Bayesian Analysis - Learning about Winners of the Eurovision Song Contest and Investigation of Viewership Trends

Henrik Andreas Grenersen

May and June 2024

Abstract

In this project, I will consider data on the Eurovision song contest. Attempts will be made to estimate the probabilities of different countries winning the countries, both with and without covariates. Here, the models are mainly implemented in STAN, with some comparisons made to standard functions in R. Viewership data for previous years in selected countries is used to learn about the overall variation in viewership rates, through a Bayesian approach. Lastly, simulations are carried out in order to leverage spatial relationships to identify countries that have unique viewership rates compared to their neighbours. This was originally done to assess the effect of potential boycotts in the different countries, but this attempt was unsuccessful, mainly due to a lack of incorporation about prior knowledge about the countries viewership rates.

Contents

1 Introduction

1.1 Problem Description and Choice of Problem

This project has two, somewhat separate, objectives, that both deal with the Eurovision Song Contest (ESC). As you might be aware, this is a song contest, where the participating countries, mainly European ones, send a participant to the country of the previous year's winner in order to compete for the title. Viewers from all participating countries are then able to vote for their favourite artist, as long as they are not from the same country. The points from the viewers are then combined with the points from the juries in each country, and the artist that receives the most points is crowned as the winner.

Winning the competition of course brings fame and glory to the artist, but it also means that the broadcasting union of the winning country will have to host the next edition of the competition. As this is rather costly, some unions have been known to send contributions that are far from likely to win. The contributions that make an honest attempt at winning are however far more interesting.

Another interesting question in this context is what it is that characterises the winning contributions in the competition. To identify this will be the first objective of this project, and might hopefully provide some insight such that my country, Norway, does not end up in last place again, as we did this year.

Although the ESC is normally a great celebration, this year's edition has been quite the opposite in the eyes of many, due to the horrible situation in Gaza. Many have called for an exclusion of Israel from the competition, but their attempts were unsuccessful and Israel were allowed to compete by the hosting organisation, the European Broadcasting Union (EBU). In Norway, and other European countries, this lead to boycotts, both by establishments such as bars that usually shows the competition, but also from individuals that usually watch from home.

Much can be said about boycotts as a form of demonstration, but in order for a boycott to have the desired effect, it seems reasonable that enough people need to participate in order to catch the attention of those in power. To quantify how much viewership numbers, here given by the number of people that watched the competition divided by the population in the country, have to change in order for the boycott to have been significant seems like a more complex task. The second objective of this project deals with this theme, and I will compare viewership data from previous years for some of the participating countries in order to explore if it is possible to quantify the effects of the boycotts.

So, to summarise this section, this project has two main objectives, the first being learning more about the characteristics of ESC winners, so that I can hopefully showcase incredible prediction skills to my friends before the competition in Switzerland in 2025. The second objective on the other hand, has a more serious character, motivated by the horrible backdrop the contest was held against this year.

1.2 Scope of the Project

Regarding the first objective, I have restricted the scope to consider two methods, the first being modelling the probability of a win for each country that has ever won, by only using the number of times each country has won. This is done by considering the result of each edition of the competition as a multinomial random variable, with the possible outcomes being the countries that have won the ESC at least once.

Because of potential problems with the implemented model in STAN for the above approach, I have also considered the same data as independent binomial variables. In this approach, we consider one random variable for each country, all having 66 trials, corresponding to number of times the ESC has been held and only had one winner, i.e. except 1969 and 2020, but where each variable has its own success probability. Here, the success probability is also the parameter of interest.

Lastly, I model the success probability through a logistic regression, using the same approach as in the last paragraph. Now however, other covariates as the language of the song and the gender of the singer are included.

For the second objective, regarding the viewership data, the scope has been restricted to a Besag model, described in Section **??**, for the proportion of viewers in a few selected European countries. The reason for only using a subset here is that the model assumes that we have a connected graph, i.e. all countries that are included share borders, but also the fact that I have only found data for

these countries. Here, I will use a Bayesian model to learn about the precision parameter τ, which is a measure of the precision in the response in the different regions, given the realised value in one or more other regions.

Finally, by using the viewership data for this year's competition, I will calculate the posterior predictive probabilities of observing the given viewership rate, or something lower, in all other countries. The simulations for a given country are done conditionally on knowing the rates in all other countries, and the reason for this technicality will also be explained later. Through doing this, I aim to be able to say something about where it is easiest to identify countries where effective boycotts have taken place.

This section can be summarised by the following list of the models that will be considered.

1. To learn more about the characteristics of ESC winners
   (a) Multinomial model for estimating the probability of winning for all countries that have won the ESC.
   (b) Binomial model for estimating the probability of winning for all countries that have won the ESC.
   (c) Logistic regression for the probability that a given song wins the ESC.
2. To quantify the effects of boycotts of the competition in 2024
   (a) Bayesian Besag model for the precision parameter τ.
   (b) Using the posterior for τ to calculate posterior predictive probabilities for the viewership rate in all countries.

1.3 Response and Explanatory Variables

In the context of learning more about the characteristics of ESC winners, we have three different scenarios, each with their own response and explanatory variables. In the setting of a multinomial response, we have that the response is distributed as follows

$$Y_i|\vec{\theta} \sim \text{Multinomial}(\vec{\theta}) \quad i \in \{1, 2, ..., 66\} \tag{1}$$

where the parameter vector contains the probability of winning for each of the countries that have ever won the ESC, and has length 27. Exactly which countries these are will be presented in the next section, but the other countries have been excluded to reduce the number of potential categories for the response. There are no explanatory variables in this approach.

The multinomial approach has an advantage in that it accounts for the fact that only one country can win at a time, which is something that is not automatically encoded in the following binomial model. I have however also implemented the following binomial model. Here, the response is given per country, not competition, and is distributed as

$$Y_j|\theta_j \sim \text{Binomial}(66, \theta_j), \quad \theta_j|a, b \sim \text{Beta}(a, b), \quad a, b \sim \text{Gamma} \quad j \in \{1, ..., 27\} \tag{2}$$

since this is a hierarchical model.

The last model regarding this objective is a Bayesian Generalized Linear Model (GLM). Here, the response is either 0 (did not win) or 1 (did win), and here I will consider the data on the winners that have previously been used. However, I will also include data on songs from 2019, 2021, 2022 and 2023 that did not win, as I will then also have some responses that are 0. Then, the response for song number k is

$$Y_k \sim \text{Binomial}(1, \theta_k) \tag{3}$$

where the success probability is assumed to obey the following relationship

$$\text{logit}(\theta_k) = \vec{x}_k^T \vec{\beta}$$

where it is needed to assume priors on the regression coefficients $\vec{\beta}$.

Lastly, the vector $\vec{x}_k$ is a vector of covariates for song number k, and the covariates I will consider are:

1. Gender of artist: Female, Male or Group.
2. Starting Position: Given as a number between 0 and 1, which is the starting position divided by the total number of participants, which has varied from year to year.
3. Language of the song: English or Non-English.

Then, for the objective regarding viewership data, the response in each country is the number of viewers in each country of the ESC, divided by the country's population in 2024, for the years 2018, 2019, 2021, 2022 and 2023. As far as I am aware, there have not been any drastic changes in the populations of any of the countries considered during the period, and it should be reasonable to divide by the population in 2024. However, as these rates are constrained to the interval $[0, 1]$, and the model used assumes a Gaussian response, I have transformed the rates to the interval $(-\infty, \infty)$ through the logit transformation, letting $r_{i,j}$ denote the viewership rate in country i in year j

$$y_{ij} = \text{logit}(r_{ij}) = \log\left(\frac{r_{ij}}{1 - r_{ij}}\right).$$

There are no explanatory variables in this model, but the graph structure used in the model is graph given in Figure 1.

1.4 Presentation of Data

The first edition of the ESC was held in 1956, and the competition has been held yearly ever since, except for in 2020, when it was cancelled due to covid. The year 2020 will therefore not be included in the data used in this project, and neither will the year 1969, as the four countries Spain, the UK, the Netherlands and France, all won that year in some incredible way. The year 2024 has been held out for checking models, as the competition had not yet been held when I started working with this project.

A histogram showing the number of times countries have won is given in Figure 2. Here we see that Ireland and Sweden are in the lead with the most wins, and it should also be noted that among the five most winning countries, i.e. Ireland, Sweden, Luxembourg, the UK and the Netherlands, only Sweden has won in the last decade. So, there might be a trend in more recent years that no particular country dominates the competition, like Ireland which has previously won the competition three years in a row. We also note again that there are very many categories compared to the number of data points we have.

The data depicted above is also the foundation for the binomial model, whereas for the logistic regression, we also consider some covariates. An excerpt of some of this data is given in Figure 3. It should also be noted that I have augmented the data here by also including data for songs that did not win the competition, from the years 2023 - 2019. excluding 2020. This data is examined more closely in section 2.1.1.

Figure 1: Graph depicting the neighbourhood structure of the subset of European countries included in the analysis

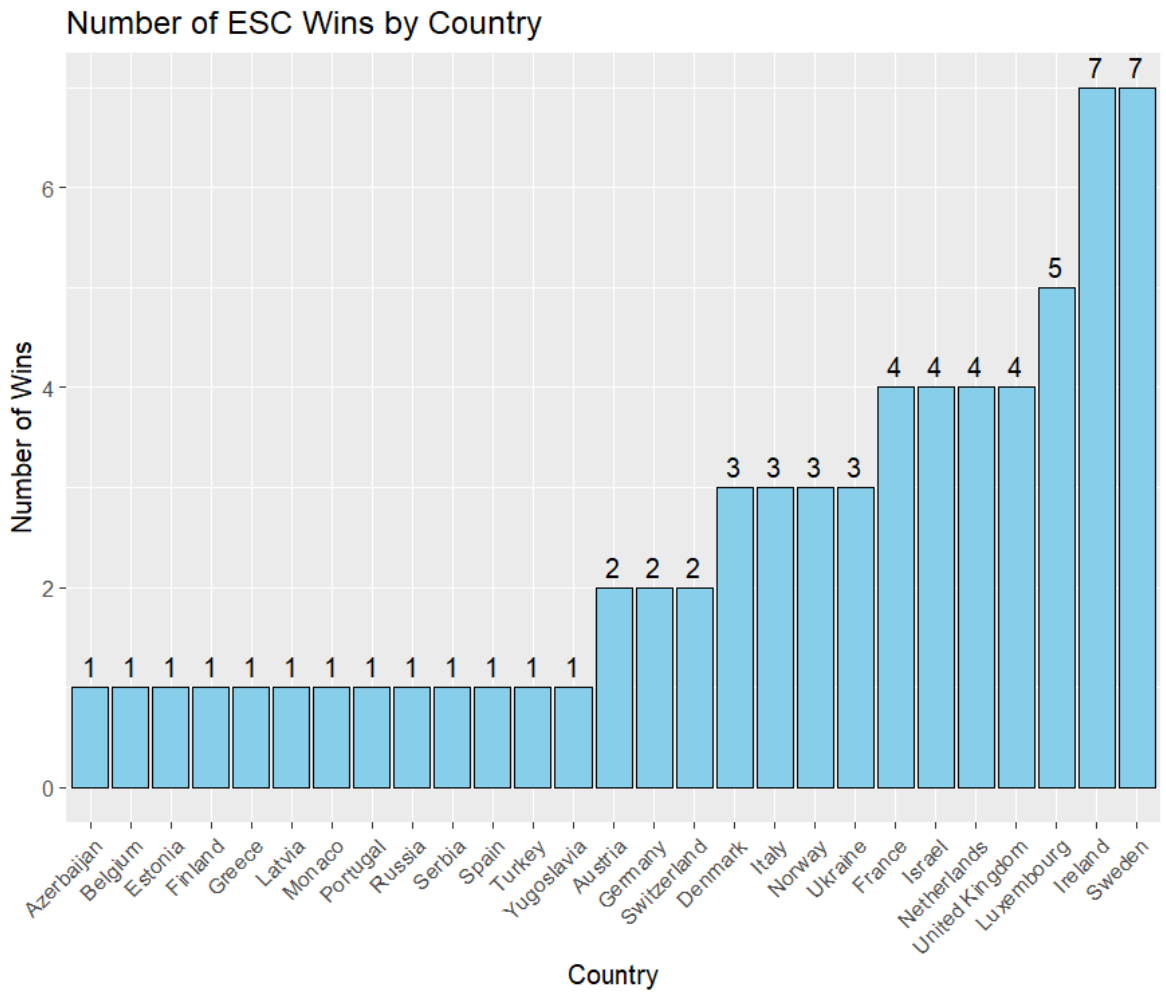


Figure 2: Only countries that have actually won the ESC are included

Winner	Singer's gender	Language	Starting position	Number of competing teams	Starting in %
Sweden	F	E	9	26	34,62
Ukraine	G	N	12	25	48,00
Italy	G	N	24	26	92,31
CANCELLED					
Netherlan	M	E	12	26	46,15
Israel	F	E	22	26	84,62
Portugal	M	N	11	26	42,31
Ukraine	F	N	21	26	80,77
Sweden	M	E	10	27	37,04
Austria	O	E	11	26	42,31
Denmark	F	E	18	26	69,23

Figure 3: Example data for the logistic regression.

2 Analysis

As was explained in the Introduction section, this project has two rather different objectives, although they both deal with the ESC. In this section, I have therefore chosen to separate the analysis regarding the different objectives into two sections, in which I will perform exploratory data analysis and also Bayesian analyses.

2.1 Objective 1 - Learning About the Probability of Winning the ESC

2.1.1 Exploratory Data Analysis

For the multinomial and binomial models for the probability of winning for different countries, all the relevant data is presented in Figure 2 as mentioned in the introduction.

Regarding the logistic regression that will be performed, a pairs plot of the data is presented in Figure 4. From the plots along the diagonal we can learn more about the distribution of the covariates and the response. We can for instance see that the most frequent gender amongst the participants is female, the second most is male, and the last is participants that were groups. Additionally, we see that there are more songs sung in English than not sung in English, and there are more observations of songs that did not win than songs that did win. From the plot in the lower left corner, i.e. the plot of gender against number of wins, we see that among the winners, which is the lower of the two plots, there are very many more female participants than participants from the other groups. Thus, it seems likely that women have a higher chance of winning the ESC.

