Final Project

WT 2024

Nimble/Stan dataset

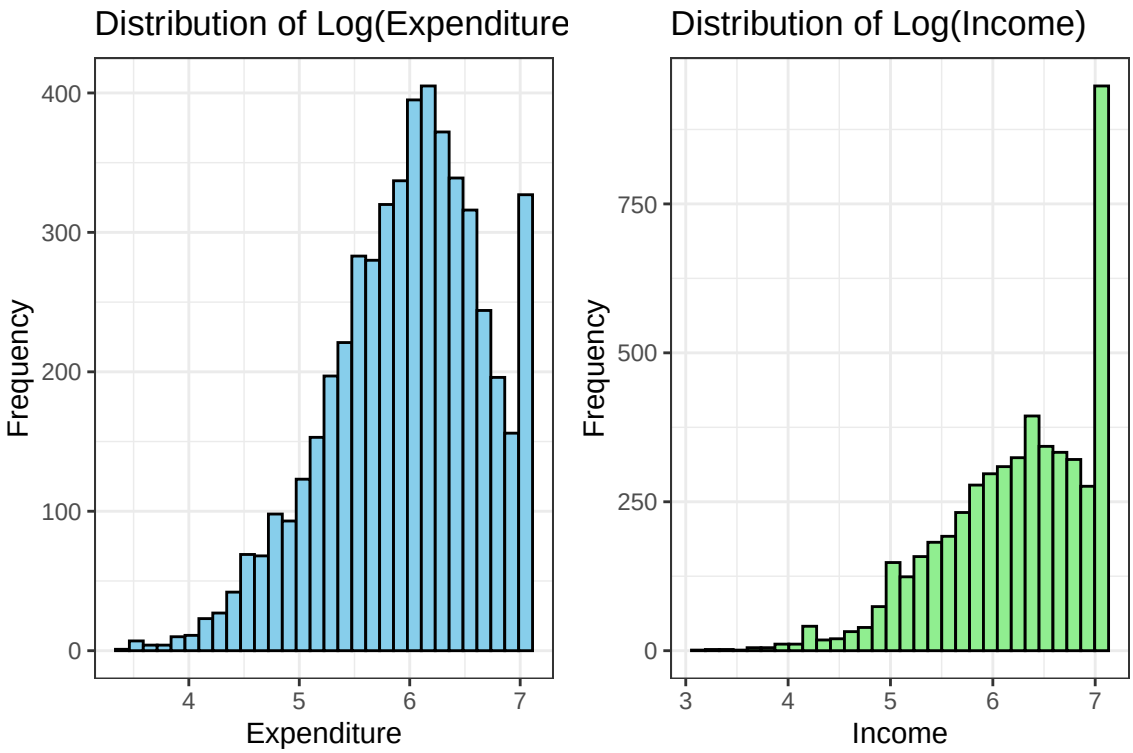
1. Introduction

This project aims to address the question of which factors impact the expenditure patterns of British households. Furthermore, it seeks to examine whether disparities exist in household expenditure across various regions, and if so, how regional factors influence the expenditure. To answer these questions, I trained Bayesian non-hierarchical models and Bayesian hierarchical models both with nimble and stan. After conducting feature selection and comparing model performance, it is concluded that the Bayesian non-hierarchical model trained with Stan, utilizing variables `A172`, `A094r`, `A121r`, and `income`, offers the most optimal solution for this case.

I utilized the 2013 Living Costs and Food Survey's educational dataset, where expenditure served as the outcome, and I incorporated 12 government office regions as random effects within the Bayesian hierarchical model. During data preprocessing, I logged both income and expenditure, and removed observations with a log(income) less than 2.5. Additionally, I excluded duplicated or irrelevant variables from the analysis, such as `weighta`, `P550tpr`, and `P344pr`. Unlike the prior project, I examined all variables and made feature selection based on "significance".

