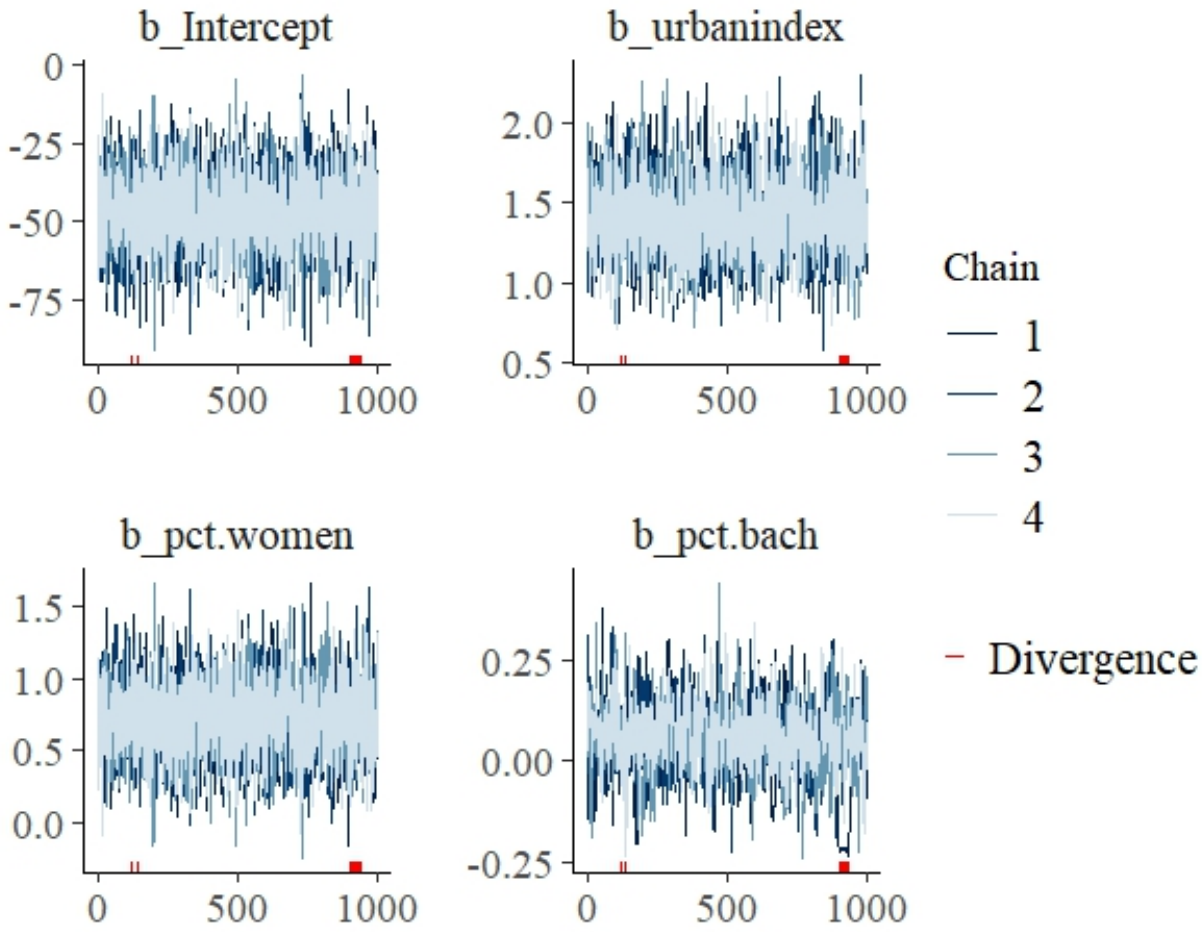
Model 3 Pt 1:


b_Intercept
b_urbanindex
b_pct.retirees
l_State__urbanind
-10
-15
-20
-25
2.5
2.0
1.5
1.0
0.0
-0.1
-0.2
-0.3
-0.4
0.4
0.3
0.2
0.1
0.0
0
500
1000
Chain
1
2
3
4
Divergence