It seems hard to see any clear relationships between the starting position, given by the covariate "prop". For the response however, we see that the median starting position, given by the horizontal line in the box plot, is higher for the winners than the losers. Thus, a possible hypothesis could be that having a later starting position in the competition, i.e. a greater value of "prop", increases the chances of winning the competition. Lastly, from the plot of language against the response, we can see from the lower plot that the frequency of the two groups is fairly similar between winners. Therefore, we might expect a regression coefficient for this covariate that is close to zero.

2.1.2 Bayesian Analysis

In order to model the data as described in equation 1, I have implemented the model given in Appendix 5.1 in STAN. This has lead to samples from the posterior of the vector $\vec{\theta}$, given in Figure 5.

If we first consider the credible intervals, we see that the model has created identical intervals for countries that have the same number of wins, which seems logical as we have no additional information. Another interesting feature that can be observed in the plot is that the width of the intervals seems to increase as the number of wins increases. Thus, there is a larger uncertainty associated with the estimates of the probabilities of Sweden or Ireland winning the competition, than that of Turkey or Spain.

In Figure 5, the red lines denote the maximum likelihood estimates MLEs, for the different countries, depending on their number of wins. This estimate is for country i given as

$$\hat{\theta}_i = \frac{y_i}{n}.$$

The reason for this is that the likelihood is given as

$$\mathcal{L}(\vec{\theta}) = p(\vec{y}|\vec{\theta}) \propto \prod_{i=1}^{n} \theta_i^{y_i}$$

Which gives the following loglikelihood

$$l(\vec{\theta}) = \sum_{i=1}^{n} y_i \log(\theta_i)$$

Now, before maximizing the loglikelihood as above, we must remember that the parameter vector always lives under the constraint that its component must sum to 1. We can respect this constraint

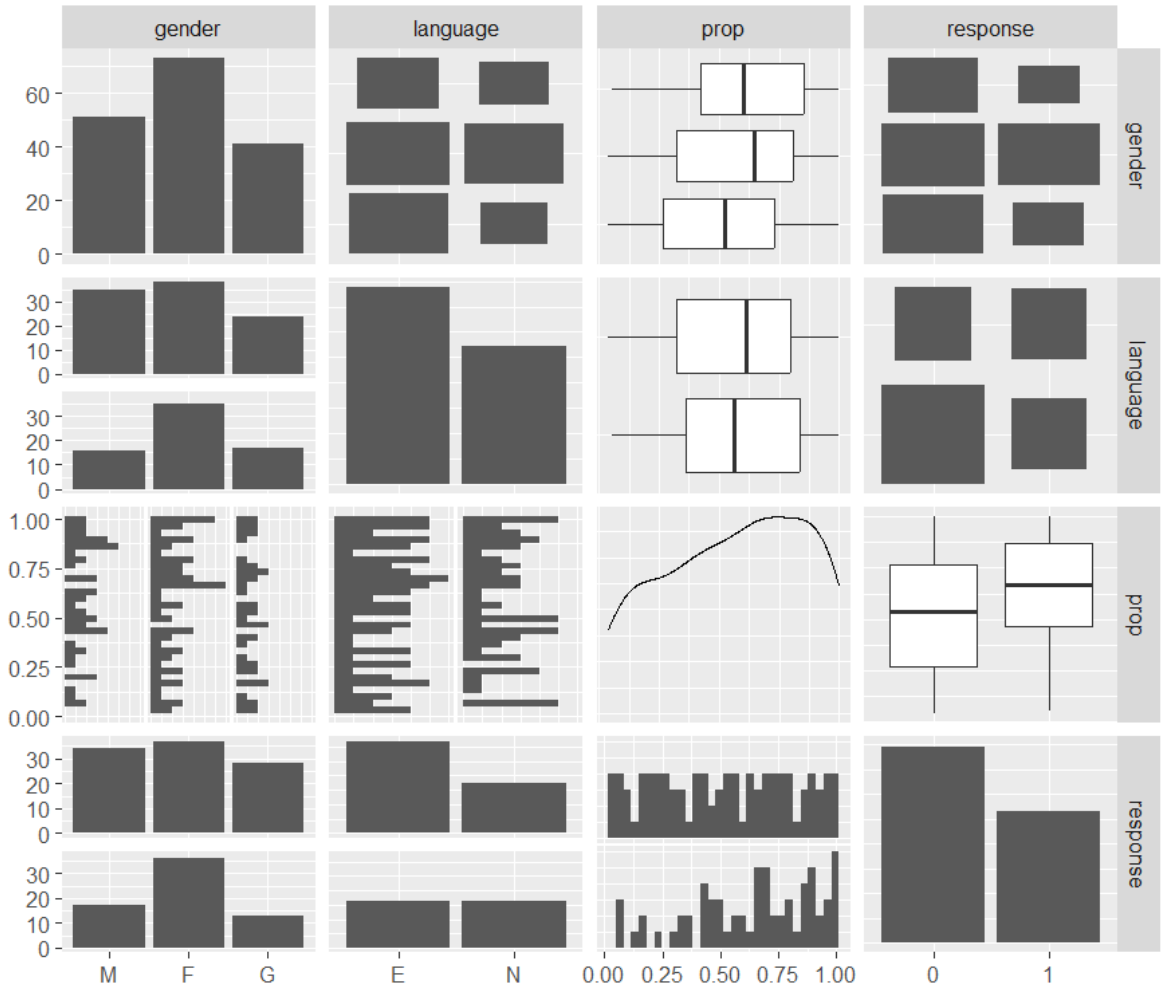


Figure 4: Pairs plot of response and covariates for the logistic regression.

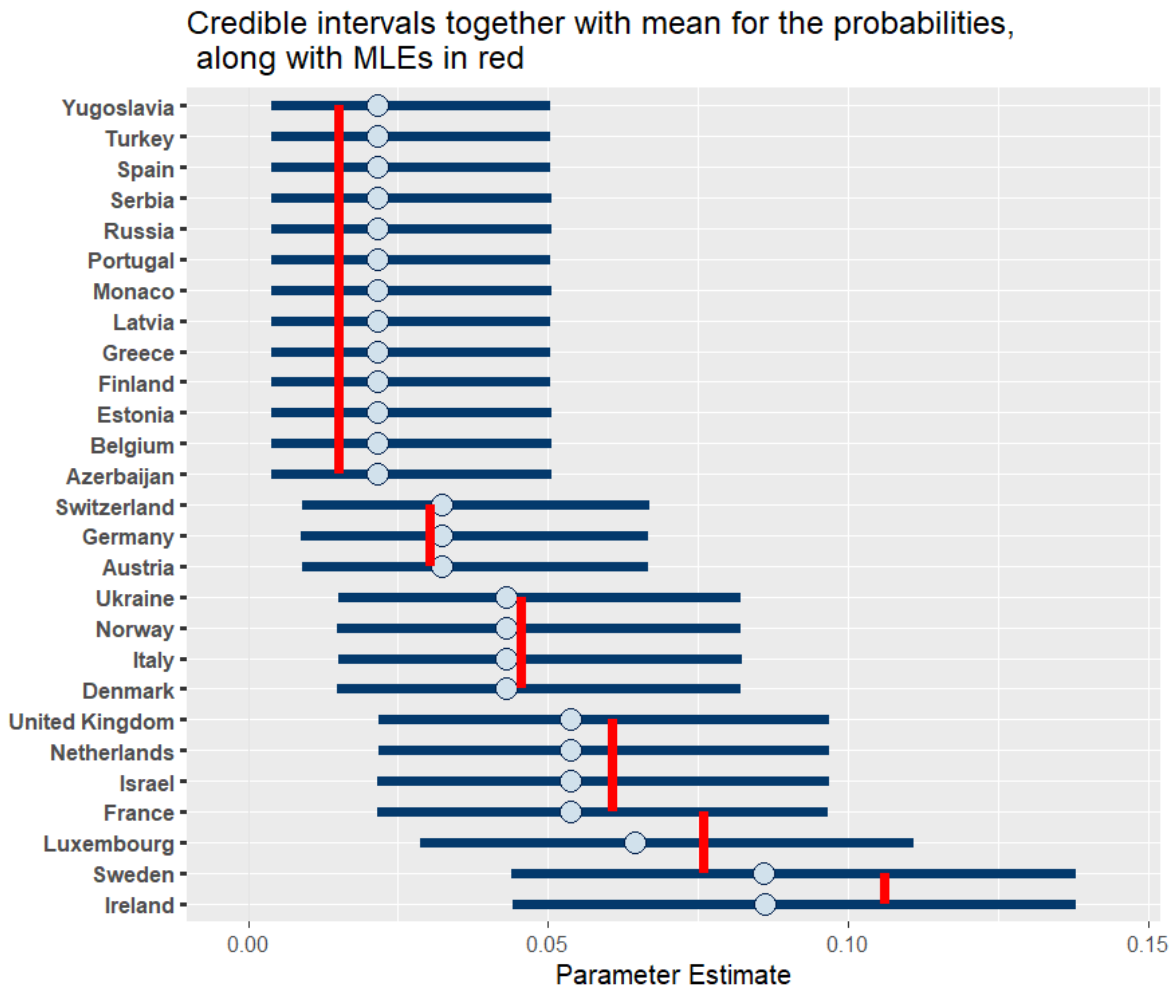


Figure 5: 90% CIs for posterior samples

by instead maximizing the following objective function

$$f(\vec{\theta}, \lambda) = \sum_{i=1}^{n} y_i \log(\theta_i) + \lambda(1 - \sum_{i=1}^{n} \theta_i).$$

Now, if we calculate the gradient of the objective function with respect to the parameters, we get that

$$\frac{df}{d\lambda} = 1 - \sum_{i=1}^{n} \theta_i \stackrel{!}{=} 0 \implies \underline{\sum_{i=1}^{n} \theta_i = 1}$$

$$\frac{df}{d\theta_i} = \frac{y_i}{\theta_i} - \lambda \stackrel{!}{=} 0 \implies \theta_i = \frac{y_i}{\lambda}$$

Now, to fulfil the underlined expression in the expressions for the gradient of λ, we have that

$$\sum_{i=1}^{n} \theta_i = \frac{1}{\lambda} \sum_{i=1}^{n} y_i = \frac{n}{\lambda} \stackrel{!}{=} 0 \implies \lambda = n$$

This then gives that

$$\hat{\theta}_i = \frac{y_i}{n}$$

The reason for bringing up the MLE here, is that my model assumes a noninformative prior such that $\pi(\vec{\theta}) \propto 1$, meaning that the posterior distribution for the parameter vector is just the normalized likelihood. Thus, I expected that the MLEs would correspond with the maximum a posteriori estimates (MAP), but this was not the case for this model, as can be seen from Figure 5, where the MLEs are represented by the red lines. Here, we see that for the countries with $1 - 4$ wins, i.e. from Yugoslavia to France on the y axis, the MLE is quite close to the mean of the posterior. The mean is of course not necessarily equal to the MAP, but as most of the posteriors are somewhat symmetric, and peaked around the mean, the mean is close to the MAP. For the most-winning countries, Luxembourg, Sweden and Ireland however, the MLEs are farther off.

It is not clear to me whether the above discrepancy between the MLE and MAP is due to errors with the implemented model, or if it is simply too much to expect that the values should be equal when we work with something numerical such as MCMC sampling. I have however thought about the possibility that the data could be the problem, as I only have 66 observations, with 27 categories. I have therefore also tested the my STAN model on simulated multinomial data, with many more observations, in the order of 1000, with just 4 categories, and success probabilities as $\frac{1}{10}\{1, 2, 3, 4\}$. The results were however pretty similar to what I have observed for the ESC data, and thus it is still unclear if the model is dysfunctional or if it is functioning as well as one might expect from a STAN model.

Because of the uncertainty regarding the functionality of the multinomial model that was just mentioned, I have also implemented a similar model where the data is thought of as binomial, given in equation 2, and a simpler non-hierarchical model. The results of the non-hierarchical model, using a non-informative prior, is presented in Figure 6. Here, we see the same trend as for the multinomial model, where we get identical estimates identical data points, as expected. We also have greater uncertainty for countries that have won the competition more times. Lastly, we can note that the means here are quite far off from the MLEs for all countries, and all probabilities seem to be overestimated compared to the MLE. This can be caused by the loss of information caused by ignoring the fact that we can only have one winner in each edition of the competition.