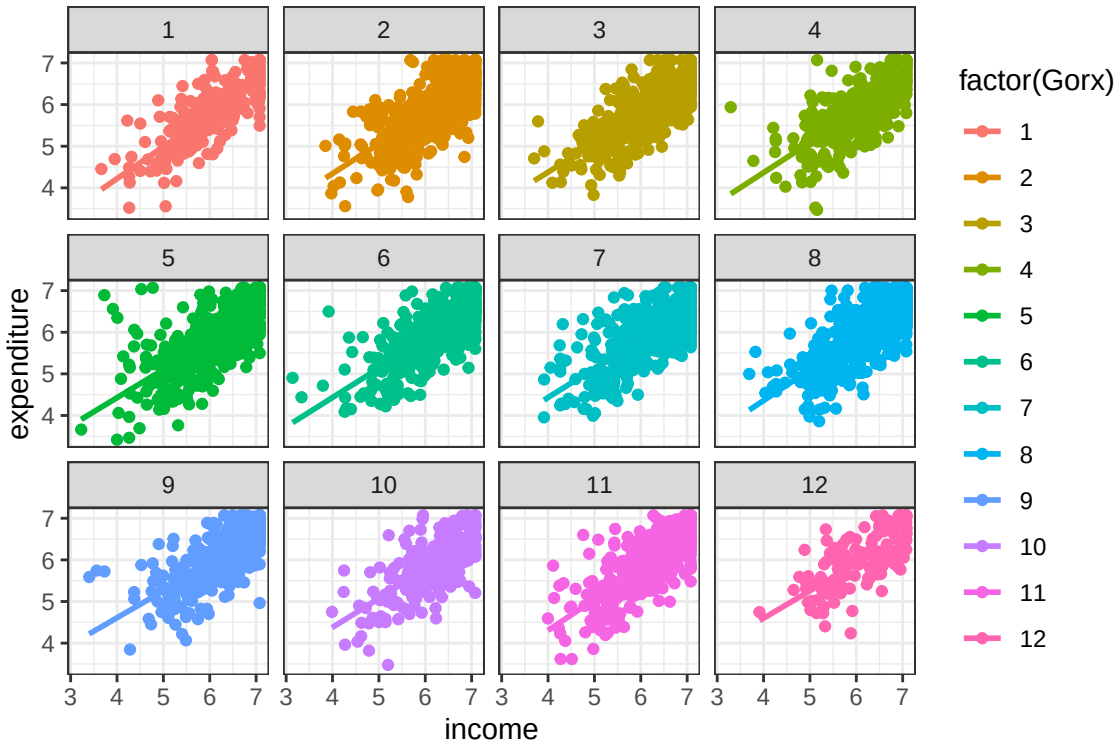
2. Summary of EDA

Region 8 has the highest income and expenditure while region 1 has the lowest. There are substantial discrepancies of expenditure across regions, which may impact the performance of non-hierarchical models. Notably, the distributiuon of both expenditure and income are skewed, with large samples clustered between 6 and 7. Thus, the prior distribution should be changed to compare model performance when training the model.



When visualizing the distribution of predictors and regions, it becomes evident that the influence of region on expenditure primarily affects the intercepts rather than the slopes. Hence, in Bayesian hierarchical models, the model featuring random effects only at the intercept may yield superior performance.

```
## `geom_smooth()` using formula = 'y ~ x'
```

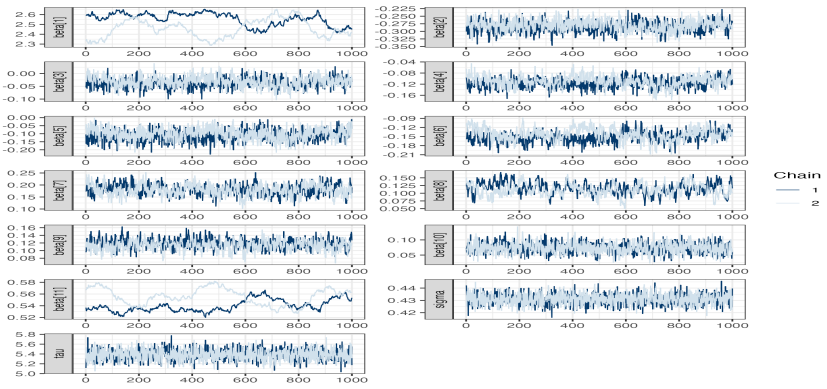


3. Analysis of non-hierarchical models

Frenquentist model: First, I fit the model in a frequentist regression approach. Only `A172`, `A094r`, `A121r`, and `A049r` variables are important with p values less than 0.05. This model explains 41.78% of the variance in expenditure.

Non-hierarchical model with nimble: First, I screened variables based on a significance level of 95%, resulting in 6 variables: `A172`, `A094r`, `A121r`, `G018r`, `G019r`, and `income`. They measure the internet connection status in the household, occupation category of the household reference person, residential status, number of adults in the household, number of children in the household, and household income.

Retraining Bayesian non-hierarchical model on these variables, I found that the mixing of `Intercept`, `A121r3`, and `Income` had the worst effect. Furthermore, their elevated rhat values and diminished neff values suggest significant chain variances and inadequate convergence. The covergence of these preditctors also affected the intercept, which acted as a dummy for the baseline level.



Given that the distribution of `expenditure` is skewed, with a significant concentration of samples around 7, I tried to change the outcome distribution to a lognormal distribution with heavier tails. However, the convergence of `income` and `A121r3` didn't improve.

Non-hierarchical model with stan: I also conducted variable screening for the Stan model, yielding similar variables to Nimble. With the exception of `G018r` and `G019r`, all other variables were retained, resulting in a total of 4 variables. In comparison to Nimble, the Bayesian non-hierarchical model trained by Stan exhibited significantly better convergence performance. The rhat values of all coefficients ranged between 1 and 1.01, indicating good convergence, and the neff values were larger generally. The acf plots show that the autocorrelation went away quite quickly within 3/4 lags, suggesting that the chains converge well within the current iterations. Notably, both `Income` and `A121r3`, which exhibited poor mixing effects in Nimble, appeared to have stationary distributions of both chains as shown in traceplots of beta[9] and beta[8]. And thus, the convergence performance of the `Intercept` was also enhanced. Regarding the model fitting, I computed the WAIC value of 6028.3, which could be used for the comparison later.