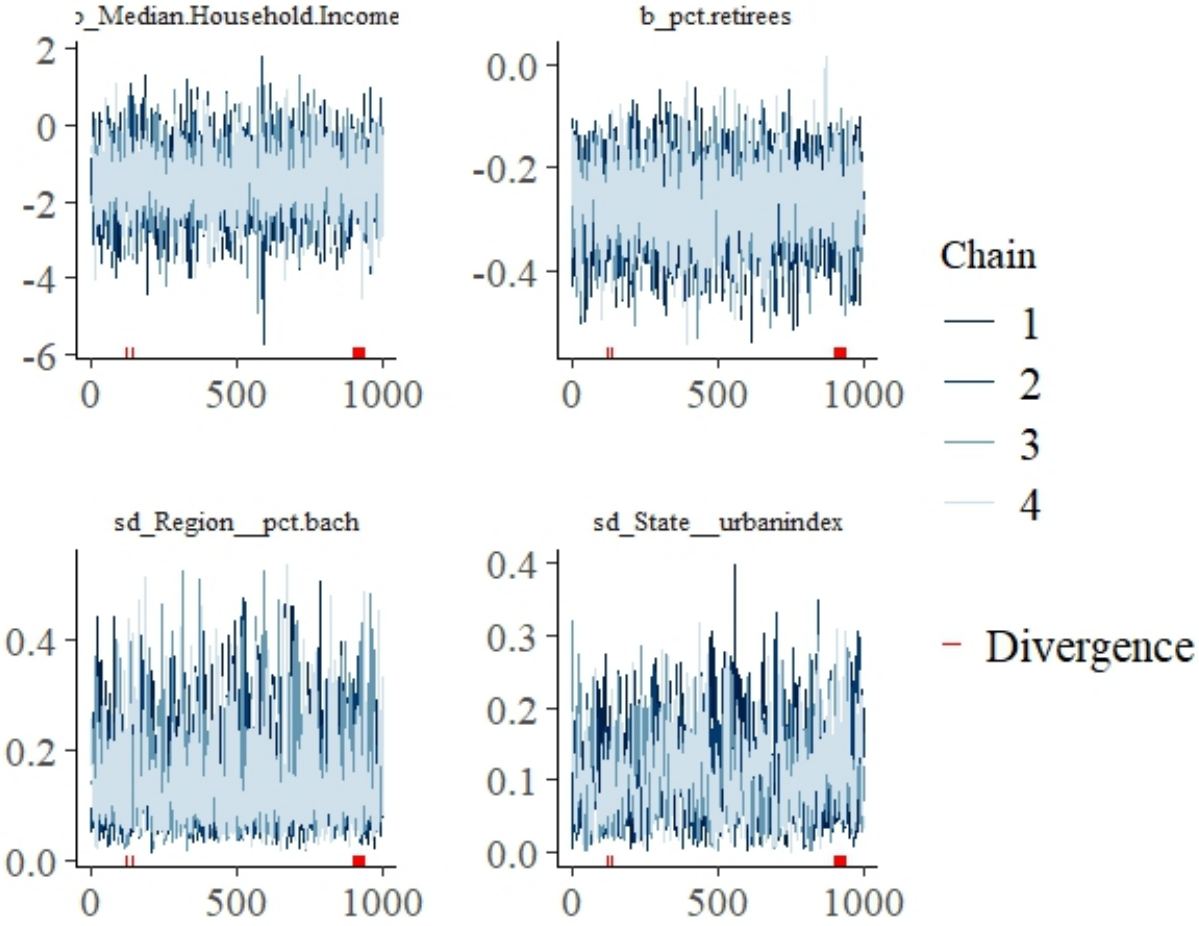
Model 3 Pt 2:


sd_State:Region__urbanindex
0.4
0.3
0.2
0.1
0.0
0
500
1000
Chain
1
2
3
4
Divergence