The hierarchical binomial model was also implemented in STAN, with the following set of parameters

$$a \sim \text{Gamma}(1, 1)$$

$$b \sim \text{Gamma}(32.33, 1)$$

which gives a prior mean of approximately 0.03 for θ_i, which is close to the mean of the observed fractions. The posteriors for these parameters are given in Figure

Since the means of these posteriors are 3.079 and 31.651, we can estimate the mean of the posterior as $\frac{3.079}{3.079+31.651} = 0.089$. A vertical line for this expected value is included in Figure 8, which depicts

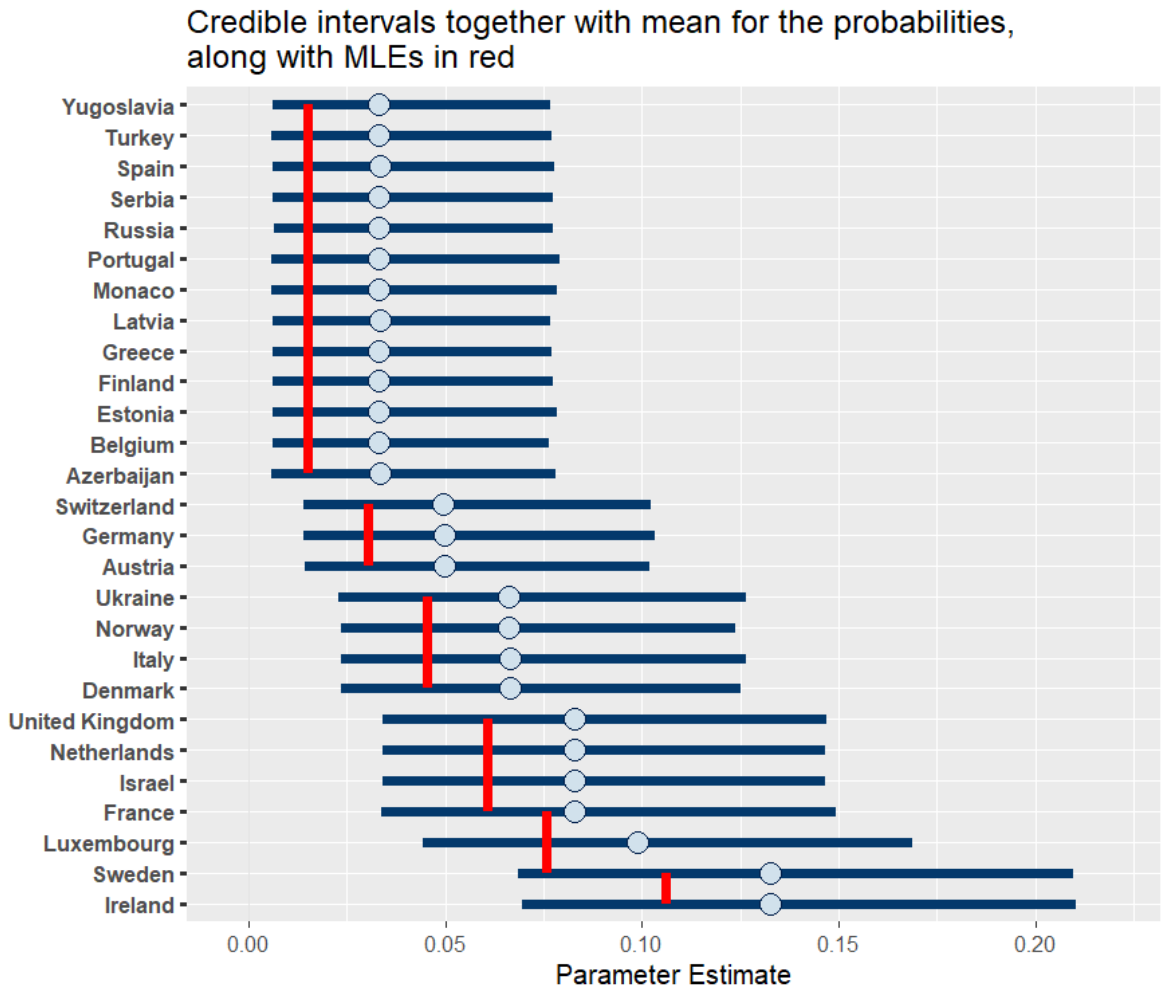
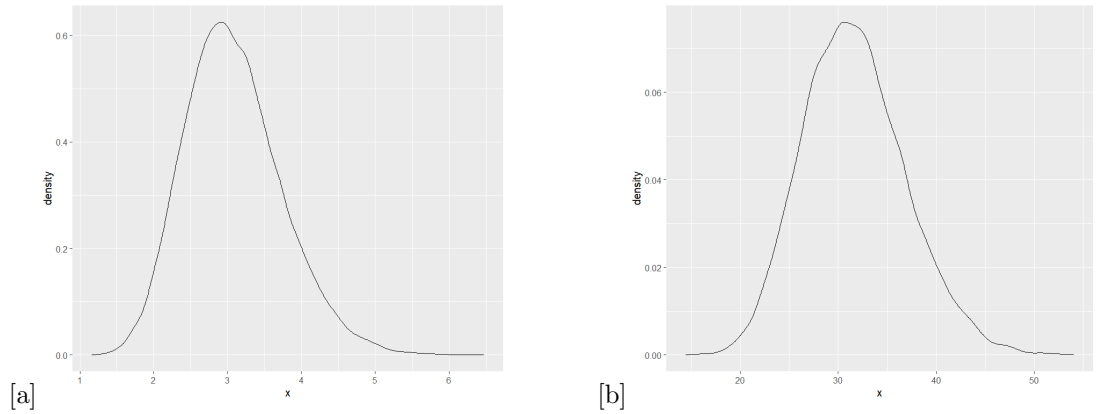


Figure 6: CIs for the non-hierarchical binomial model



[a] [b]

Figure 7: Posterior for hyperparameters for the hierarchical binomial model

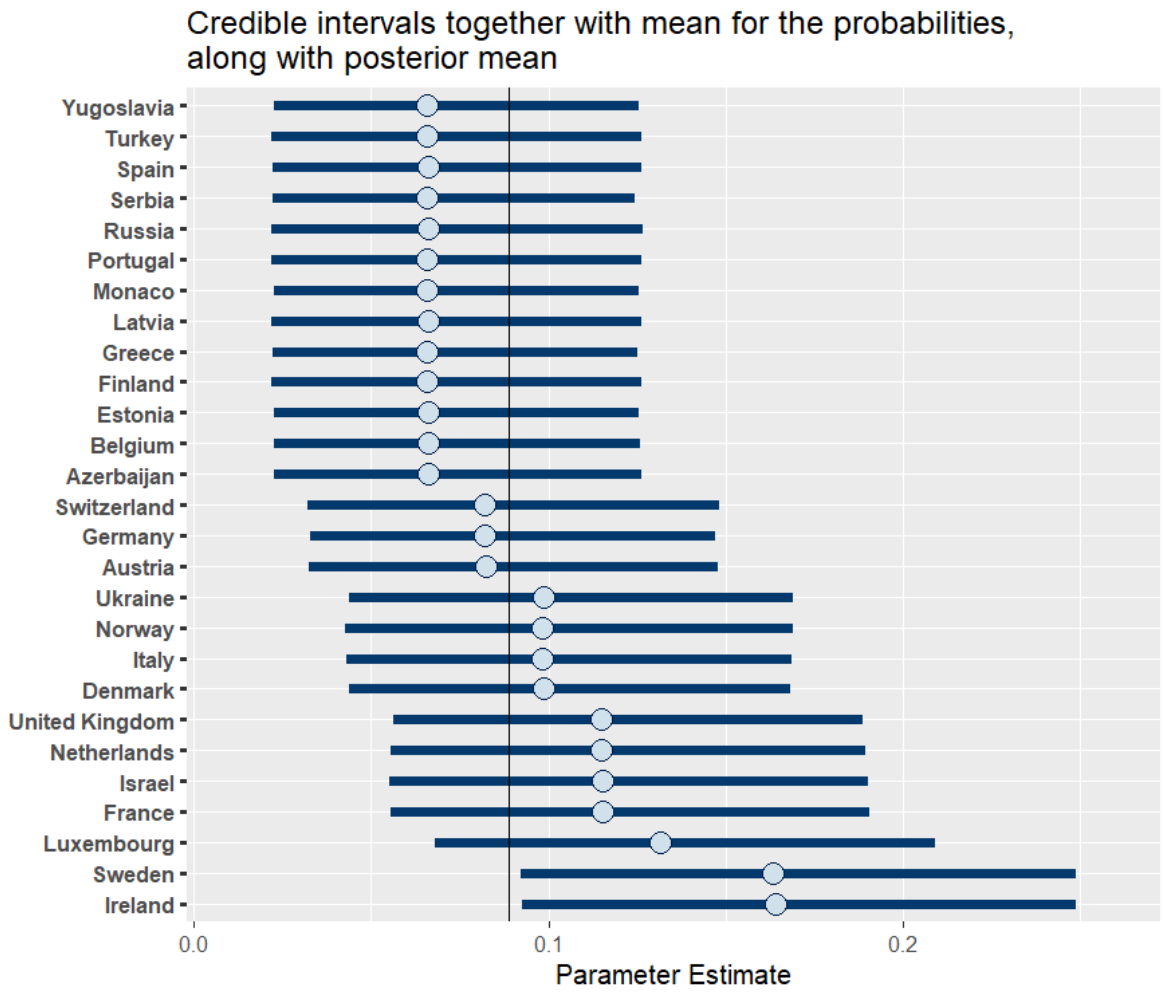


Figure 8: Hierarchical Binomial Model

the results of running this model. Here I have not included the MLEs as previously, since the MLEs no longer correspond to the MAP, as we have a hierarchical model with an informative prior. A general trend in this model is that all posterior means are higher than the corresponding values for the previously considered models. Thus, the model is likely overestimating the probability of winning for all countries. However, the model has a possible advantage, in that it is also possible to predict new θ_i, if a country that has never won the ESC before, such as Iceland, were to win one year. This might however not be very useful here, as the most logical thing to do seems to be to estimate it as for the other countries that have won once, which should be also updated.

As previously mentioned, I have also performed logistic regression in a Bayesian setting. Here, I have assumed noninformative priors on all coefficients, such that

$$\pi(\vec{\beta}) \propto 1.$$

This model has been implemented in the file 5.4. In order to asess the model, I have performed best subset selection, i.e. I have trained a model for all possible combinations of including and excluding regression coefficients, starting at the null model with only an intercept, β_0, ending at the full model with coefficients

1. β_1: Coefficient for performances by a woman.
2. β_2: Coefficient for performances by groups.
3. β_3: Coefficient for non-English songs.

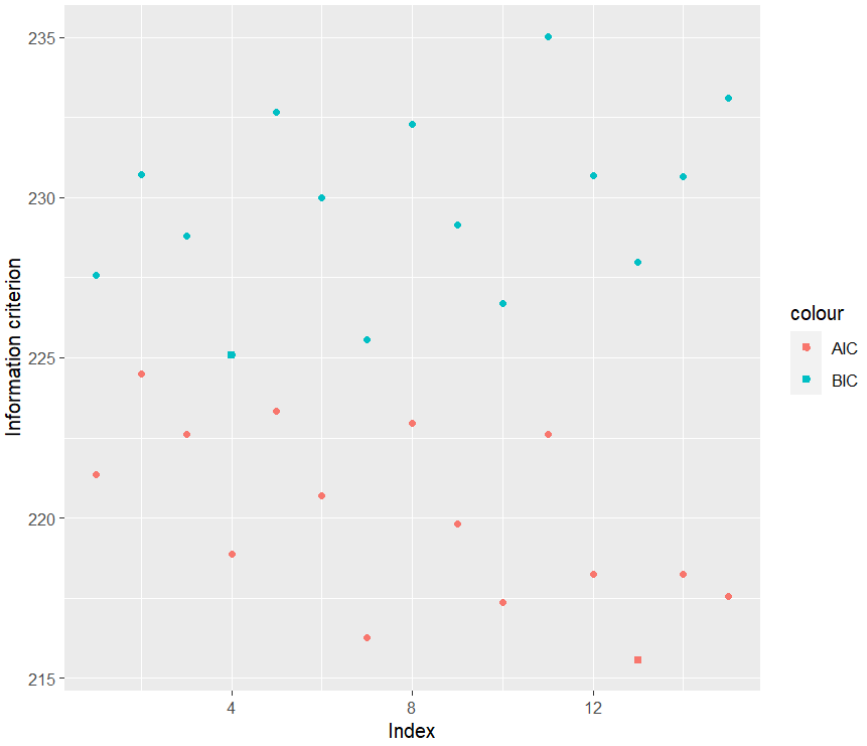


Figure 9: Best subset selection for the logistic regression

4. β_4: Coefficient for starting position.

Then, for each of these models, I have calculated the information criterion AIC and BIC, and these results are presented in Figure 9.

Here, we see that according to the BIC, we should choose subset 4, which corresponds to only including the coefficients β_0 and β_1. This seems like a too simple model, and the AIC on the other hand prefers subset number 13, which corresponds to only excluding the coefficient β_2, i.e. to not distinguish between males and groups, but include all other covariates.

I have chosen to use the model that performs best according to AIC, as the model that has the lowest BIC seems too simplistic. The posteriors for the regression coefficients for this model are presented in Figure 10. I have also fitted the corresponding model using the glm() function in R, and the summary of this model is presented in Figure 11. Here we see that the estimated coefficients from glm() are quite similar to the means of the posteriors for the intercept and gender. For the coefficients for language and starting position, the situation is different however, where the Bayesian estimates are lower than those produced by glm() in both cases.

Having performed model selection, I have also tried to validate the model. Due to time constraints I have unfortunately only been able to collect test data from this year's competition, meaning that there is only one observation that can be correctly classified as winner. Therefore, it is easier to assess the classifier's specificity, than its sensitivity, although the sensitivity is probably the most interesting quantity. Instead of doing this however, I have chosen to compare my estimated probabilities, to the probabilities that several betting companies have produced in advance of the competition. These predictions gave high probabilities for Croatia, Israel, Switzerland, France, Ireland and Ukraine. The probablities are compared to each other in Figure 12. Here, we see that my model grossly overestimates almost all probabilities, except for that of Croatia and Israel, which were actually ranked as most likely, and second-to-most likely, to win this years competition by the betting company. It should however be noted that the probabilities from the betting company have been normalized to add to 1, but this was not done by me. This could be possible to do if I had done multinomial logistic regression instead, but I was a bit put of by modelling things as multinomial from my previous potential issues with the STAN model for multinomial data.