4. Analysis of hierarchical models

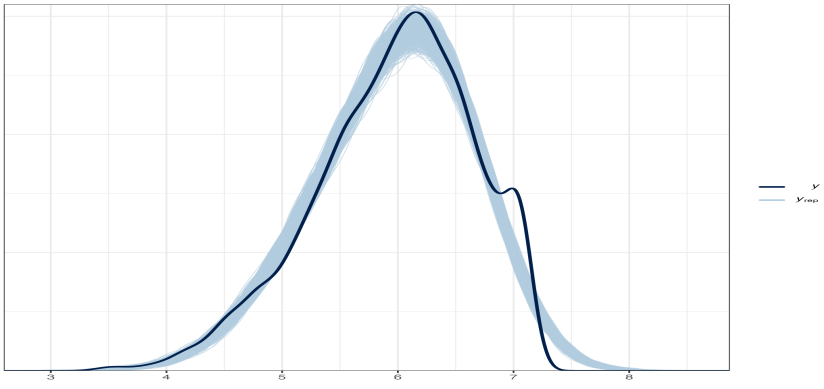
Hierarichical models in Nimble: The model trained with independent random effects on all coefficients in nimble, including intercepts and slopes, exhibits poor predictive power. Moreover, when comparing WAIC values between this model and the Bayesian non-hierarchical model trained by Nimble, the former yields a higher WAIC value of 57731.96 compared to 5925.279 for the latter. These results suggest a bad fitting effect of the fully hierarichical model.

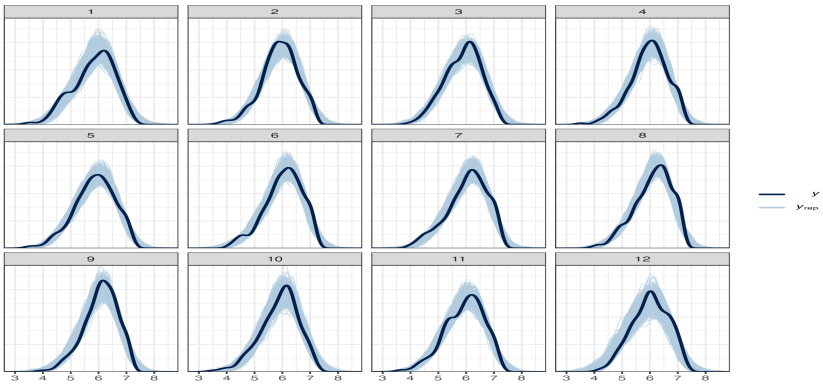
Then I tried to run a hierarichical model with random effect only on intercept. It achieved a WAIC value of 45771.92, slightly better than the fully hierarichical model. However, the acf plot shows that the chain was not moving around enough and it didn't converge well.

Hierarichical models in Stan: Given that the random effects of region primarily exert on the intercept, I trained 2 hierarchical models with selected features by using stan: **random intercept + fixed slope**, and **random slope of continuous predictor and fixed slope of categorical predictors**. In the

hierarchical model with random intercept, the negative random effects observed for regions 1-5, 7, and 10 suggest that households located in these regions tend to have lower expenditures compared to households in other regions, all else being equal. Notably, region 12 exhibits the largest absolute random effect, at 0.07, indicating that households residing in this region are estimated to spend approximately 7.25% more than identical households without region-specific information. In general, there is little difference among the random effects across regions in both models, indicating that the regional information is limited.

Compared to the models in nimble, they both converged much better, converging after approximately 4600 iterations. They performed comparable in terms of fitting, with WAIC values around 6004. In terms of prediction, both the non-hierarchical and hierarchical models in Stan demonstrate similar performance as shown below. However, they fail to accurately predict the turning points, particularly around the expenditure value of 7. Notably, the small double peaks observed in regions 1, 3, and 11 are not captured effectively in the predicted posterior distribution.





5. Discussion

After careful consideration of convergence, fitting, and prediction performance, as well as the interpretability and complexity of the models, the Bayesian non-hierarchical model trained in Stan emerges as the optimal choice for this case. Although the hierarchical model marginally outperforms the non-hierarchical model in terms of fitting, the negligible difference in prediction and convergence suggests that the addition of regional random effects may not be justified. This implies that the explanatory contribution of areas on household expenditure is relatively minor compared to other variables. Consequently, the final model is formulated as follows: `expenditure = 2.39 - 0.3*A1722 - 0.33*A094r2 - 0.1*A094r3 - 0.11*A094r4 - 0.15*A094r5 + 0.17*A121r2 + 0.11*A121r3 + 0.58*income`.

All coefficients in the model represent fixed effects. Analysis of the coefficients reveals insightful patterns. For instance, the coefficients associated with variable `A094r` are consistently negative, indicating that compared to households with reference personnel in higher managerial, administrative, and professional occupations (`A094r1`), households with reference personnel in other occupation categories tend to have lower expenditures. Specifically, if all other variables remain constant, households with reference personnel in intermediate occupations (`A094r2`) are estimated to spend approximately 71.89% of households of `A094r1`. Similarly, holding all other variables constant, households without internet connection (`A1722`) are estimated to spend approximately 74.08% of the households with internet connection (`A1721`). On the other hand, the positive coefficient associated with variable `A121r` suggests that households in private rental or owner-occupied tenure spend more compared to households in publicly rented tenure (`A121r1`). Furthermore, households with higher incomes tend to have higher expenditures.

It's clear that there's still room for improvement in the current model. While the impact of regions on the model may be limited, it doesn't necessarily imply that there are no differences in household expenditure between regions. One potential avenue for improvement could involve refining the regional hierarchy, such as focusing on household spending within specific London boroughs. Narrowing down the analysis to smaller geographic areas could reveal more nuanced and intriguing insights into household expenditure patterns, potentially uncovering subtle differences that were previously overlooked and improving model's prediction

on the turning points of each region.