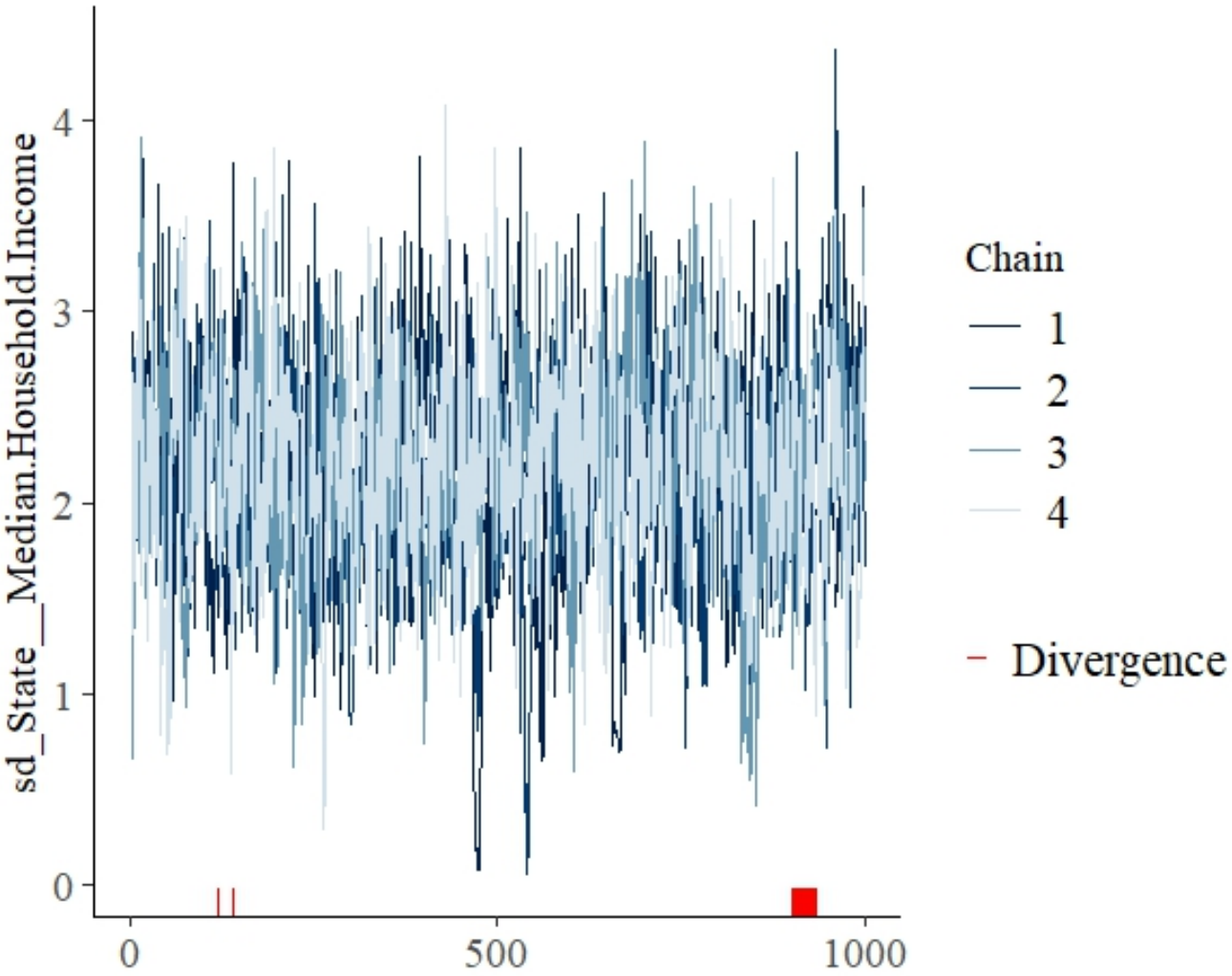
Model 4 Pt 1:


b_Intercept
b_urbanindex
b_pct.women
b_pct.bach
Chain
1
2
3
4
Divergence

Model 4 Pt 2:


b_Median.Household.Income
b_pct.retirees
sd_Region__pct.bach
sd_State__urbanindex
Chain
1
2
3
4
Divergence

Model 4 Pt 3:



Model Estimates

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Regression Coefficients:							
Intercept	-15.75	2.63	-20.96	-10.73	1.00	7397.71	6010.33
urbanindex	1.66	0.21	1.28	2.10	1.00	5191.43	5331.31
pct.retirees	-0.14	0.06	-0.27	-0.02	1.00	6086.48	5724.70
Multilevel Hyperparameters: State (Number of levels: 50)							
sd(urbanindex)	0.19	0.05	0.11	0.30	1.00	1569.53	3108.00

Table A2: Parameter estimates obtained by Model 1 with the 435 observations in our dataset

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Regression Coefficients:							
Intercept	-12.01	1.78	-15.58	-8.60	1.00	2892.71	3356.25
urbanindex	1.22	0.15	0.95	1.53	1.00	1294.59	473.21
pct.retirees	-0.09	0.04	-0.16	-0.01	1.00	4773.23	4217.12
Multilevel Hyperparameters: Region (Number of levels: 4)							
sd(urbanindex)	0.09	0.06	0.02	0.27	1.00	892.96	465.15