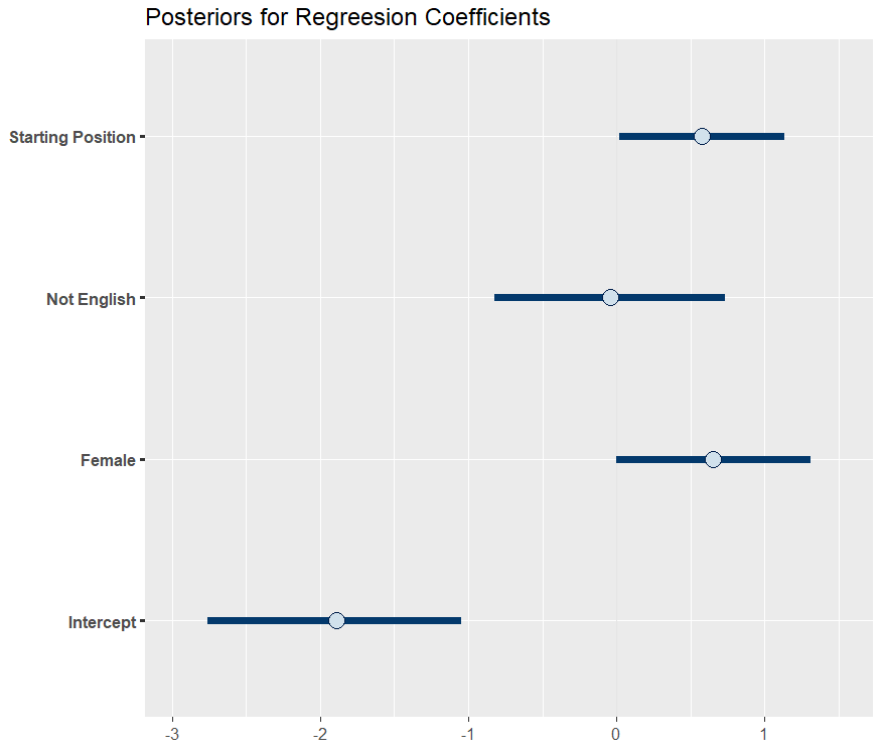


Figure 10: Posteriors for regression coefficients in the logistic regression

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.8415     0.4532  -4.064 4.83e-05 ***
genderF       0.6511     0.3343   1.948  0.05143 .
languageN     0.5518     0.3366   1.639  0.10119
prop          1.5566     0.5983   2.602  0.00927 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 222.09  on 164  degrees of freedom
Residual deviance: 207.54  on 161  degrees of freedom
AIC: 215.54