6. Summary

In summary, we can attribute differences in household spending to factors such as income level, internet connectivity, occupation of the household reference person, and tenure status. At the current regional level, there isn't a significant disparity or correlation in household expenditure across regions. However, households with higher incomes, owning their residencies, having internet access, and being engaged in professional or managerial occupations tend to exhibit higher expenditures. Among these factors, household income exerts the most significant influence on spending behavior. Holding other variables constant, a unit increase in income corresponds to approximately 78.6% growth in spending.

Spatial dataset

1. Introduction

This project aims to understand the changing pattern of Covid-19 deaths in the United States spanning 2020 Febuary to 2021 September. It seeks to address problems regarding the variations in deaths across different regions (states/counties) during distinct timeframes and the influence of demographic factors on these fluctuations. This study incorporates data source from multiple channels, including poverty ratio `PropPov`, white population ratio `PropWhite`, male ratio `PropMen` as independent variables and population at county level `TOT_POP` as offset variable. The monthly death is utilized as the dependent variable for analysis.

Initially, I build spatio-temporal models to analyze the death toll in Alabama, incorporating both models without space-time interaction and those with interaction terms. Then, I explore the fluctuations in Covid-19 deaths within New York State. Given its distinct political-economic landscape, higher population density, and an earlier onset of the outbreak, I compare these trends with those observed in Alabama. I aim to study the primary factors influencing death counts across different states, considering the significance of demographic attributes in this context.

```
## Joining with `by = join_by(FIPS.Code)`
## `summarise()` has grouped output by 'Month.num', 'Month', 'County.Name'. You
## can override using the `.groups` argument.

## Joining with `by = join_by(NAME)`
```

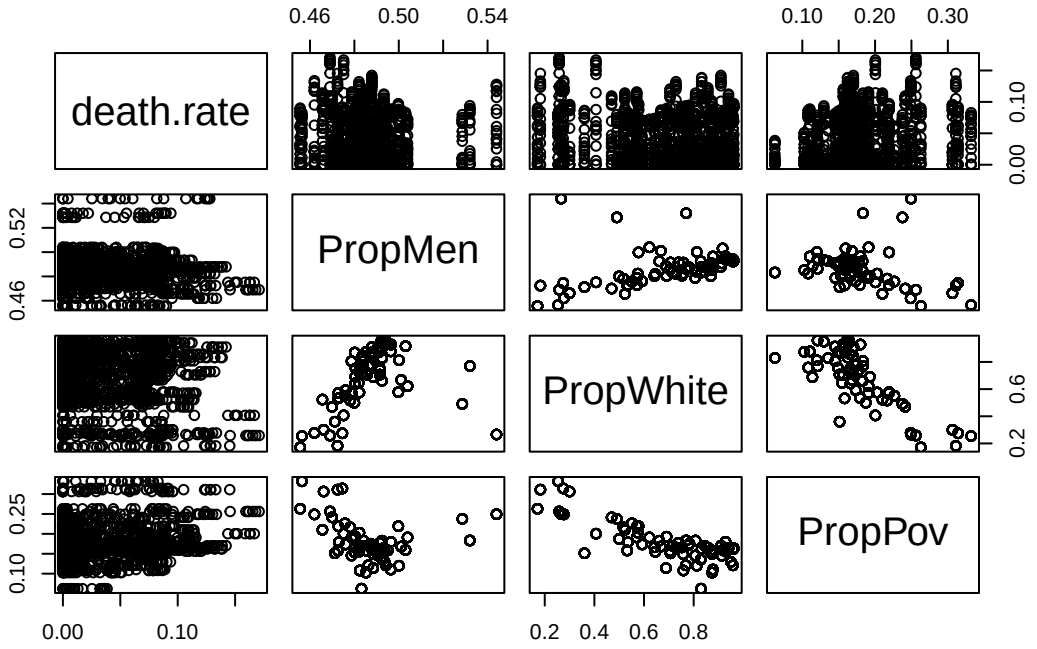
2. Exploratory Data Analysis (Alabama)

We observe a consistent increase in the total number of deaths statewide, indicative of the persistent impact of the outbreak. The surge in deaths happened in the 10th-14th month period. Although the number of deaths remained stable after the 15th month, it stayed at a high level of more than 300,000, suggesting that the spread of Covid-19 has not been effectively controlled during the observation period.

Jefferson and Mobile counties, being among the most populous, record significantly higher death counts. The mortality rate (deaths/total population) paints a nuanced picture, counties with smaller population exhibit elevated mortality rates.

The scatter plot shows that there is correlation between the male ratio, white population ratio, and poverty ratio, which implies that demographic characteristics intersect with socioeconomic status. For example, certain demographic groups, such as white males are less likely to live at a poverty level and they may face larger or smaller challenges of Covid-19.

In addition, the distribution of monthly deaths is over-dispersed with a large variance, so a Negative binomial distribution will be used for modeling.