Table A3: Parameter estimates obtained by Model 2 with the 435 observations in our dataset

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Regression Coefficients:							
Intercept	-15.02	2.69	-20.32	-9.87	1.00	2837.78	5324.03
urbanindex	1.64	0.23	1.21	2.10	1.00	2787.63	2875.43
pct.retirees	-0.16	0.06	-0.29	-0.04	1.00	5424.17	5097.97
Multilevel Hyperparameters: Region (Number of levels: 4)							
sd(urbanindex)	0.21	0.16	0.03	0.67	1.01	348.01	120.67
Region:State (Number of levels: 50)							
sd(urbanindex)1	0.19	0.05	0.11	0.29	1.00	1976.07	3875.21

Table A4: Parameter estimates obtained by Model 3 with the 435 observations in our dataset

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Regression Coefficients:							
Intercept	-45.84	12.13	-70.40	-22.99	1.00	6668.56	5729.57
urbanindex	1.41	0.24	0.95	1.90	1.00	7707.06	6016.13
pct.women	0.70	0.26	0.21	1.24	1.00	6155.31	4936.03
pct.bach	0.06	0.09	-0.11	0.27	1.00	1593.83	662.43
Median.Household.Income	-1.43	0.87	-3.12	0.25	1.00	6817.58	3411.18
pct.retirees	-0.26	0.08	-0.42	-0.12	1.00	6809.18	5806.15
Multilevel Hyperparameters: Region (Number of levels: 4)							
sd(pct.bach)	0.15	0.09	0.04	0.39	1.00	1855.37	1959.66
Multilevel Hyperparameters: State (Number of levels: 50)							
sd(urbanindex)	0.08	0.06	0.00	0.21	1.00	866.41	1941.43
sd(Median.Household.Income)	2.24	0.51	1.19	3.26	1.00	2627.23	2493.48

Table A5: Parameter estimates obtained by Model 4 with the 435 observations in our dataset

Model Comparison Statistics

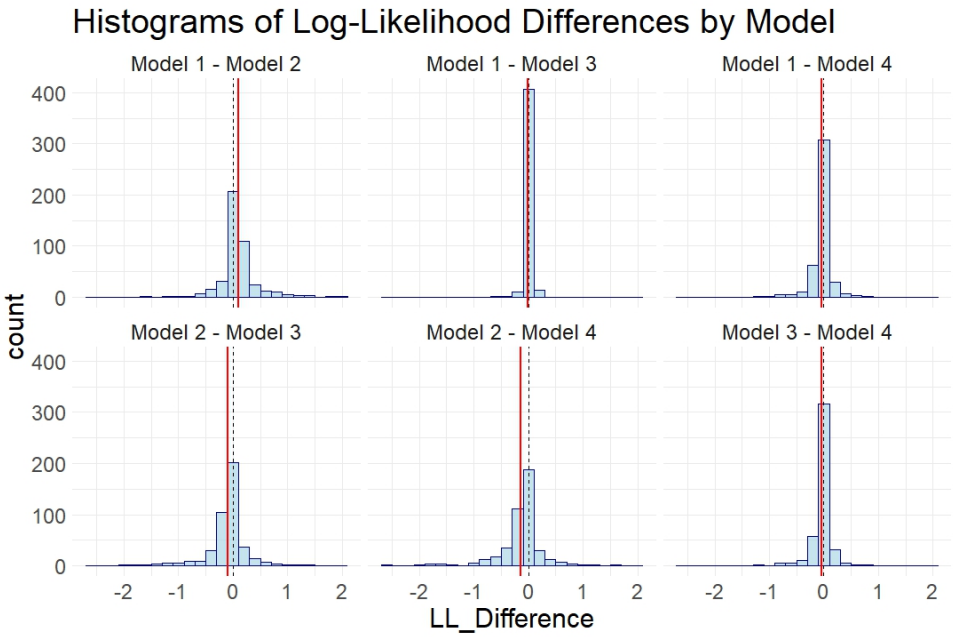


Figure A1: Histograms of differences in Log-Likelihood Scores for each observation between all 4 models; black dotted line marks zero; red vertical line marks the mean of the difference between the corresponding models

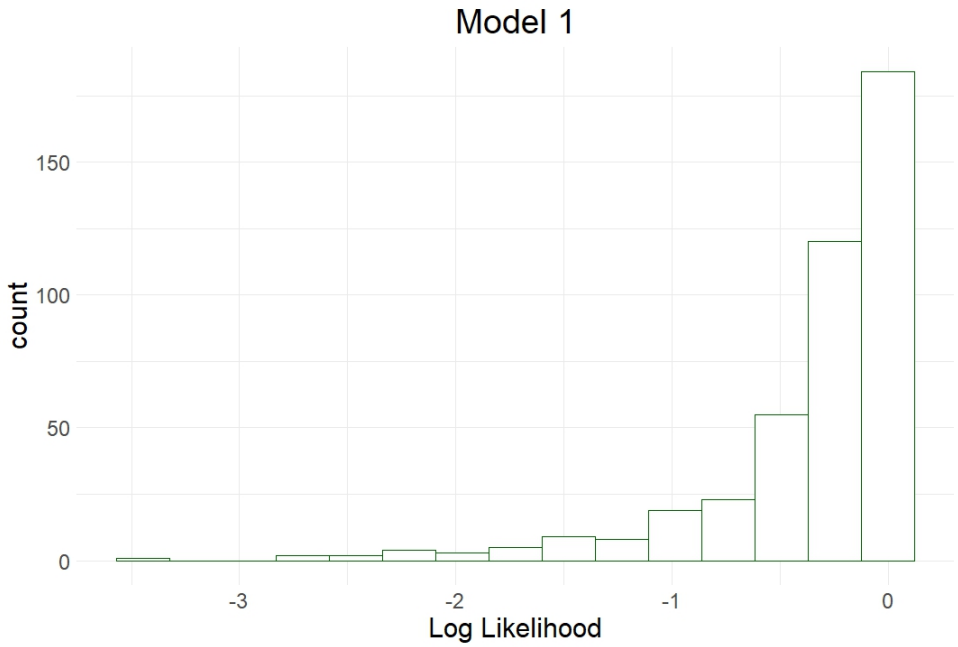


Figure A2:

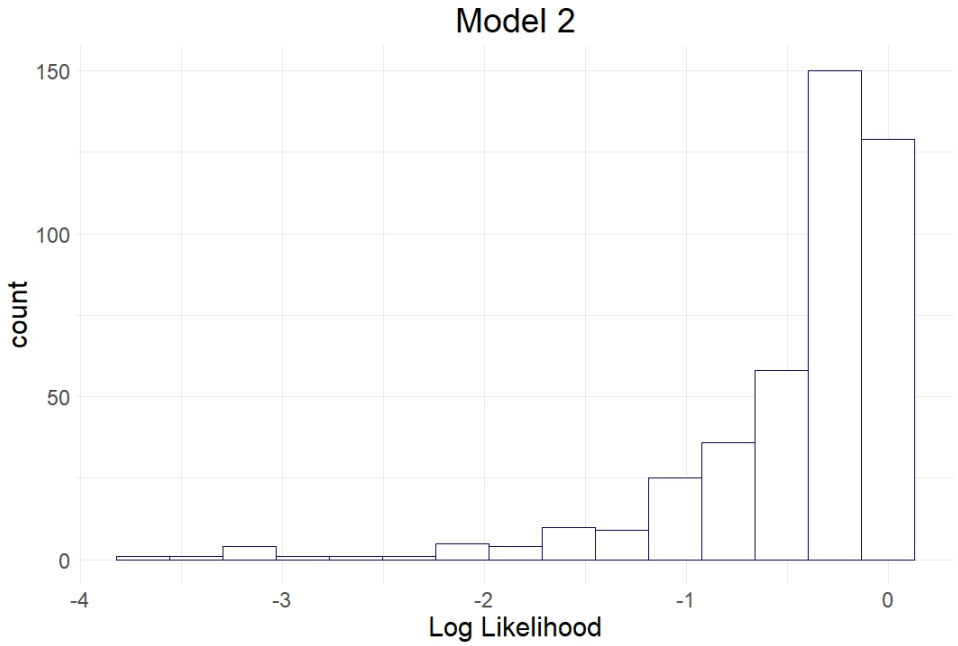


Figure A3:

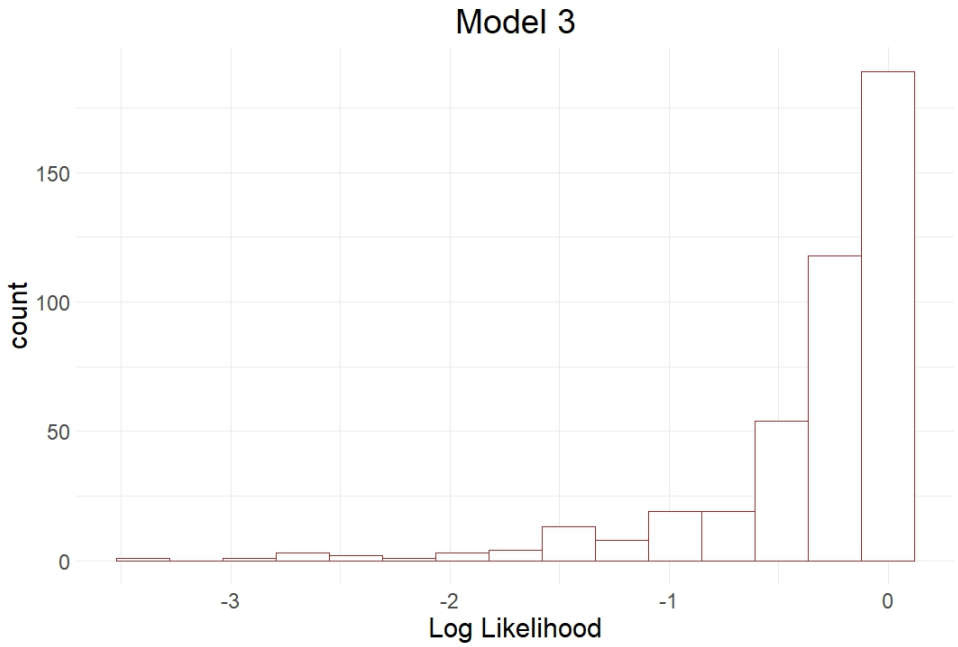


Figure A4:

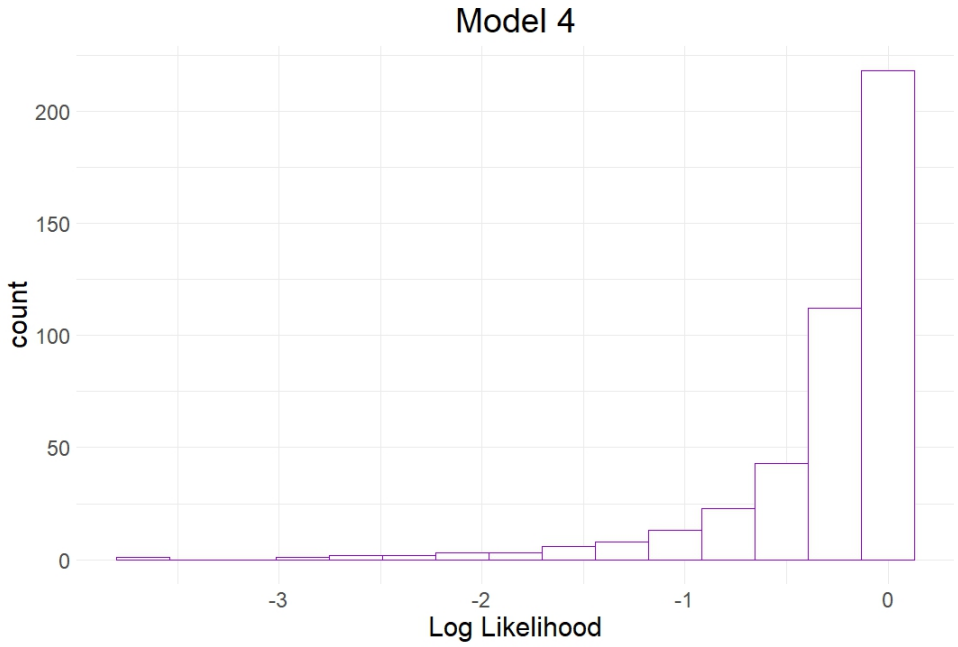


Figure A5:

References

Holly Fuong, G. S. (2022). District urbanization index 2022. https://github.com/fivethirtyeight/data/tree/master/district-urbanization-index-2022

Mehta, D. (n.d.). Election results. https://github.com/fivethirtyeight/election-results/blob/main/README.md

Skelly, G. (n.d.). *The republican path to a house majority goes through the suburbs*. https://fivethirtyeight.com/features/the-republican-path-to-a-house-majority-goes-through-the-suburbs/