Number of Fisher Scoring iterations: 4
```

Figure 11: Summary of the GLM created by using glm() and the variables from best subset selection

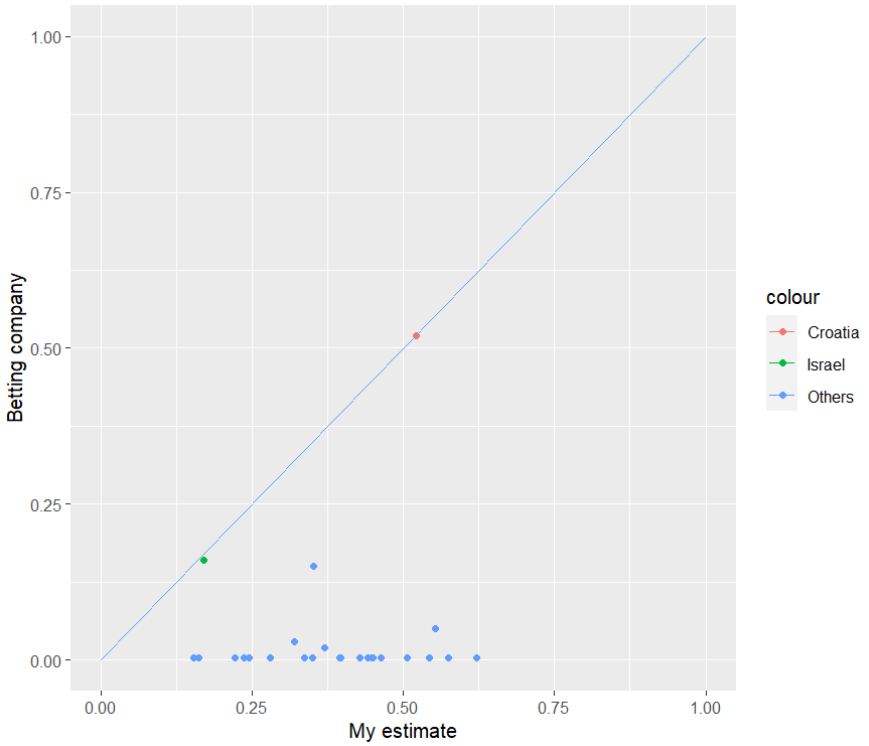


Figure 12: Comparison between my calculated probabilities and the probabilities published by a betting company

2.2 Objective 2 - Inference about Viewership Data in 2024

2.2.1 Exploratory Data Analysis

In this subsection, I will consider data on the percentage of the population that watched the ESC finale in 18 European countries, from the years 2018-2024, excluding 2020 due to covid. I have illustrated this data as time series for each country in Figure 13. By studying the series closely, and comparing the data points for 2024 versus those of 2023, we can note that the following holds:

1. Countries that experienced a decrease in viewership rate from 2023 to 2024: Austria, Belgium, the Czech Republic, Finland, Iceland, Norway, Poland and the UK

2. Countries that experienced an increase in viewership rate from 2023 to 2024: Denmark, Estonia, France, Germany (marginally, around 0.0002 difference), Ireland, Italy, the Netherlands, Spain (marginally, around 0.0009 difference) and Switzerland

Sweden had virtually identical viewership rates both in 2023 and 2024, which might be caused by their high chances in winning in 2023, and high interest in the competition in 2024 since it was in fact hosted in Sweden. It should also be noted that just comparing the numbers for 2023 and 2024 only serves as an introductory analysis, as illustrated by the data for Finland. Because, although Finland experienced a much lower viewership rate this year compared to last year, it is still higher than that of 2022. This might also be an indication that identifying viewership rates as unusual might be a challenging task.

I have also created maps that display the same data as in Figure 13, and these are given in Figure 14. The aim of this figure was to help illustrate trends in space, and something one can note here is that the viewership rates are generally highest in the Nordic countries, i.e. Scandinavia, Finland and Iceland. Iceland also stands out with achieving viewership rates of up to 50% some year. Eastern Europe on the other hand, here represented by Poland and the Czech Republic, have very low viewership rates. The remaining countries seem to be somewhat similar for most of the years. This presence of spatial trends further motivates the use of a spatial model, which will be the goal of this section.

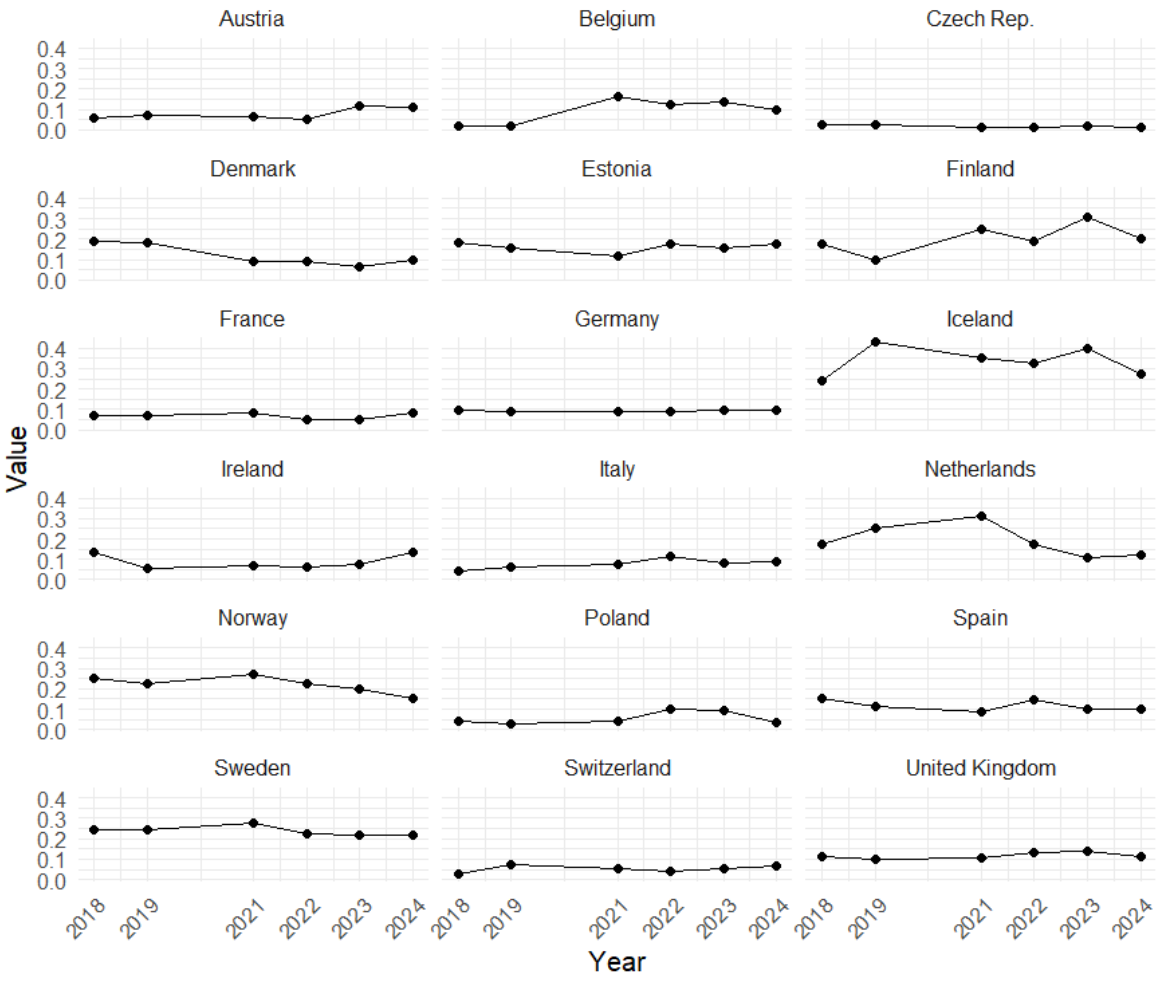


Figure 13: Depiction of viewership rates for the years 2018-2024, except 2020

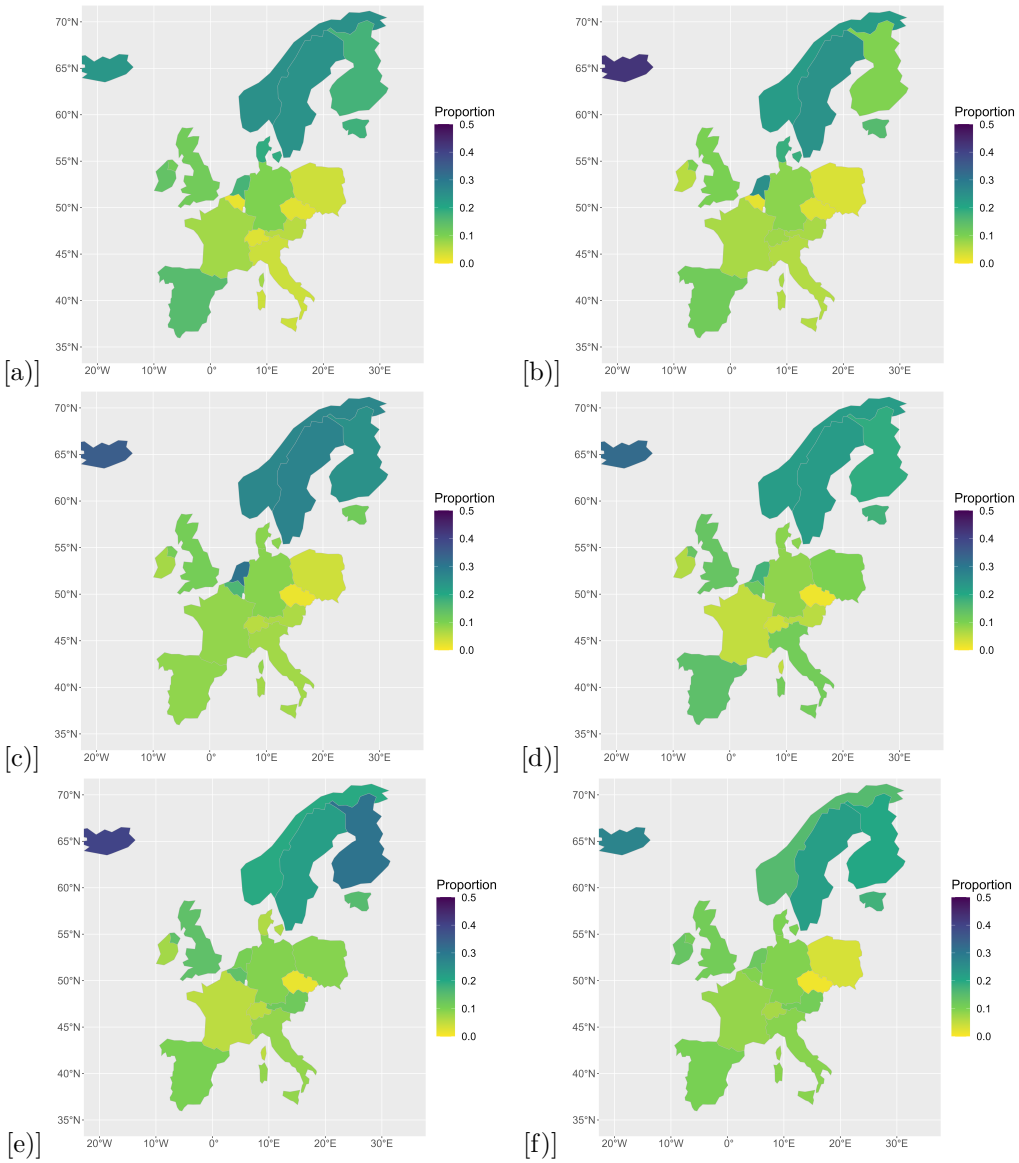


Figure 14: Viewership rates in selected European countries for the ESC finale (a) 2018 (b) 2019 (c) 2021 (d) 2022 e) 2023 f) 2024

2.2.2 Bayesian Analysis

The work in this subsection will mainly be based on what I have learned about spatial statistics in the course TMA4250 Spatial Statistics at my home university, NTNU. A specific model that I was introduced to in this course is called the Besag model, which is an improper intrinsic Gaussian Markov Random Field, and the theory here is based on the book 1.

Starting off, a Gaussian Markov Random Field (GMRF) is defined as follows, from definition 2.1 in 1.

Definition .1. A random vector $\vec{x} = (x_1, ..., x_n)^T$ is called a GMRF with respect to a labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\vec{\mu}$ and precision matrix Q, that is positive definite, iff its density is on the form

$$p(\vec{x}) = (2\pi)^{-\frac{n}{2}} |Q|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T Q(\vec{x} - \vec{\mu})\right)$$

and $Q_{ij} \neq 0 \iff \{i, j\} \in \mathcal{E}$ for all $i \neq j$

So, this is essentially a multivariate Gaussian, parameterised by the inverse of the covariance matrix, the precision matrix, but with the additional constraint that the precision matrix has a structure given by a graph $\mathcal{G}$. Here, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of the nodes in the graph and $\mathcal{E}$ is the set of edges. An example of such a graph is given in Figure 1, which I can finally refer to, where the set of nodes is all 18 countries, whereas the edges are given by lines between flags.

Although it is possible to model directly with a GMRF, a common problem with this approach is that the density given in the definition can often be improper. In order to avoid this, one must ensure that the matrix Q is positive definite, which is often easier said than done. This can however be worked around by using a Besag model, but then we also need the definition of an intrinsic GMRF, given as definition 3.2 in 1. This definition is virtually identical to the definition of a GMRF, but now we allow for a symmetric positive semidefinite matrix Q, that has rank $n - k$. Then the corresponding form for the density is

$$p(\vec{x}) = (2\pi)^{-\frac{n-k}{2}} (|Q|^*)^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T Q(\vec{x} - \vec{\mu})\right)$$

where $|Q|^*$ is the product of the non-zero eigenvalues of Q.

A special class of these improper GMRFs, are so-called intrinsic improper GMRFs, which are defined as follows, from definition 3.2 in 1,

Definition .2. An intrinsic GMRF (of first order) is an improper GMRF of rank $n-1$, i.e. $\text{rank}(Q) = n - 1$, where $Q\vec{j} = 0$. Here, $\vec{j}$ is a vector of ones.

Since we know from the definition that $Q\vec{j} = 0$, we know that if we assume that $\vec{\mu} = \vec{0}$, we see that for all $\vec{x} = \text{constant} \cdot \vec{j}$, we have that

$$p(\vec{x}) = \exp\left(-\frac{1}{2} c\vec{j}^T Q c\vec{j})\right) = \exp(0) = 1$$

Meaning that this density is in fact not proper. This creates problems when trying to sample from intrinsic GMRFs, but we will not need to do that in this project, as we will work with the Besag model. Having teased this model for a while now, the following definition finally illustrates the properties of the Besag model, which is an intrinsic GMRF, with density given by equation 3.30 in 1, i.e. that

$$\pi(\vec{x}) \propto \kappa^{\frac{n-1}{2}} \exp\left(-\frac{\kappa}{2} \sum_{i \sim j} (x_i - x_j)^2\right)$$

where $i \sim j$ means that there is an edge between node i and j, and we only count each pair of neighbours once in the sum. This density arises from assuming that the difference in two regions is distributed as follows

$$X_i - X_j \sim \mathcal{N}\left(0, \frac{1}{\kappa}\right) \quad \text{if } i \sim j.$$

Thus we can see that if we condition on knowing the value in a neighbouring region, we have that

$$X_i - X_j | X_j = x_j \sim \mathcal{N}\left(0, \frac{1}{\kappa}\right) \implies X_i | X_j = x_j \sim \mathcal{N}\left(x_j, \frac{1}{\kappa}\right)$$

Additionally, from the assumption of the distribution of differences, we see that we can write

$$X_i - X_j = \epsilon_{ij} \sim \mathcal{N}\left(0, \frac{1}{\kappa}\right)$$

Thus, by letting $\text{ne}(i)$ denote all the neighbours of node i, we see that we can write

$$\sum_{j \in \text{ne}(i)} X_i - X_j = \sum_{j \in \text{ne}(i)} \epsilon_{ij}$$

where the right hand side is distributed as $\mathcal{N}\left(0, \frac{|\text{ne}|(i)}{\kappa}\right)$, since all error terms are independent. Thus, if we condition on knowing the values in all regions except region i, denoted by $\vec{X}_{-i}$, we have that

$$\text{Var}[\sum_{j \in \text{ne}(i)} X_i - X_j | \vec{X}_{-i} = \vec{x}_i] = \text{Var}[\sum_{j \in \text{ne}(i)} X_i | \vec{X}_{-i} = \vec{x}_i] = \text{Var}[|\text{ne}(i)| \cdot X_i | \vec{X}_{-i} = \vec{x}_i] = |\text{ne}(i)|^2 \text{Var}[X_i | \vec{X}_{-i} = \vec{x}_i]$$

$$\implies \text{Var}[X_i | \vec{X}_{-i} = \vec{x}_i] = \frac{1}{|\text{ne}(i)|\kappa} \tag{4}$$

In addition, since the right hand side has mean zero, we also have that

$$\text{E}[X_i | \vec{X}_{-i} = \vec{x}_i] = \frac{\sum_{j \in \text{ne}(i)} x_j}{|\text{ne}(i)|} \tag{5}$$

These identities will later be used to simulate the proportion in each region, conditional on knowing the proportion in all other regions. In order to get there however, we also need to define the precision matrix. Its form can be found by considering a fully connected graph with three nodes, i.e. there are edges between all three nodes. Then, we have that

$$\sum_{i \sim j} (x_i - x_j)^2 = (x_1 - x_2)^2 + (x_1 - x_3)^2 + (x_2 - x_3)^2 = x_1^2 - x_1 x_2 + x_1^2 - x_1 x_3 + x_2^2 - x_1 x_2 - x_2 x_3 + x_2^2 + x_{32} - x_1 x_3 - x_2 x_3 + x_{32}$$

$$= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

This can be generalised to more complex graphs, and in general, the Besag model is characterised by the following precision matrix

$$Q_{ij} = \kappa \begin{cases} |\text{ne}(i)| & \text{if } i = i \\ -1 & \text{if } i \sim j \\ 0 & \text{else} \end{cases}$$

Now, as most of the theoretical groundwork is done, I am sure you are wondering where the Bayesian part of all this is, and rightly so. Well, the idea here has been to consider κ, often called the precision parameter, as a random variable, such that we now consider

$$\pi(\vec{x} | \kappa) \propto \kappa^{\frac{n-1}{2}} \exp\left(-\frac{\kappa}{2} \sum_{i \sim j} (x_i - x_j)^2\right)$$

The reason for following this approach, has been to use the viewership data from previous years, in order to obtain a posterior for κ. Thereafter, one can sample from this posterior, to obtain samples which one can then use to sample from one region, conditional on knowing the values in other regions. Through doing this, I hope to be able to identify some countries as sticking out, by either having very few or very many viewers in 2024, compared to what we would expect given the viewership rates in

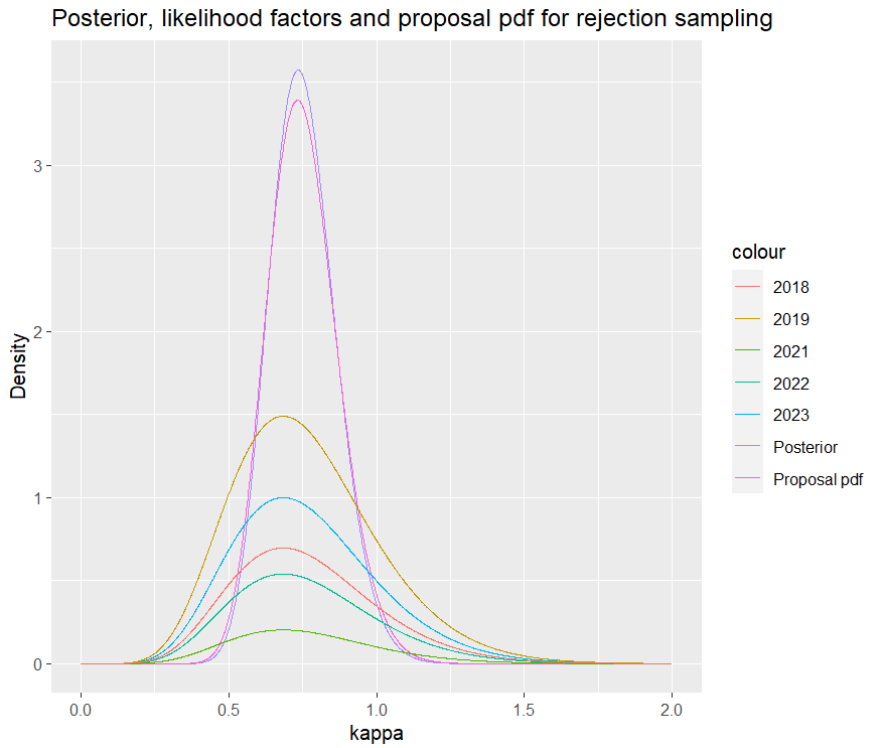


Figure 15: Posterior density together with a proposal density for rejection sampling, and likelihood functions for the different years that contribute as factors to the posteriors. The likelihoods have not been normalized, just scaled to fit in the plot, and should thus not be interpreted as densities, although their shapes are illustrative.

neighboring countries. Another possibly interesting hypothesis to investigate here could be to compare possible values for κ in 2024 compared to the previous years, through the posterior.

To start the Bayesian analysis, I have assumed a uniform prior on the precision parameter, which is restricted to be non-negative, as variance also must be non-negative. If we let Y denote a matrix with columns $\vec{y}_i$, we have that each column represents the observed viewership rates in Europe in a given year, so we have five columns in total. We can then formulate the likelihood as follows

$$\mathcal{L}(\kappa; Y) = \prod_{i=1}^{5} \kappa^{\frac{18-1}{2}} \exp\left(-\frac{\kappa}{2}\vec{y_i}^T Q \vec{y_i}\right)$$

By the choice of a uniform prior on the parameter, we then have, through Bayes' theorem, that the posterior has the following form

$$p(\kappa|Y) \propto \kappa^{\frac{5\cdot 17}{2}} \exp\left(-\frac{\kappa}{2}\sum_{i=1}^{5}\vec{y_i}^T Q \vec{y_i}\right)$$

I have calculated and normalized this posterior, and the results are presented in Figure 15. This plot also contains the likelihood factors for the different years that make up the posterior, and from these plots, we can see that for each likelihood, possible values for the precision parameter lie in the interval $(0.25, 1.5)$. By combining the knowledge from all years, we see however that the posterior has the majority of its mass within the interval $(0.5, 1)$. A 95% credible interval based on this posterior is also given in Figure 16.

In order to calculate the credible interval mentioned at the end of the last paragraph, I have sampled from the posterior. In order to do this, I have used rejection sampling, which involves a proposal density, also given in Figure 15. This is a Gamma(shape $= 40$, rate $= 0.01670278$) density. The algorithm consists of drawing a number y from the proposal density before drawing a $u \sim \text{Uniform}(0, 1)$, and

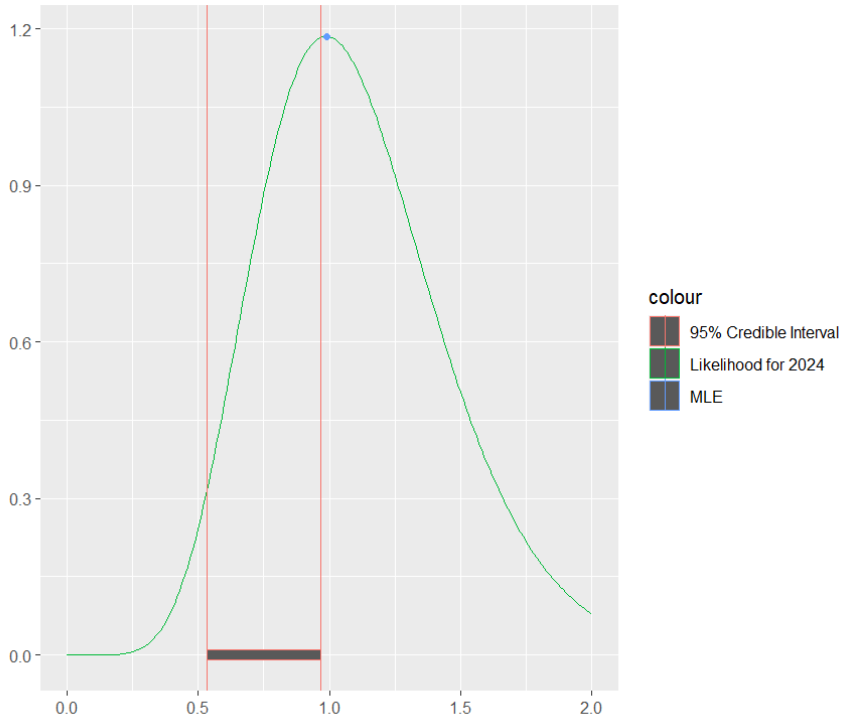


Figure 16: Posterior credible interval along with likelihood from 2024

then accepting y if

$$u \leq M\text{posterior}(y)/\text{proposal density}(y)$$

where $M = \text{supposterior}(x)/\text{proposal density}(x)$, and was here estimated to be 1.16. Thus we on average will need 1.16 iterations to obtain a sample from the posterior, meaning that it is computationally feasible to sample from the distribution.

The curve in Figure 16, is the (unnormalised) posterior for the κ from 2024, given an uniform prior, so it is really just the likelihood for this year. Here, we can note that the MLE, or equivalently, the MAP, lies outside the credible interval. This means that if we use the MLE as an estimate of the precision parameter in 2024, we would say that the posterior probability of observing a κ as large as, or greater than, our estimate, is less than 2.5%. From the figure, we also see that according to the data from 2024, we have a rather large probability, approximately one half,of observing something greater than the MLE. Therefore, it seems reasonable to conclude that the precision parameter this year likely has an exceptionally high value. This means that we expect less variance in realisations in one country, given the realised values in all other countries. This conflicts with my initial belief that one might expect more variation this year, because of the controversy surrounding the competition.

Furthermore, I have also simulated the viewership rate in each country, conditional on knowing the rate in all other countries, by using the distribution given in equations 4 and 5. As these expressions involve the precision parameter, I have for each simulation, drawn a κ from the posterior, where I have not included the data from 2024 in the posterior. This was done because these data contained proof of an exceptionally high precision, but one could also argue that the data should be included exactly because of this. I have collected 10 000 samples for each country, and histograms of the rates are presented in Figures 17 and 18.

I have then tried to summarize these results even further, by calculating the probability of observing a rate as low as the observed rate, or lower. This was done by counting the number of simulations where the simulated rate was lower than or equal to the observed rate, and then dividing by the number of simulations. These results are given in Figure 19. First of all it should be noted that the Czech Republic has an extremely low value, and this could be because of the low viewership rates, that are virtually zero, which has been observed in the country in the last years, given in Figure 13. This is a potential weakness with the model, as it does not incorporate any information about the

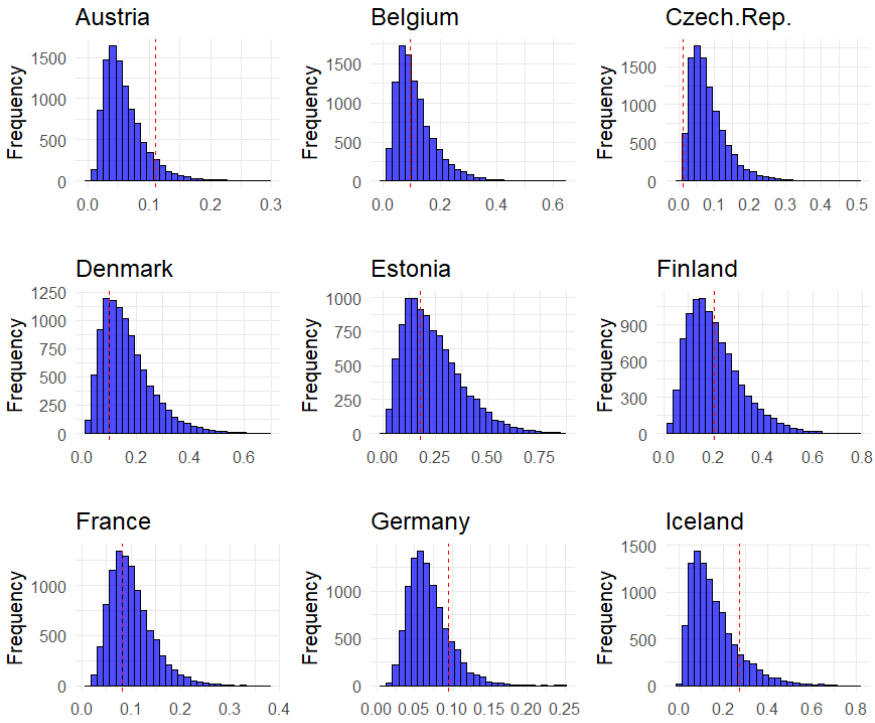


Figure 17: Histogram of simulated viewership rates for the first 9 countries. Red dotted lines represent the observed rate in 2024.

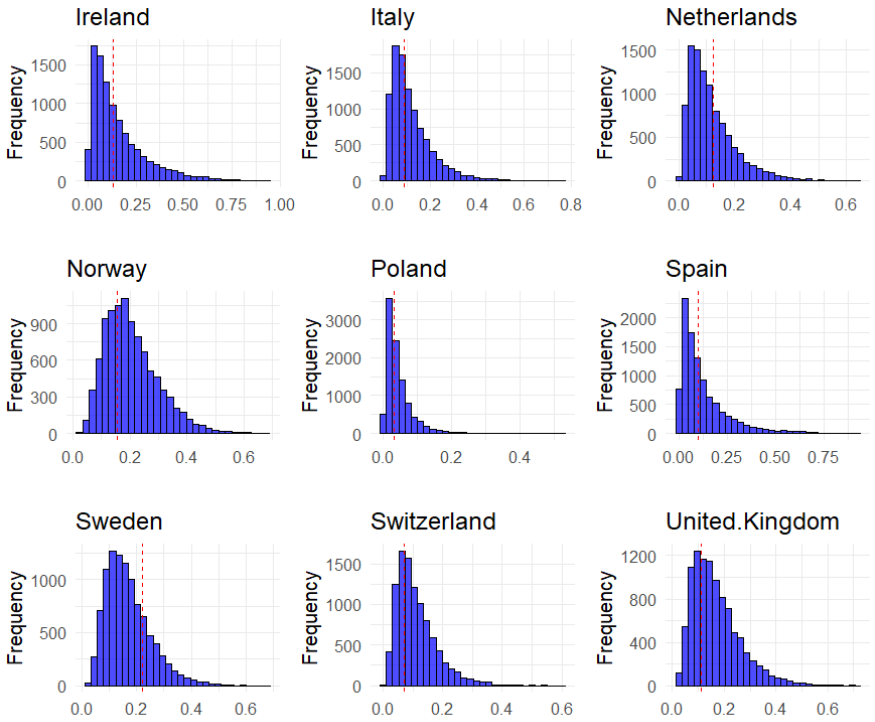


Figure 18: Histogram of simulated viewership rates for the last 9 countries.Red dotted lines represent the observed rate in 2024.

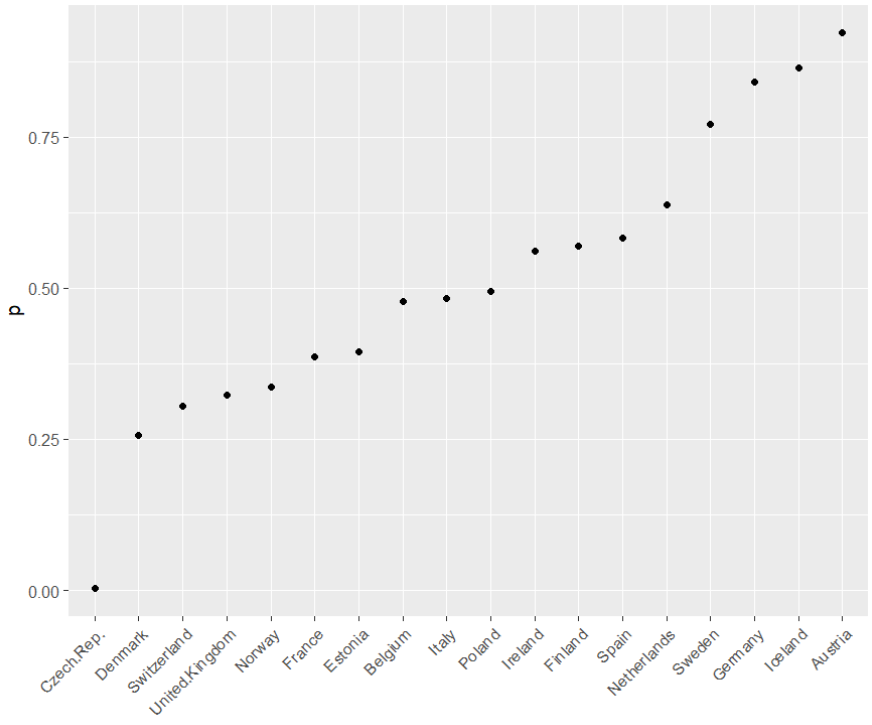


Figure 19: Estimated probabilities of observing the rates we have observed, or something even smaller

mean for each country, the mean is only dependent of the realizations in the neighboring countries. The Czech Republic only has two neighbours, Austria and Poland, which both seem to generally have higher means than the Czech Republic.

The four countries, Denmark, Switzerland, the United Kingdom and Norway, also have quite unlikely observed rates, given the rates in the neighboring countries. Thus, one could say that these countries stand out amongst their neighbours, by having fewer viewers than we would expect. Another weakness of this model is however that we cannot decide exactly what the reason for the low rates is, it could be anything from boycotts to the country not qualifying, as was the case for Denmark.

On the other hand, we can note that Sweden, Germany, Iceland and Austria experienced rather high viewership rates, given the rates in the neighbouring countries. This also illustrates the lack of information about means, as Iceland experienced a dip in the viewership rates, illustrated in Figure 13, but because so many people normally watch the ESC, this decrease is not large enough to be significant when comparing to its neighbours, Norway, the UK and Ireland, which all generally experience lower viewership rates.

3 Conclusion

3.1 Main Conclusions

The first part of this project consisted of modelling the probability of winning the ESC, first by only using the number of times a country has won. The different estimates for these probabilities are presented in Figures 5, 6 and 8. Here, we saw a general trend with larger uncertainty associated with the probabilities of the countries that have won the most times, i.e. Sweden and Ireland. Another experience I earned while working with this, is how the estimates are affected by the underlying assumptions for the distribution of the data. The multinomial approach lead to estimates that correspond quite well with the MLEs, while the standard binomial model generally produced estimates that were generally larger than the MLEs, and the MLEs are the same in these models. Lastly, the hierarchical binomial model, produced estimates that were generally higher than all the previous estimates. It should of course also be noted that modelling the data as binomial, as was done in the last two models, ignores the fact that only one country can win, and we do not get the regularising or shrinking effect we get in the multinomial model, where all probabilities must add to one.

Having estimated these probabilities, the spotlight was directed towards estimating probabilities for a song to win the ESC, using information on the language of the song, the gender of the performer, and its starting position in the competition. Through model selection and the information criterion AIC, it was found that a model that does not discriminate between songs performed by men and groups, but otherwise includes all available information, performs best. Here, the estimated coefficients tell us that songs performed by female singers and songs that have later starting position, have greater chances of winning the competition. This performance was then assessed by comparing the estimated probabilities for winning for the participating songs in 2024, where the model only performed well for 2 out of the 25 songs unfortunately.

For the second part of the project, data on the number of viewers of the ESC finale in 18 European countries was analysed. Here, a Besag-model was implemented and used, in order to first learn about the precision parameter κ. Here, the viewership rates from 2018-2019 and 2021-2023, gave the posterior for κ presented in 15. Furthermore, we saw evidence that the precision parameter was unexpectedly large in 2024 compared to previous years. This means if we consider for example Norway, we would expect a lower deviance from the mean in 2024, than in previous years, given that we know the viewership rates in Iceland, Denmark, Sweden and Finland in 2024.

Lastly, the viewership rate in 2024 in each country was simulated, conditional on knowing the rates of all other countries. Thereafter, they were compared with the actually observed rates, and these results were summarised in Figure 19. From this figure, we saw that some countries stick out as having unexpectedly low or high viewership rates compared to their neighbours.

3.2 Difficulties Encountered

One potential difficulty encountered in this project, has been implementing the multinomial and binomial models in STAN, mainly because of the discrepancy between MLEs and MAPs. It is however unclear to me still, if this is an actual problem with the model implemented, due to a lack of data, or if it is a reasonable result given the potential limitations of MCMC sampling. Additionally, it has been hard to validate the logistic regression model, as I have used all winning performances for training the model, and I was thus unable to quantify the most interesting property of the model, its sensitivity.

Another difficulty has been to gather data, especially on viewership. As previously mentioned, only 18 European countries have been included in the analysis, and this was mainly due to not finding data for the remaining countries. I also considered adding Greece and Romania, as I found data for these countries, but since they do not share borders with any of the other countries in the model, they were excluded. I was also not able to find data for years before 2018, which prevented me from learning even more about the precision parameter.

Additionally, I have realized that the Besag mdoel might be too simplistic, in that the conditional distribution given in 4 and 5 only depends on the number of neighbours, the precision parameter and the realizations in the neighboring countries. This does not account for the fact that countries might have different means for the viewership rates, where Iceland for example stands out as normally having exceptionally high rates.

3.3 Possible Extensions

When it comes to the multinomial and binomial models, the only extension I have been able to think of is that one could have worked more with the STAN models, especially if there truly are problems with them. Regarding the logistic regression, one possibility could be to instead do a multinomial logistic regression, where we account for the fact that there can only be one winner. If one were to do this however, I believe one would also need information on all the losers from the years one chooses to include. However, with those additional data points, it might still be possible to learn about the parameters, while also excluding some years, that can be held out for testing.

Regarding the Besag model, it would be possible to learn more about the precision parameter if one were able to collect samples also from the years before 2018, but this seems hard. Poland for instance, was not included in the statistic I found for 2024, so I used ChatGPT to translate "How many people watched the ESC finale in 2024 in Poland" to Polish, before googling with that translated phrase, and then translating the first website that popped up to English again. Another possible extension here could be to include information about the means in the countries, so that we can avoid the potential problems mentioned about Iceland.

Finally, one could also use an intrinsic GMRF which allows for more complex spatial relationships, by having an expected value that is a weighted sum of the realized values for the neighbours for example. This can be done by choosing the following density, which corresponds to equation 3.33 in 1

$$p(\vec{x}) \propto \kappa^{\frac{n-1}{2}} \exp\left(\frac{\kappa}{2}\sum_{i\sim j} w_{ij}(x_i - x_j)^2\right).$$

Then, with more data, one could assume priors for the weights w_ij, and also obtain posteriors for them, to see if there are any relationships between neighbours that are especially strong or weak.

4 Bibliography

1. Rue, H., Held, L. (2005). Gaussian Markov random fields: theory and applications. CRC press
2. https://en.wikipedia.org/wiki/List_of_Eurovision_Song_Contest_winners List of previous winners. Accessed 20.05.24
   (a) Several Wikipedia articles with tables on the editions from 2018, 2019, 2021, 2022 and 2023 have also been useful to collect data on the songs that did not win.
3. Previous viewership rates
   (a) https://eurovisionworld.com/esc/viewing-figures-how-many-people-watched-eurovision-2023 Viewership rates 2023 Accessed 31.05.24
   (b) https://eurovisionworld.com/esc/viewing-figures-how-many-people-watched-eurovision-2022 Viewership rates 2022 Accessed 31.05.24
   (c) https://eurovisionworld.com/esc/here-are-the-viewing-figures-for-eurovision-song-contest-20 Viewership rates 2021 Accessed 31.05.24
   (d) https://eurovisionworld.com/esc/here-are-the-viewing-figures-for-eurovision-song-contest-20 Viewership rates 2019 and 2018 Accessed 31.05.24
4. https://www.esc-plus.com/viewing-figures-eurovision-2024-grand-final-ratings-across-europe/ Viewershiprates2024 Accessed 03.06.2024
5. https://eurowizja.org/eurowizja-2024-163-miliony-widzow-konkursu/ Polish blogs for viewing figures in the ESC finale in 2024. Accessed 04.06.24
6. https://eurovisionworld.com/odds/eurovision Winning Probabilities ESC 2024 Accessed 05.06.24

5 Appendix A - STAN Models

5.1 Multinomial Model

```
data {
  int<lower=2> num_of_cats;
  int<lower=0> N;
  int y[N, num_of_cats];
}


parameters {
  simplex[num_of_cats] theta;
}

model {
  for(i in 1:N){
    y[i,] ~ multinomial(theta);
  }

}

```

5.2 Binomial Model

```
data {
  int<lower=0> N;
  int<lower=1> num_of_trials;
  real<lower=0> a;
  real<lower=0> b;
  int y[N];
}

parameters {
  vector<lower=0, upper=1>[N] theta;
}

model {
  for(i in 1:N){
    y[i]~binomial(num_of_trials, theta[i]);
    theta[i]~beta(a, b);
  }
}
```

5.3 Hierarchical Binomial Model

```
data {
  int<lower=0> N;
  int<lower=1> num_of_trials;
  int y[N];
}

parameters {
  vector<lower=0, upper=1>[N] theta;
  real<lower=0> a;
  real<lower=0> b;
}

model {
  for(i in 1:N){
    y[i]~binomial(num_of_trials, theta[i]);
    theta[i]~beta(a, b);
  }
  a~gamma(1, 1);
  b~gamma(32.33, 1);
}
```

5.4 Logistic Regression

```
data {
  int<lower=0> N;
  int <lower=0, upper=1> y[N];
  int <lower=1> num_of_params;
  matrix [N, num_of_params] x;
}

parameters {
  vector [num_of_params] beta;
}

model {
  for(i in 1:N){
    y[i]~binomial(1, 1/(1+exp(-dot_product(x[i,], beta))));
  }
}
```

6 Appendix B - R files

I was not sure if these files needed to be included, but I have done so, just in case.

6.1 Multinomial Analysis

```r
library(ggplot2)
library(dplyr)
library(rstan)
library(rstanarm)
library(bayesplot)


setwd("C:/Users/47980/OneDrive - NTNU/Documents/FYSMAT/Fjerde/Spring/Analisi Bayesia

data<-read.csv2("Data/Data.csv", sep=";", header = TRUE)

data<-data[-c(4, 55), ] #Dropping 2020 and the TIE-year

df <- data.frame(country = data$X...Winner) %>%
  group_by(country) %>%
  summarize(wins = n())

df <- df[order(-df$wins), ]

# Plot the bar plot
ggplot(df, aes(x = reorder(country, wins), y = wins)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  geom_text(aes(label = wins), vjust = -0.5) +
  labs(title = "Number of ESC Wins by Country",
       x = "Country", y = "Number of Wins") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#Bayesian Analysis

options(mc.cores = parallel::detectCores())

unique_countries <- unique(data$X...Winner)

# Create a mapping between countries and numbers
country_to_number <- match(data$X...Winner, unique_countries)

input_array<-country_to_number
```

```
N←nrow(data)

num_of_cats←length(unique_countries)

y←matrix(rep(0, N*num_of_cats), nrow=N, ncol=num_of_cats)
#Filling up the winners into the matrix. y has N rows, and one column
#for each country
for(i in 1:N){
  j←input_array[i]

  y[i, j]←1
}
"""
generate_multinomial_matrix ← function(n, k, p) {



  matrix ← matrix(0, nrow = n, ncol = k)


  for (i in 1:n) {

    category ← sample(1:k, size = 1, prob = p)

    matrix[i, category] ← 1
  }

  return(matrix)
}

p_vec←df$wins/sum(df$wins)

N←1000

y←generate_multinomial_matrix(N, 4, rep(1,4)/sum(rep(1,4)))
```

```
num_of_cats=4
"""
data_list=list(N=N, y=y, num_of_cats=num_of_cats)
#Running the model
mod<-stan("stan_models/multinomial.stan", iter = 2*50000, chains = 4,
     data = data_list, seed = 1, warmup=10000)
#Model diagnostics
traceplot(mod)

#Code for including all parameters in plot
pars<-c()
for(i in 1:num_of_cats){
  st<-paste0("theta[", i, "]")
  pars<-c(pars,st)
}

#plot(mod, pars=pars)

posterior_samples<-extract(mod)

theta_samples <- posterior_samples$theta

# Create a data frame to hold the samples

theta_df <- as.data.frame(theta_samples)
colnames(theta_df)<-unique_countries
theta_df <- theta_df[, df$country]
mcmc_intervals(
  theta_df,
  prob = 0.9 , # 95% credible intervals,
  point_est="mean",
) +
  scale_y_discrete(labels = colnames(theta_df)) +
  labs(x = "Parameter Estimate")+
  geom_segment(aes(x = 0.10606061 , xend = 0.10606061 , y = "Sweden", yend = "Irela
  geom_segment(aes(x = 0.07575758, xend = 0.07575758, y = "France", yend = "Sweden"
  geom_segment(aes(x = 0.06060606, xend = 0.06060606, y = "United Kingdom", yend =
  geom_segment(aes(x = 0.04545455, xend = 0.04545455, y = "Ukraine", yend = "Denmark
  geom_segment(aes(x = 0.03030303, xend = 0.03030303, y = "Switzerland", yend = "Aus
  geom_segment(aes(x = 0.01515152, xend = 0.01515152, y = "Yugoslavia", yend = "Aze
  ggtitle("Credible intervals together with mean for the probabilities,\nalong with
```

6.2 Binomial Analysis

```r
library(ggplot2)
library(dplyr)
library(rstan)
library(rstanarm)
library(bayesplot)

setwd("C:/Users/47980/OneDrive - NTNU/Documents/FYSMAT/Fjerde/Spring/Analisi Bayesia

data<-read.csv("Data/Data.csv", sep=";")

data<-data[-c(4, 55), ] #Dropping 2020 and the TIE-year

df <- data.frame(country = data$X...Winner) %>%
  group_by(country) %>%
  summarize(wins = n())

df <- df[order(-df$wins), ]

options(mc.cores = parallel::detectCores())

#We want theta to have a prior mean of 0.03
plot(df$wins/sum(df$wins))
a<-1
b<-a/0.03 - a

prior_tibble<-tibble(x=seq(0, 0.25, by=0.001)) %>% mutate(prior=dbeta(x, a, b))

ggplot(prior_tibble, aes(x=x, y=prior))+geom_line()

ggplot(tibble(x=seq(0, 20, by=0.01), y=dgamma(x, shape=a, rate=1)), aes(x=x, y=y))+g
ggplot(tibble(x=seq(0, 60, by=0.01), y=dgamma(x, shape=b, rate=1)), aes(x=x, y=y))+g

data_list=list(N=nrow(df), num_of_trials=nrow(df), y=df$wins, a=a, b=b)
#Running the model
mod<-stan("stan_models/binomial.stan", iter = 10000, chains = 4,
          data = data_list, seed = 1)
```

```r
#Model diagnostics
traceplot(mod)

#Code for including all parameters in plot
pars<-c()
for(i in 1:nrow(df)){
  st<-paste0("theta[", i, "]")
  pars<-c(pars,st)
}

posterior_samples<-extract(mod)

theta_samples <- posterior_samples$theta

theta_df <- as.data.frame(theta_samples)

colnames(theta_df)<-df$country
theta_df <- theta_df[, df$country]

mcmc_intervals(
  theta_df,
  prob = 0.9 , # 95% credible intervals,
  point_est="mean"
) +
  scale_y_discrete(labels = df$country) +
  labs(x = "Parameter Estimate")+
  geom_segment(aes(x = 0.10606061 , xend = 0.10606061 , y = "Sweden", yend = "Irelar
  geom_segment(aes(x = 0.07575758, xend = 0.07575758, y = "France", yend = "Sweden"
  geom_segment(aes(x = 0.06060606, xend = 0.06060606, y = "United Kingdom", yend = '
  geom_segment(aes(x = 0.04545455, xend = 0.04545455, y = "Ukraine", yend = "Denmark
  geom_segment(aes(x = 0.03030303, xend = 0.03030303, y = "Switzerland", yend = "Aus
  geom_segment(aes(x = 0.01515152, xend = 0.01515152, y = "Yugoslavia", yend = "Azer
  ggtitle("Credible intervals together with mean for the probabilities,\nalong with


#Hierarchical Binomial model
data_list=list(N=nrow(df), num_of_trials=nrow(df), y=df$wins)
#Running the model
mod<-stan("stan_models/hierarchical_binomial.stan", iter = 10000, chains = 4,
          data = data_list, seed = 1)
#Model diagnostics
traceplot(mod)

#Code for including all parameters in plot
pars<-c()
for(i in 1:nrow(df)){
  st<-paste0("theta[", i, "]")
  pars<-c(pars,st)
}

plot(mod, pars=c("a", "b"))

posterior_samples<-extract(mod)

ggplot(tibble(x=posterior_samples$a))+geom_density((aes(x=x)))
ggplot(tibble(x=posterior_samples$b))+geom_density((aes(x=x)))

theta_samples <- posterior_samples$theta

# Create a data frame to hold the samples
theta_df <- as.data.frame(theta_samples)

colnames(theta_df)<-df$country
theta_df <- theta_df[, df$country]
```

```
mcmc_intervals(
  theta_df,
  prob = 0.9 , # 95% credible intervals,
  point_est="mean"
) +
  scale_y_discrete(labels = df$country) +
  labs(x = "Parameter Estimate")+
  geom_vline(aes(xintercept = mean(posterior_samples$a)/(mean(posterior_samples$a)+
  ggtitle("Credible intervals together with mean for the probabilities,\nalong with
```

6.3 Logistic Regression

```
library(tidyverse)

library(rstan)

library(rstanarm)

library(bayesplot)

options(mc.cores = parallel::detectCores())

setwd("C:\\Users\\47980\\OneDrive - NTNU\\Documents\\FYSMAT\\Fjerde\\Spring\\Anal

winners<-read.csv2("Data\\data.csv")

winners<-na.exclude(winners)

colnames(winners)<-c("Country", "gender", "language", "start_pos",
                     "total_number_of_countries", "prop")
winners<-winners %>% select(c("Country", "gender", "language", "prop"))

winners<-winners %>% mutate(response=rep(1, nrow(winners)))

#For the losers

losers<-read.csv2("Data\\logreg.csv")

colnames(losers)<-c("Year","Country", "gender", "language", "start_pos",
                     "total_number_of_countries", "prop")

losers<-losers %>%
  filter(str_to_upper(gender) != "WINNER")

losers<-losers %>% select(c("Country", "gender", "language", "prop"))

losers<-losers %>% mutate(response=rep(0, nrow(losers)))

y<-rbind(winners, losers)
```

```
y<-y[, -1] #Dropping the name of the country

y$gender <- factor(y$gender)

y$gender<-relevel(y$gender, "M")

y$language<-factor(y$language)

y$language<-relevel(y$language, "E")

y$prop<-y$prop/100

N<-nrow(y)

x<-model.matrix(response ~ .,data=y)

data_list<-list(N=N, x=x, y=y$response, num_of_params=ncol(x))

mod<-stan("stan_models/logreg.stan", iter = 10000, chains = 4,
          data = data_list, seed = 1)
#Model diagnostics
traceplot(mod)

posterior_samples<-extract(mod)

beta_samples<-posterior_samples$beta
mcmc_intervals(
  mod,
  prob = 0.9 , # 95% credible intervals,
  point_est="mean",
  pars=c("beta[1]",
         "beta[2]",
         "beta[3]",
         "beta[4]",
         "beta[5]")
) + scale_y_discrete(labels=c("Intercept", "Female", "Group", "Not English", "Sta


summary(glm(response ~ ., data=y, family = "binomial"))

#Forward selectoin

BICs<-c()

AICs<-c()
i<-1
list_of_subsets<-list()
```

```
for(m in 1:4){
  cols ← combn(2:5, m)

  for (col in 1:ncol(cols)){
    subset←cols[, col]

    list_of_subsets[[i]]←subset

    print(i)

    x_sub←x[, c(1, subset)]

    data_list←list(N=N, x=x_sub, y=y$response, num_of_params=ncol(x_sub))

    mod←stan("stan_models/logreg.stan", iter = 10000, chains = 4,
             data = data_list, seed = 1)

    posterior_samples←extract(mod)

    beta_samples←posterior_samples$beta

    beta_params←apply(beta_samples, 2, mean)

    loglike←sum(dbinom(y$response, 1, 1/(1+exp(-x_sub%*%beta_params)), log=TRUE))

    BIC←length(beta_params)*log(nrow(x_sub)) - 2*loglike

    AIC←2*length(beta_params)-2*loglike

    BICs←c(BICs, BIC)

    AICs←c(AICs, AIC)

    i←i+1
  }
}
```

```
ggplot(tibble(x=1:length(list_of_subsets), AIC=AICs, BIC=BICs), aes(x=x, y=AIC))
  geom_point(aes(color="AIC"))+geom_point(aes(y=BIC, color="BIC"))+
  ylab("Information criterion")+xlab("Index")+
  geom_point(aes(x=which.min(AICs), y=min(AICs), color="AIC"), shape=15)+
  geom_point(aes(x=which.min(BICs), y=min(BICs), color="BIC"), shape=15)

list_of_subsets[[4]]

best_subset<-list_of_subsets[[13]]

x_sub<-x[, c(1, best_subset)]

data_list<-list(N=N, x=x_sub, y=y$response, num_of_params=ncol(x_sub))

optimal_mod<-stan("stan_models/logreg.stan", iter = 10000, chains = 4,
          data = data_list, seed = 1)

posterior_samples<-extract(optimal_mod)

beta_samples<-posterior_samples$beta

mcmc_intervals(
  mod,
  prob = 0.9 , # 95% credible intervals,
  point_est="mean",
  pars=c("beta[1]",
         "beta[2]",
         "beta[3]",
         "beta[4]")
) + scale_y_discrete(labels=c("Intercept", "Female", "Not English", "Starting Po
  ggtitle("Posteriors for Regreesion Coefficients")

beta_vec<-apply(beta_samples, 2, mean)
#Comparison with standard GLM
sd(beta_samples[, c(2)])
summary(glm(response ~ ., data=cbind(x_sub[, -1], data.frame(response=y$respons

test<-read.csv2("Data\\LogReg Test.csv")

comparison_p<-test$Probability

test<-test[, -c(1, 4, 5, 7)]

test<-cbind(test, data.frame(response=c(rep(0, 19), 1, rep(0, 5))))

x_test<-model.matrix(response ~ .,data=test)
x_test<-x_test[, c(1, best_subset)]

predictions<-x_test%*%beta_vec
probs<-1/(1+exp(-predictions))

ggplot(tibble(x=probs, y=comparison_p), aes(x=x, y=y, colour="Others"))+geom_po
  geom_point(aes(x=0.1710575, y=0.160000000 , colour=test$Country[5]))+
  geom_point(aes(x=0.5218147 , y=0.52 , colour=test$Country[22]))+
  geom_segment(aes(x=0, xend=1, y=0, yend=1))+
  labs(x="My estimate", y="Betting company")


y$response<-relevel(factor(y$response), 1)

GGally::ggpairs(y)
```

6.4 Besag Model

```r
library(ggplot2)
library(sf)
library(dplyr)
library(viridis)
library(stringr)

###Code for plotting the maps with proportions

plotMap = function(fName, dims, Val, Map, leg, colLim = NULL) {
  if(is.null(colLim)){
    colLim = range(estVal, na.rm = TRUE)
  }

  x_limits <- c(-20, 35)
  y_limits <- c(35, 70)
  # Plot
  map = ggplot() +
    geom_sf(data = Map,
            aes(fill = Val),
            color = 'gray', size = .2) +
    coord_sf(xlim = x_limits, ylim = y_limits)+
    scale_fill_viridis_c(direction = 1,
                         begin = 1,
                         end = 0,
                         limits = colLim,
                         name = leg)
  map = map +
    theme(text = element_text(size=20),
          legend.key.height = unit(1.5, 'cm'),
          legend.key.width  = unit(1, 'cm'))
  ggsave(filename = fName,
         plot = map,
         width = dims[1],
         height = dims[2])
}

# Load the Europe map data
```

```r
europe <- st_read("https://raw.githubusercontent.com/datasets/geo-boundaries-worl
europe <- europe %>% filter(continent == "Europe")

setwd("C:\\Users\\47980\\OneDrive - NTNU\\Documents\\FYSMAT\\Fjerde\\Spring\\Anal

data<-read.csv2(file="Data\\Viewership data.csv", header=TRUE)
rownames(data)<-data$X
data<-data %>% select(-X)
data<-remove_missing(data)
data<-data[-1,]

data<-data %>% filter(!rownames(data) %in% c("Greece", "Romania"))

colnames(data) <- str_replace(colnames(data), "^X", "")

data<-data %>% mutate(across(all_of(c("2023", "2022", "2021",
                       "2019", "2018")), ~ . / Population, .names = "Prop{col}"))

data<-data %>% mutate(across(all_of(c("Prop2023", "Prop2022", "Prop2021",
                                      "Prop2019", "Prop2018")), ~ log(./(1-.)), .
countries<-rownames(data)


df<-read.csv2("Data\\Viewers 2024.csv", header = FALSE)
data$Prop2024<-df$V2/df$V3
europe_filtered <- europe %>% filter(name %in% countries)

# Match proportions to countries

years<-c(2018, 2019, 2021, 2022, 2023, 2024)

for(year in years){
  colname<-paste0("Prop", year)
  pop_data <- data.frame(name = countries, response=data[, colname])

  # Merge the population data with the map data
  new_df <- merge(europe_filtered, pop_data, by.x = "name", by.y = "name")
```

```
  plotMap(fName = paste0("Pictures\\Proportions in ", year, ".png"),
              dims = c(10,8),
              Val = data[, colname],
              Map = new_df,
              leg = "Proportion",
              colLim = c(0, 0.5))
}
#-------------------------------------------------------

#Code for Bayesian Analysis:
#Neighbourhood matrix
N_mat<-read.csv2("Data\\Neighbors.csv", header=TRUE)
N_mat<-N_mat[,-1]
rownames(N_mat)<-colnames(N_mat)

N_mat<-N_mat[-c(9, 16), -c(9, 16)]

#Structure matrix
R<- -(N_mat+t(N_mat))
diag(R)<-abs(apply(R, MARGIN=1, sum))
R<-as.matrix(R)

#Determining the posterior for kappa
kappa<-seq(0, 2, by=0.001)

plot_df<-tibble(kappa)

#Likelihoods for all years
y<-data$GRFProp2023
plot_df$GRF2023<-kappa^((18-1)/2)*exp(-1/2*kappa*t(y)%*%R%*%y)

y<-data$GRFProp2022
plot_df$GRF2022<-kappa^((18-1)/2)*exp(-1/2*kappa*t(y)%*%R%*%y)

y<-data$GRFProp2021
plot_df$GRF2021<-kappa^((18-1)/2)*exp(-1/2*kappa*t(y)%*%R%*%y)
```

```r
y<-data$GRFProp2019
plot_df$GRF2019<-kappa^((18-1)/2)*exp(-1/2*kappa*t(y)%*%R%*%y)

y<-data$GRFProp2018
plot_df$GRF2018<-kappa^((18-1)/2)*exp(-1/2*kappa*t(y)%*%R%*%y)

#Plot of likelihoods
ggplot(plot_df, aes(x=kappa))+geom_line(aes(y=GRF2023, colour="2023"))+
  geom_line(aes(y=GRF2022, colour="2022"))+
  geom_line(aes(y=GRF2021, colour="2021"))+
  geom_line(aes(y=GRF2019, colour="2019"))+
  geom_line(aes(y=GRF2018, colour="2018"))+
  ggtitle("Likelihood for kappa for the observed years")

prior<-rep(1, nrow(plot_df))
plot(plot_df$kappa, prior)
likelihood<-apply(plot_df, 1, prod)
plot(plot_df$kappa, likelihood)

post<-prior*likelihood #Now, to normalize the posterior

library(pracma)
#Tools for approximating the posterior
approx_params <- function(pdf_values, locations) {
  # Trapezoidal rule for numerical integration
  normalize_constant <- trapz(locations, pdf_values)

  # Mean calculation
  mean <- trapz(locations, pdf_values * locations) / normalize_constant

  # Variance calculation
  variance <- trapz(locations, pdf_values * (locations - mean)^2) / normalize_co

  # Standard deviation
  sd <- sqrt(variance)

  return(list(mean = mean, sd = sd, normalize_constant=normalize_constant))
```

```r
}

results<-approx_params(post, plot_df$kappa)

mean_post<-results$mean #Tried to use this to find a proposal density for rejectior
#sampling

sd<-results$sd #Tried to use this to find a proposal density

plot_df$post<-post/results$normalize_constant
posterior_plot<-ggplot(plot_df,aes(x=kappa)) + geom_line(aes(y=post, color="Poster:

a<-40

b<-mean_post/a
ggplot(plot_df,aes(x=kappa)) + geom_line(aes(y=post, color="Posterior"))+
  geom_line(aes(y=dgamma(kappa, shape=a, scale=b), col="Proposal pdf"))+
  geom_line(aes(y=GRF2023/max(GRF2023), col="2023"))+
  geom_line(aes(y=GRF2023/max(GRF2022), col="2022"))+
  geom_line(aes(y=GRF2023/max(GRF2021), col="2021"))+
  geom_line(aes(y=GRF2023/max(GRF2019), col="2019"))+
  geom_line(aes(y=GRF2023/max(GRF2018), col="2018"))+
  labs(x="kappa", y="Density")+
  ggtitle("Posterior, likelihood factors and proposal pdf for rejection sampling")

#-----------------------------------------
#Rejection sampling for sampling from the posterior
#First I find the normalizing constant for the posterior again

GRF_responses<-data %>% select(c("GRFProp2023",
                                 "GRFProp2022",
                                 "GRFProp2021",
                                 "GRFProp2019",
                                 "GRFProp2018"))
kappa<-seq(0, 2, by=0.01)
posterior_vec<-rep(1, length(kappa))
```

```
for(i in 1:ncol(GRF_responses)){
  y<-GRF_responses[, i]
  posterior_vec<-posterior_vec*kappa^((18-1)/2)*exp(-1/2*kappa*t(y)%*%R%*%y)
}

normalization_constant<-trapz(kappa, posterior_vec)

posterior_pdf<-function(kappa){
  posterior<-rep(1, length(kappa))
  for(i in 1:ncol(GRF_responses)){
    y<-GRF_responses[, i]
    posterior<-posterior*kappa^((18-1)/2)*exp(-1/2*kappa*(t(y)%*%R%*%y))
  }
  return(posterior/normalization_constant)
}
plot(kappa, posterior_pdf(kappa))
lines(kappa,posterior_vec/normalization_constant) #The methods are equivalent

#Rejection sampling

#Proposal density: dgamma(shape=a=40, scale=mean/a=0.01670278)
lr<-function(kappa){
  return(posterior_pdf(kappa)/dgamma(kappa, shape=a, scale=b))
}
M<-optimize(lr, c(0.01, 2), maximum=TRUE)$objective
plot(kappa, lr(kappa))
rejSample<-function(num_of_samples){
  a<-40
  b<-mean_post/a
  samples<-c()
  while(length(samples)<num_of_samples){
    y<-rgamma(1, a, scale = b)
    u<-runif(1)

    if(u<posterior_pdf(y)/(M*dgamma(y, a, scale = b))){
      samples<-c(samples, y)
    }
```

```
  }
  return(samples)
}

sample<-rejSample(10000)
ggplot()+geom_density(aes(x=sample, colour="Sampled")) + geom_line(aes(x=kappa, y=

#https://www.esc-plus.com/viewing-figures-eurovision-2024-grand-final-ratings-acr

data_2024<-read.csv2("Data\\Viewers 2024.csv", header = FALSE)
data_2024[1, 1]<-"Austria"
rownames(data_2024)<-data_2024$V1

response_2024<-logit(data_2024$V2/data_2024$V3)

likelihood_2024<-kappa^((18-1)/2)*exp(-1/2*kappa*t(response_2024)%*%R%*%response_

post_2024<-likelihood_2024/trapz(kappa, likelihood_2024)

set.seed(2)
CI<-quantile(rejSample(100000), c(0.025, 0.975))



ggplot(tibble(x=kappa, y=post_2024), aes(x=x, y=y, colour="Likelihood for 2024"))
  geom_point(aes(x=x[which.max(y)], y=max(y), colour="MLE"))+
  geom_rect(aes(xmin = CI[1], xmax = CI[2], ymin = -0.01, ymax = 0.01, colour="95%
  geom_vline(aes(xintercept = CI[1], colour="95% Credible Interval"))+
  geom_vline(aes(xintercept = CI[2], colour="95% Credible Interval"))+
  labs(x="", y="")

#Sampling
set.seed(2024)
num_of_sims<-10000

sampled_2024<-as.data.frame(matrix(rep(0, length(response_2024)*num_of_sims), nrow
for(j in 1:length(response_2024)){
  country<-data_2024$V1[j]
  print(country)
  R_restricted<-as.data.frame(t(R[country,])) %>% select(-all_of(country))
  R_restricted<-as.matrix(R_restricted)
  samples_country<-c()

  observed_values<-response_2024[-which(data_2024$V1==country)]
  mean_country<- -1/R[country, country]*(R_restricted%*%observed_values)

  for(i in 1:num_of_sims){
    kappa_sample<-rejSample(1)
    print(i)
    country_sample<-rnorm(1, mean_country, 1/sqrt(R[country, country]*kappa_sample
    samples_country<-c(samples_country, country_sample)
  }
  sampled_2024[, j]<-samples_country
}

sampled_2024<-1/(1+exp(-sampled_2024))

colnames(sampled_2024)<-data_2024$V1

library(gridExtra)

plots <- list()

# Loop through each column (country) in the dataframe
```

```r
for (i in 1:9) {
  country<-colnames(sampled_2024)[i]
  print(country)
  p <- ggplot(sampled_2024, aes_string(x = country)) +
    geom_histogram(fill = "blue", color = "black", alpha = 0.7) +
    theme_minimal() +
    geom_vline(xintercept = 1/(1+exp(-response_2024[which(data_2024$V1==country)])))
    labs(title = country, x = "", y = "Frequency")
  plots[[country]] <- p
  print(1/(1+exp(-response_2024[which(data_2024$V1==country)])))
}

do.call("grid.arrange", c(plots, nrow = 3, ncol = 3))

plots <- list()

for (i in 10:18) {
  country<-colnames(sampled_2024)[i]
  print(country)
  p <- ggplot(sampled_2024, aes_string(x = country)) +
    geom_histogram(fill = "blue", color = "black", alpha = 0.7) +
    theme_minimal() +
    geom_vline(xintercept = 1/(1+exp(-response_2024[which(data_2024$V1==country)])))
    labs(title = country, x = "", y = "Frequency")
  plots[[country]] <- p
  print(1/(1+exp(-response_2024[which(data_2024$V1==country)])))
}

do.call("grid.arrange", c(plots, nrow = 3, ncol = 3))

p_values<-c()
for(i in 1:18){

  observed_value<-1/(1+exp(-response_2024[i]))

  p<-sum(sampled_2024[, i]<observed_value)/num_of_sims

  p_values<-c(p_values, p)
}

df<-tibble(x=colnames(sampled_2024), y= p_values) %>% arrange(y)
df$x <- factor(df$x, levels = df$x)

ggplot(df, aes(x=x, y=y))+
  geom_point()+theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  ylab("p")+
  xlab("")

#EDA

times_series_df<-cbind(data %>% select(c("Prop2018", "Prop2019", "Prop2021",
                                         "Prop2022", "Prop2023")),
                       data.frame("Prop2024"=1/(1+exp(-response_2024))))
library(tidyr)
library(tibble)

df_long <- times_series_df %>%
  rownames_to_column(var = "Country") %>%
  pivot_longer(cols = starts_with("Prop"), names_to = "Year", values_to = "Value")
```

```r
# Create the plot
p ← ggplot(df_long, aes(x = as.integer(gsub("Prop", "", Year)), y = Value)) +
  geom_line() +
  geom_point() +
  theme_minimal() +
  labs(x = "Year", y = "Value") +
  scale_x_continuous(breaks = c(2018, 2019, 2021, 2022, 2023, 2024)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  facet_wrap(~Country, nrow = 6)
p
```