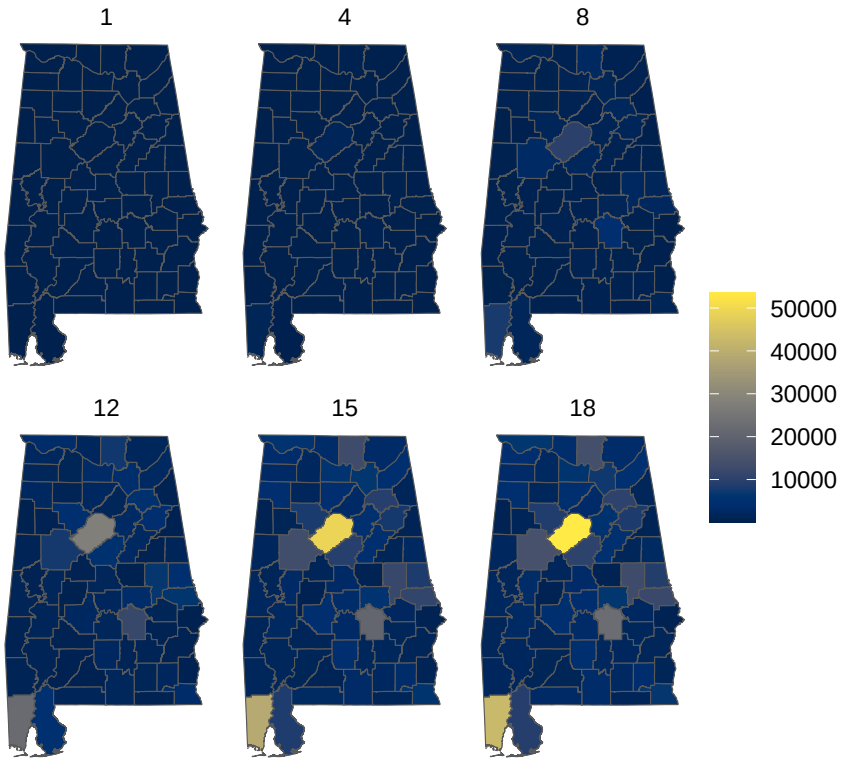
3. Spatio-Temporal Models

3.1 Spatio-Temporal Models without Interaction In the full model incorporating three explanatory variables, I found that the coefficients of `PopWhite` and `PopMen` were not significant at the level of 95%, so only `PropPov` was retained as the explanatory variable. This suggests that `PropPov` potentially picks up the explanatory contribution of `PopWhite` and `PopMen` to the outcome, and thus we don't need to consider interactions between these variables further.

The plot of time component shows the effect of time component on the outcome in the original scale. The "rw2" line exhibits the most pronounced fluctuations and displays an overall upward trend, mirroring the previous EDA time trend chart for statewide deaths. This suggests that the impact of time on monthly deaths increases over time, with notable spikes occurring around the 10th month. Conversely, the "iid" and "lin" lines remain stable around 1, indicating minimal impact on the number of deaths. This could be attributed to the dominance of the random walk component in the RW2 model, which captures most of the time effect. Similarly, in the linear model, the linear fixed effect of time plays a primary explanatory role.

Given the absence of periodicity in the model with months as the unit, I change the time unit from months to weeks later to discern subtle differences in time changes. This approach aims to investigate whether any seasonal or cyclical variations manifest differently at the weekly level.

In the fitted value plot, it's evident that temporal changes had the most significant impact on Jefferson County and Mobile County. The number of deaths in these two counties exhibited a notable increase starting around the 8th month, marking them as early hotspots within Alabama to experience a surge in fatalities. As time progressed, the epidemic expanded, leading to a gradual rise in deaths in the surrounding counties of Jefferson. Positioned in the southwest corner of Alabama, Mobile County has fewer neighboring counties, resulting in a more localized impact compared to Jefferson. This highlights the substantial influence of both time and space on the deaths in Alabama. Particularly, the evolving patterns of fatalities in the counties surrounding Jefferson County highlight the interaction between time and space with discernible neighborhood structure and time lag.




```
## Error in inla.inlaprogram.has.crashed() :
##   The inla-program exited with an error. Unless you interupted it yourself, please rerun with verbos
##   If this does not help, please contact the developers at <help@r-inla.org>.
##
##  *** inla.core.safe:  inla.program has crashed: rerun to get better initial values. try=1/2
##
##  *** inla.core.safe:  rerun with improved initial values
```

