YOUR TITLE

— Project Report —
Advanced Bayesian Data Analysis

Eldaleona Odole, Leonor Cunha, and Anarghya Murthy

March 3, 2025

TU Dortmund University

1 Introduction

notes to ourselves about things that should go in each section can be in red like this so we dont forget to delete them :)

1.1 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 other org tend to analyze their polls by breaking respondants into smaller demographic groups. 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 put it more concreatly we want to investigate the question; "How does urbanization of a particular US House distrcit affect the resulting party that is elected?".

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. When then use each one the districts as a single replicate since they are approximately equivalent in population size, but the characteristics of voters living in each district varies.

Inherent in our analysis is the assumption that voters for each 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, and assume that the districts are drawn fairly.

As we are using a custom dataset it was difficult to find directly comparable research, 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 show that Democrats tend to more often be located quite densly in the urban core or surrounding nearby suburbs of a city, whereas Republicans tend to live further from city centers. Rather than trying to predict the partisan outcome of a particular district, their analysis focuses on describing partisal spatial efficency. what is partisan spatial efficency/inefficency? This is related to our analysis because where these authors were trying to understand the outcomes as related to distribution of voters within a city we are tyring to do somehting similar but rather look to predicting the outcomes of particular districts with respect to their level of urbanization. Whereas the authors assume some sort of causal mechanism in the urbanization being a determinate variable for partisanship be assume rather that both partisanship and urbanization are independent given some other third latent variable.

conclusions of the introduction or somehing basically 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.

this is something like assumptions but im not really sure what to say Since the districts are allocated based on population, they are approximately the same size, which allows us the make the assumption the differences in voting behavior have something do with people within the districts rather than simply the size of each district. This also allows us the make exchangability assumptions with respect to state and region.

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 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. As this dataset includes all elections going back to 1998, we only used a subset of the data relevant to the 2022 House Election. From this dataset we used the party, state, and winner variables.

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 used the following variables for each House district;

add variables that we used

The fourth dataset was the region dataset, which was put together manually by us following the region designations of I forgot where we got this info tbh

2.1 Data Cleaning

Since we used four different data sources, this meant merging different datasets on shared variables. As previously said, each observation represents a particular house district, so for the first three dataset, 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 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 that are in the range of zero to one as they are percentages.

3 Models

Two party system

Leonor: here is an explaination for it being two parties, kinda long but yeah 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, which led to using logistic regression as a natural choice for modeling.

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: NOTE ON HOW WE KNOW IT IS NOT A 2 PARTY SYSTEM EXCEPT THAT OOPS IT IS

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

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, interchangeability, 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 a common assumption in the literature (cite). We can assume interchangeability because for complicated historical reasons certain regiions 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, as a result the parameter associated with region would likely be smaller or more negative as compared to other regions. The idea behind the differing effect strength of urban index values per region is that a city in a rural area will likely have stronger signal than an city among a bunch of other cities. [rewrite some of this paragraph]

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. Urban index was modeled hierarchically, with the coefficient varying by state, with a prior dependent on common parameters β_{urb} and σ_j, 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 hierarchical 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

	Model 3	Model 4	Model 2	Model 1
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	$\beta_{urb} \sim N(0, 1)$	$\beta_{urb} \sim N(0, 1)$	$\beta_{urb} \sim N(0, 1)$	$\beta_{urb} \sim N(0, 1)$
	$\beta_{urb,j} \sim N(0, \sigma_j)$,	$\beta_{urb,k} \sim N(0, \sigma_k)$,	$\beta_{urb,k} \sim N(0, \sigma_k)$,	$\beta_{urb,j} \sim N(0, \sigma_{urb,j})$,
	$\sigma_j \sim Halfcauchy(10)$	$\sigma_k \sim Halfcauchy(10)$	$\sigma_k \sim Halfcauchy(10)$	$\sigma_{urb,j} \sim Halfcauchy(10)$
			$\beta_{urb,j:k} \sim N(0, \sigma_{j:k})$,	
			$\sigma_{j:k} \sim Halfcauchy(10)$	
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				$\beta_{bsc} \sim t(1, 0, 1)$
				$\beta_{bsc,k} \sim N(0, \sigma_{bsc,k})$,
				$\sigma_{bsc,k} \sim Halfnormal(0, 1)$
median income				$\beta_{inc} \sim N(0, 1)$
				$\beta_{inc,j} \sim N(0, \sigma_{inc,j})$,
				$\sigma_{inc,j} \sim Halfnormal(0, 1)$

Table 1: Prior summary table

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'.

The 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. 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).

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

6 Results

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

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 only consider graphical and summary output based diagnostics, namely: the MCMC trace plots as provided by BRMS, and the Effective Sample Size as provided by the summary output command.

For the first model, we see that all 4 chains are relatively horizontal, and each chain appears to be 'centred' around a particular value. There are no divergent transitions for any coefficient for this model.

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 predictive performance, then at relative and finally at the Leave-One-Out statistics to compare out-of-sample Predictive Performance.

To measure absolute predictive performance we used the Root Mean of Squared Error (RMSE) (FORMULA???). This 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, meaning it has a direct interpretation in the context of our problem. It takes into account the uncertainty of the posterior distribution by...

[plots of RMSE draws, overlap for comparison]

[interpretation]

To assess relative predictive performance, we looked at log-likelihood scores, that is, the average of posterior draws' log-likelihoods for each observation (FORMULA???).

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.

[plot of ll scores or likelihood differences?, overlap]

[interpretation]

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 tend to favor more complex models. (In our case, the bigger mmodel (Model 1) 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 assessing the likelihood (ELPD) of the observations in the latter, given the model estimates with the training set. ELPD FORMULA??????????????

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 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 (PSIS).

[LOO statistics table]

According to the LOO statistics, model 1 is the preferred one.

[pareto k estimates issue and momnet matching?]

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 Conclusion

11.1 Reflection on own learnings

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

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/