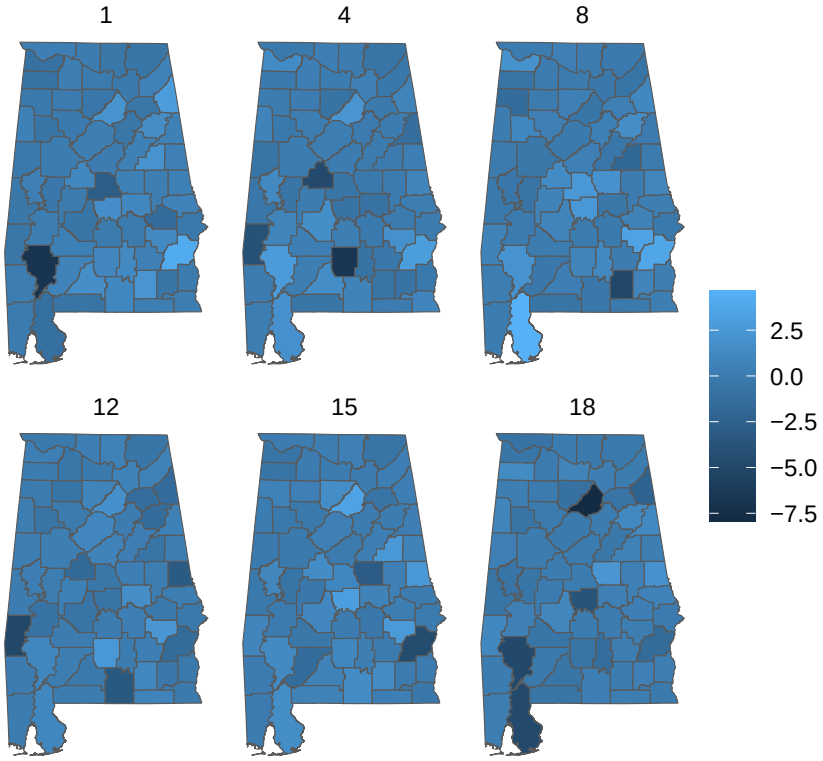
3.2 Spatio-Temporal Models with Interaction Due to the evident structural spatio-temporal interaction observed in the Covid-19 deaths in Alabama, I opted for the Type4 spatio-temporal interaction model, and the interaction term is `structured spatial component (besage) * structured temporal component (RW2)`. Notably, the Type4 model demonstrated superior fitting performance compared to other spatio-temporal models with the smallest WAIC value. Particularly, its fitting efficacy significantly surpassed that of the Type1 model, which featured an interaction term of `iid`.

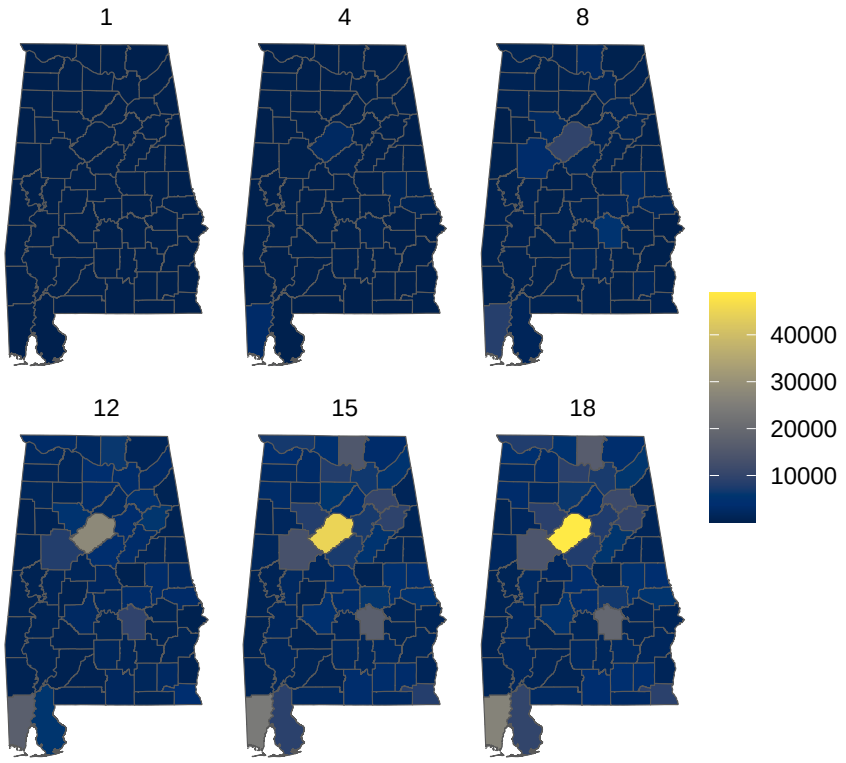
The figure of fitted values illustrates the disparities in fitted values among different models. While the contrast between the Type1 model and the non-interactive model is marginal, with a difference of less than 10, a substantial gap emerges between Type1 and Type4 models. This discrepancy highlights that the significant variance between interactive and non-interactive models predominantly originates from the inherent structured interaction mode within the interactive components.

The following chart illustrates the impact of interaction terms on deaths across regions at different time points. A notable observation is the fluctuation observed in Baldwin County, which shares a significant border with Mobile County. Around the 8th month, the interaction between spatial adjacency and time had a discernible influence on the death toll in Baldwin County, which might lead to an increase in fatalities afterwards (as shown in the fitted plot of time). Around the 18th month, the impact of the interaction term on Baldwin County diminished. This could be attributed to the stabilization of the epidemic's spread, leading to a weakening effect of space-time interaction. Other factors began to play a more prominent role in monthly deaths in this area.

Notably, counties surrounding Jefferson, such as Tuscaloosa and Shelby, exhibit stable patterns under the influence of spatio-temporal interaction terms, with the log of interaction term hovering around 0. This, coupled with the sustained high number of outbreak-related deaths in the region, suggests ongoing epidemic impact throughout the observation period. This underscores the interconnected nature of these regions, where Covid-19 continues to exert its influence intersectingly.

```
## Joining with `by = join_by(Month.num, num_id)`
```





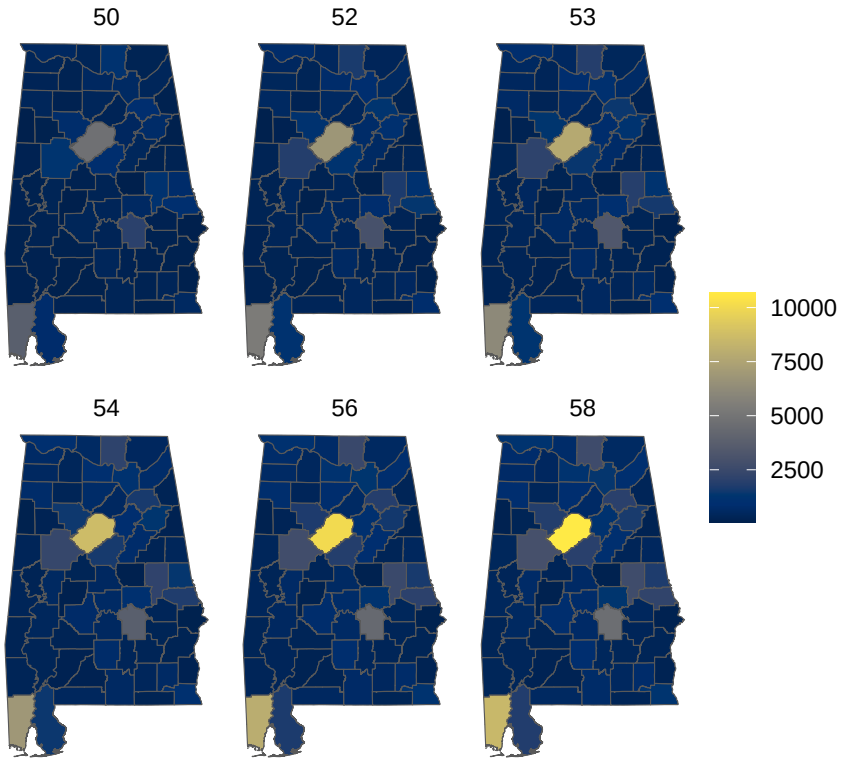
4. Spatio-Temporal Models with Week as Time Unit

After transitioning from monthly to weekly time units, no cyclical or seasonal changes were observed. The effects of the `iid`, `lin`, and `temp` terms on the number of deaths remained consistent with those in the monthly model. Interestingly, while in the monthly model, the impact of time on the outcome declined towards the end of the observation period, in the weekly model, this impact continued to strengthen significantly. This discrepancy could be attributed to the smaller time units of weeks, rendering the model more sensitive to short-term fluctuations. Consequently, the model is able to detect abrupt increases or decreases occurring towards the conclusion of the observation period.

```
## `summarise()` has grouped output by 'Week.num', 'Week', 'County.Name'. You can
## override using the `.groups` argument.
```

```
## Joining with `by = join_by(NAME)`
```

The figure below provides a breakdown of how the outbreak of deaths unfolded during the 50th-60th week period, offering a nuanced pattern to comprehend the fluctuations in the number of deaths.



5. Discussion

There are some limitations regarding the current analysis.

- Firstly, the observation period is restricted, and the chosen time points lack background information, particularly at the endpoint. Consequently, the analysis fails to capture a full cycle of the epidemic's spread, development, and mitigation efforts. This results in a simplistic portrayal of the role of time in the analysis. Addressing this limitation would require extending the observation period and ensuring that the selected time nodes are well-grounded in relevant contextual information.
- Secondly, in areas characterized by low economic levels and population density, additional factors such as accessibility to healthcare services and the effectiveness of epidemic detection measures should be considered. It's plausible that there could be underreporting or underdetection of epidemic-related deaths due to limited access to medical services or deficiencies in detection mechanisms. This could introduce bias into the results.

6. Covid-19 Deaths in New York State

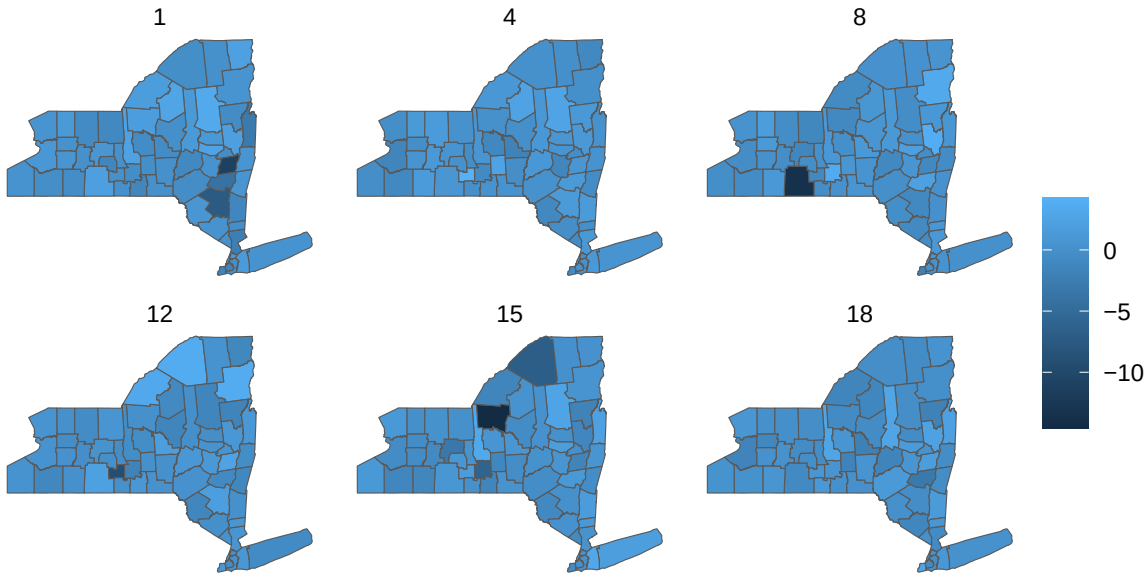
In contrast to Alabama, New York showed no significant correlation between the male ratio, white ratio, and poverty ratio variables. Deaths were predominantly concentrated in highly urbanized, densely populated areas such as New York City, Queens, and Kings County. Given the over-dispersion nature of monthly deaths in New York State, negative binomial regression was employed for modeling purposes. In the full model incorporating 3 predictors, none of the three variables reach significance, and thus they were not included in the spatio-temporal models. This suggests that time and space may play a more significant role in understanding the changing patterns of Covid-19 deaths compared to other demographic factors, especially in metropolitan areas with a greater racial and cultural diversity where death tolls are elevated.

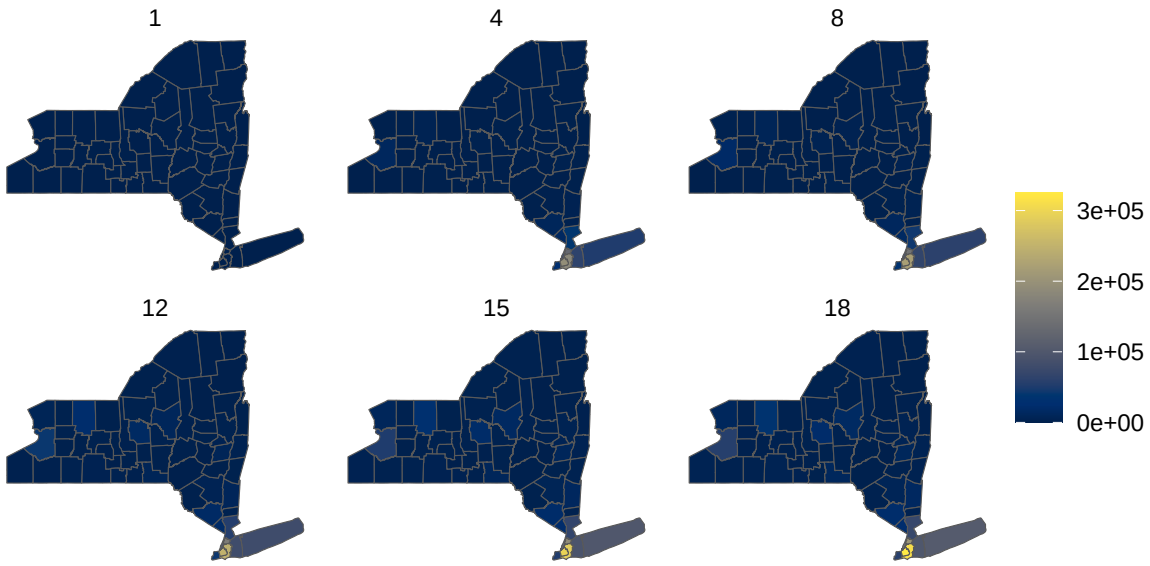
```
## Joining with ‘by = join_by(FIPS.Code)‘
## ‘summarise()‘ has grouped output by ’Month.num’, ’Month’, ’County.Name’. You
## can override using the ‘.groups‘ argument.
```

```
## Joining with ‘by = join_by(NAME)‘
```

Though the Type4 spatio-temporal model exhibits the best fitting performance with the smallest WAIC value, in general, the impact of the interaction term remained consistent and moderate. Except for a few counties at specific time points, the effect of the interaction term in other regions exhibited minimal variation throughout the observation period. When examining the fitted value maps, it becomes apparent that most of the central region consistently reported relatively low death counts, maintaining a stable level over time. Conversely, the area surrounding New York City continued to grapple with the epidemic's spread in New York, witnessing a continual surge in the number of deaths over time. Nevertheless, the impact remained localized, primarily affecting the southeastern corner of the state. This may be because the population distribution of New York State is concentrated, resulting in less significant changes in spatio-temporal interaction terms between different regions.

```
## Joining with ‘by = join_by(Month.num, num_id)‘
```





Although the three demographic characteristics are insufficient to explain the number of deaths in New York State, there are other factors that need to be taken into account. For instance, the age structure of the population in the region may play a significant role, given that older individuals generally face a higher risk of mortality from Covid-19. Moreover, although the Type4 model exhibits the highest fit with the highest WAIC value in this case, the discrepancy in WAIC between the Type4 and Type2 models is not substantial. Thus, there exists a trade-off between model complexity and interpretability. Further research is required to ascertain if interaction terms are essential and if they can be selectively integrated into specific regions, such as the southeast corner of New York State instead of the whole state.

7. Summary

In summary, the factors affecting the death counts differed between Alabama and New York state, and distinct spatio-temporal patterns were observed in each state. In Alabama, there was a discernible structural feature in the spatio-temporal interaction term and space, with the epidemic progressively spreading over time from Jefferson County to neighboring areas. In New York State, deaths were primarily concentrated in the metropolitan area, displaying pronounced geographic characteristics and showing less susceptibility to other demographic factors.

Appendix: