Forest Fires in Brazil

Davide Secco, Sandro D'Andrea, Laura Fehringer

June 14, 2024

1 Introduction

The Goal of this task is to predict the number of monthly forest fires in Brazil and in specific Brazilian states. The dataset we used consisted in observations of monthly fires per state from 1998 to 2017. To better model the data we augmented it with a second dataset regarding climate measures from the National Oceanic and Atmospheric Administration (NOAA). For this task we analyzed the effectiveness of traditional bayesian approaches (plus AutoRegressive components) compared to the frequentistic one, and the use of Spatial autocorrelation modelling techniques with conditional AutoRegressive priors to take into account the adjacency structure of the states themselves.

2 Data Preprocessing

2.1 Dataset Description

The dataset consists of 6454 observations of the number of monthly forest fires in Brazil per state from June 1998 to November 2017.

The dataset is composed by 4 variables:

- **Year**
- **Month**
- **State**
- **Number:** the total number of fires recorded in thousands

2.2 Data Preprocessing

The first problem the dataset presented was that some of the states in it were referred to with the same name: for example, every recorded month had 3 different observation for the state "Rio" because "Rio de Janeiro", "Rio Grande do Sul" and "Rio Grande do Norte" were all referred to with the same name. To solve this problem we downloaded from the original source of the dataset on Kaggle the cleaned version with all the different states correctly divided.
After that we cleaned the names of the states and the months by removing characters wrongly encoded (from the Brazilian alphabet).

To better model and analyze the effect of time and seasonality on the number of fires we added a season column, which we then One-hot-encoded, and a year trend column, with values from 1 to 20 one for each year.

Since the Dataset had so few variables, enhance it by joining with an external second one: the second dataset, that we are using, is the monthly Extended Reconstruction Sea Surface Temperature version 5 (ERSSTv5) from the National Oceanic and Atmospheric Administration (NOAA). In particular we are using the climatological mean Sea Surface Temperature (SST) for the given month and year, adjusted to account for long-term average conditions. It reflects the expected SST value based on historical averages over a defined climatological base period. Additionally we are using the anomaly variable indicating how much the observed SST deviates from the expected climatological mean. Positive values indicate warmer-than-average conditions, while negative values indicate cooler-than-average conditions [ON].

Consequently, we applied a standard scaling to the variables year trend, nino:ANOM and nino:ClimAdjust.

In conclusion, the dataset was split into different datasets in order to prepare it for the different modelling techniques:

- **JAGS modelling:** for our first approach we decided to only focus on creating a model for only one state in order to simplify the task, in this case we extracted all the observations from "Tocatins"
- **CARBayes single value:** since this modelling technique analyzes the effect that neighboring states have on each other by only taking a single value from each, we selected the data from November 2013 (November was chosen because, as we will see in the next section, it provides high variance between the observations)
- **CARBayes multi value:** this model has the same objective as the last one, the only difference is that it takes in multiple values from the same state; we decided to only use the data from 2005 onward in order to decrease the time needed by the fitting and to decrease the effect that the year trend has on the observations

For these last two tasks we also needed to compute the neighborhood matrix of the states: this was possible thanks to the libraries **geobr**, with which we were able to download the shapes of the states as polygons, and **spdep**, which provides functions that, starting from a series of polygons, output a neighborhood matrix.

2.3 Data Exploration and Analysis

After cleaning the data we focused on exploring the dataset to find interesting patterns that could help us in better creating our models.

First of all, we wanted to explore how the number of fires vary based on the month of the year. To do so, we plotted the boxplot for each month:

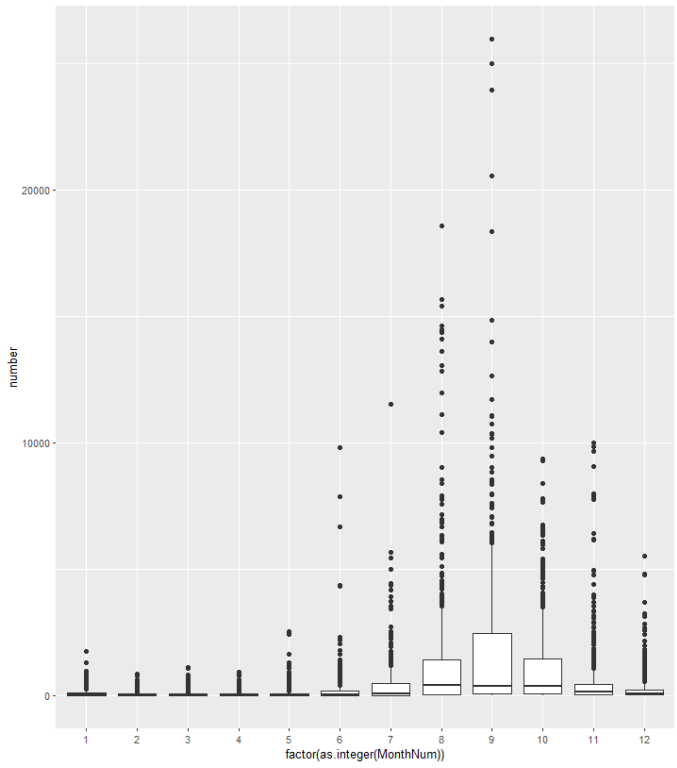

Figure 1: Boxplots of number of fires per month

As expected, the number varies a lot from month to month. Not only that, but analyzing the same boxplots for each state individually showed that this pattern differs also based on the state,

probably due to the size of Brazil and the difference in climates, temperatures and enviroments in it.

Figure 2 shows the cdf of the Number of Fires for selected states. As expected different states depict different patterns and different priors might be suitable for different states.

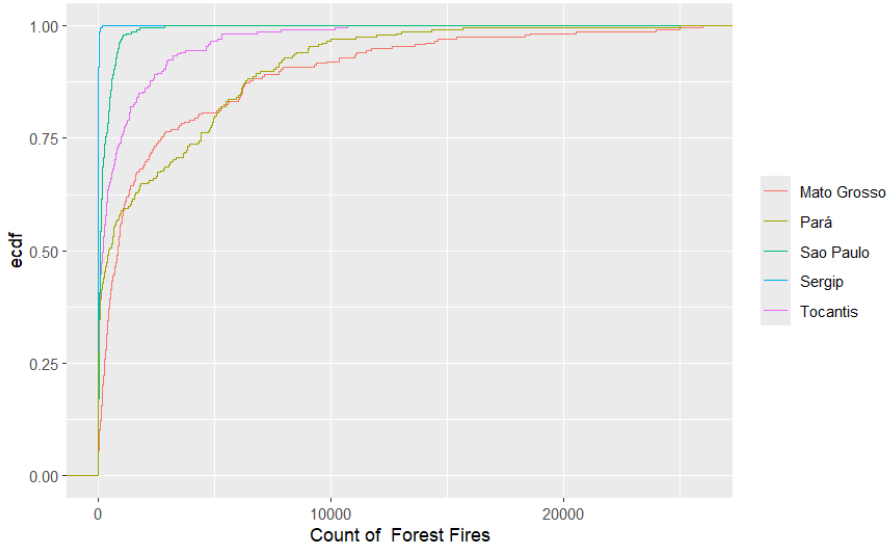

Figure 2: Distribution Per State

In fact, we noticed how some states present a lot more zeroes than the others making it viable to use a Zero Inflated Poisson model instead of a Poisson one for these specific ones.

Next, we analyzed how the values from the Nino Dataset correlate to the number of fires recorded: Figure 3 shows how the variance in the number of fires is inverse proportional to the values of the variable ClimAdjust; moreover the values are very clearly divided into three separate bins, which could make discretizing this variable an interesting option.

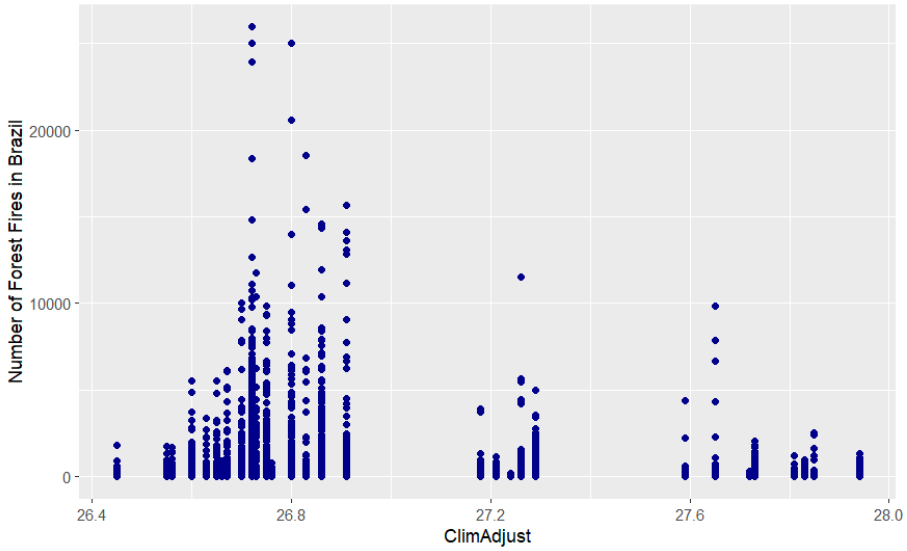

Figure 3: ClimAdjust and Number of Forest Fires

2.4 Forest Fires in Tocantis - Regression

We explored various methods to simulate the counts of Forest Fires in Tocantis. A Traditional Bayesian model that relies on the predictors,as well as several time series models. For comparison we also fitted a frequentist regression model for count data.

2.4.1 Traditional Bayesian

The Traditional Bayesian Model has following hierarchical structure. Where Y represents the count of Forest Fires in the State "Tocantis" and X is the design matrix.

$$\begin{aligned} Y &\sim Pois(\lambda) \\ \lambda &= log(\boldsymbol{X\beta}) \\ \beta_i &\sim N(0, \sigma_i^2) \\ \sigma_i^2 &\sim Unif(0.05, 10) \end{aligned} \tag{1}$$

It was necessary to restrict the variance to values between 0.05 and 10, since otherwise the variance of the β coefficients would have exploded. Hence the Uniform Distribution, instead of the more common inverse Gamma distribution. But even when using the inverse Gamma distribution as a prior the *beta* coefficients converged to the same numbers as with the more restrictive prior.

For the simulation in Jags we set the thinning value to 20, due to persistent auto correlation in the trace for the seasonal dummy coefficients. We deployed three chains in parallel, who all converge to values close to each other.

The β coefficients and therefore also the fitted values from the Standard Bayesian Model are in absolute numbers close to the ones from the frequentist Model, but they are not in the 95% Confidence Interval of the Bayesian Coefficients. Even if we initialise the Bayesian coefficients as the ones of the frequentist model, they converge back to the original Bayesian coefficients we see in Figure 4. In blue the distribution of the Bayesian Coefficients is given, the mean of the frequentist model is in red.

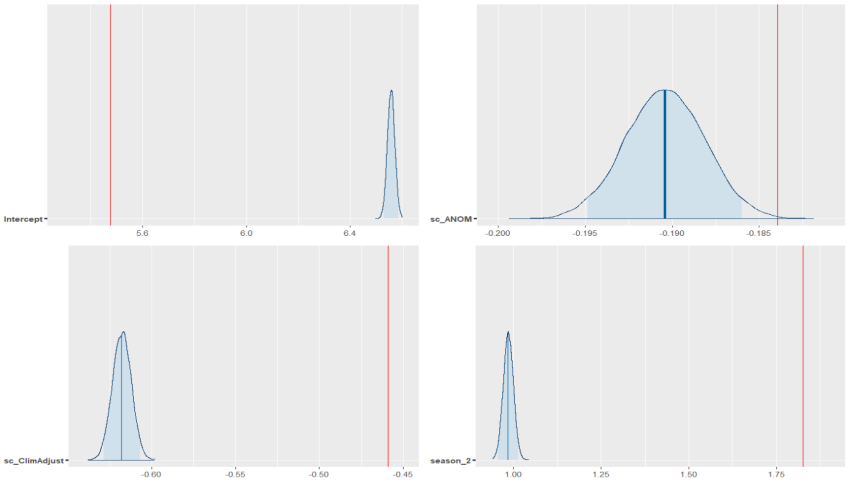

Figure 4: Frequentist and Bayesian Coefficients (in following order: Intercept, scaled Anom, scaled ClimatAdjust, dummy for Spring)

Figure 5 shows True Values as well as the Fitted Values of the Standard Bayesian Model. We can very nicely see the seasonal pattern painted by the seasonal dummy coefficients. The Model fits low numbers of Forest Fires quite well, but does not capture higher counts.

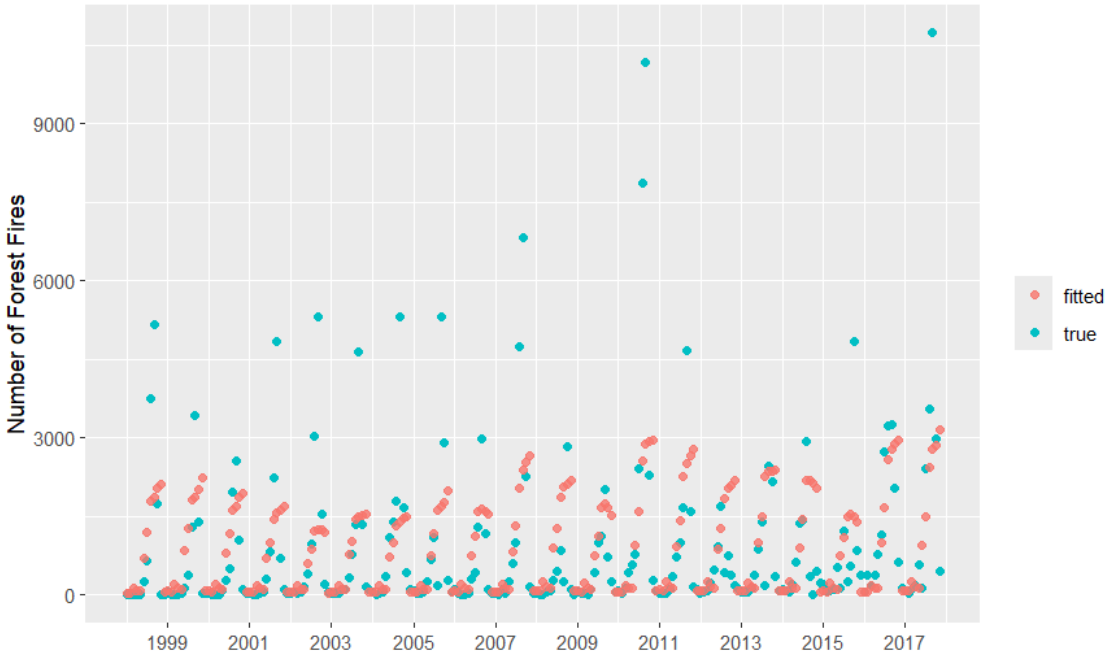

Figure 5: True VS. Fitted Values from the Standard Bayesian Model

2.4.2 Timeseries (variable selection)

We also fitted the following autoregressive timeseries model with a poisson likelihood and Uniform Priors for the coefficients to the data. Where k is the number of lags.

$$\begin{aligned} Y &\sim Pois(\lambda) \\ \lambda &= log(\sum_{i}^{k} Y_{t-i} * \alpha_{t-i}) \\ \alpha_i &\sim Unif(-1, 1) \end{aligned} \tag{2}$$

We tried an AR(1), and AR(2) and AR(12) and a Model with only a lag at $k = 1$ and $k = 12$. Using the Watanabe–Akaike Information Criterion (WAIC), we can compare the models to assess whether it is worthwhile to add more time-shifts to the series. We rank the AR with poisson distribution models with the loo library:

	$elpd_{diff}$	se_{diff}
AR2 poisson	0.0	0.0
AR1 poisson	-0.5	0.5
AR12 poisson	-1.0	0.6
AR1 12 poisson	-7.9	3.9

According to the WAIC the model with 2 lags with k=1 and k=1 performs the best.

We noticed that the AR coefficients of our Poisson AR(2) model were quite small, so we also tried to fit a model with a normal likelihood. The normal distribution is definitely not a perfect fit as it also ranges negative values and draws from the realm of continuous numbers. But it is compared with a poisson distribution more flexible, since mean and variance can take different values. Comparing the autoregressive coefficients of the AR(2) model with poisson likelihood and the one with normal likelihood we see big differences depicted in figure 6. The autoregressive coefficients of the normal model are decisively larger.

When we compare the mean square error on the fitted data we can see that the gaussian model is performing much better than the poisson model. For our future analysis we nevertheless went with the poisson likelihood.

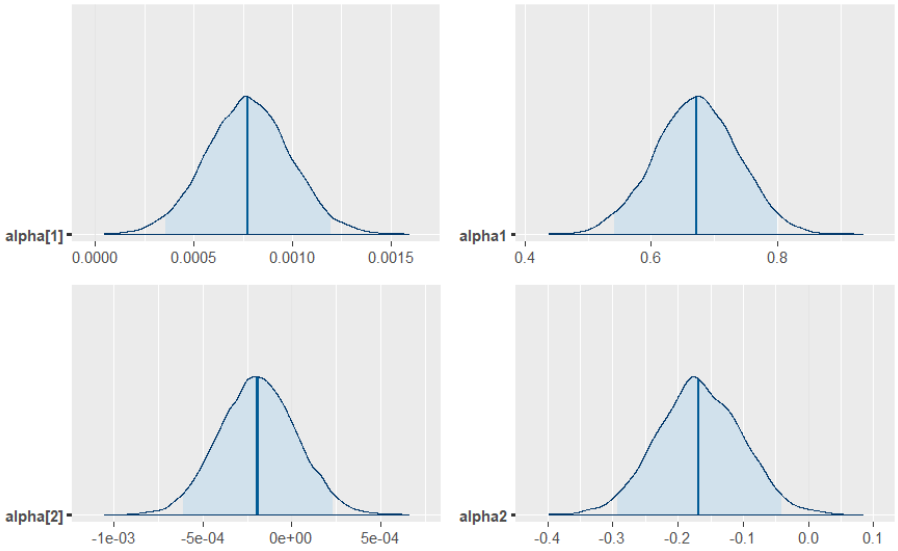

Figure 6: Coefficients with poisson likelihood (left) compared with normal likelihood (right)

Table 1: Included Variables Per Model

	Seasons	Nino: ClimAdjust	Nino: Anomaly	Year Trend	AR(1)	AR(2)	AR(3-11)	AR(12)	MSE	R Squared
Frequentist Model (poisson) [1]	X	X	X	X					1.686.478	32.97%
Standard Bayesian (poisson)	X	X	X	X					2.343.986	12.51%
Autoregressive Bayesian AR1(poisson)					X				2.503.418	30.33%
Autoregressive Bayesian AR2(poisson)					X	X			2.321.508	14.93%
Autoregressive Bayesian AR1 + 12(poisson)					X			X	42.281.014	9.14%
Autoregressive Bayesian AR12(poisson)					X	X	X	X	575.815.623	3.35%
Autoregressive Bayesian AR2(normal)					X	X			1.810.863	29.54%

As before mentioned the AR2 model is the best one among the models with only autorregressive parts and poisson likelihood, according to the WAIC.

2.4.3 Comparison

Table 1 gives an Overview of the models we investigated and their performance measured as R^2 and Mean Squared Error (MSE). The Frequentist model performs best out of all of them, followed by the normal autoregressive model. It is surprising that the AR(12) model performs so much worse than all the others, according to metrics that do not penalise additional parameters. The trace plot and the prior distribution of the variables looked quite decent, but nevertheless it seems that the inclusion of additional autoregressive covariates hindered the simulation.

2.5 Modelling Forest Fires in all of Brazil

In the previous section we were trying to fit our data to only one Brazilian state. Following we will try to model the number of forest firesnfornallnob Brazil. First we will once again use a Traditional Bayesian approach. Afterwards we try different settings to also model spatial correlation.

2.5.1 Traditional Bayesian

To model Forest Fires in Brazil we aggregated the data so hat we had the number of Fires per Month and year for the entirety of Brazil instead of each State individual. We used the same model as in equation 1, since also here the Variances were exploding from time to time. The fitted values also depict a similar behaviour as the ones of the Traditional Bayesian one-state Model. We can clearly see the differences in seasons, but the model is unable to predict higher counts of fires. All of the predictor Variables are significantly different from zero at a confidence interval of 95%. The Model captures about 56% of the Variance in the data.

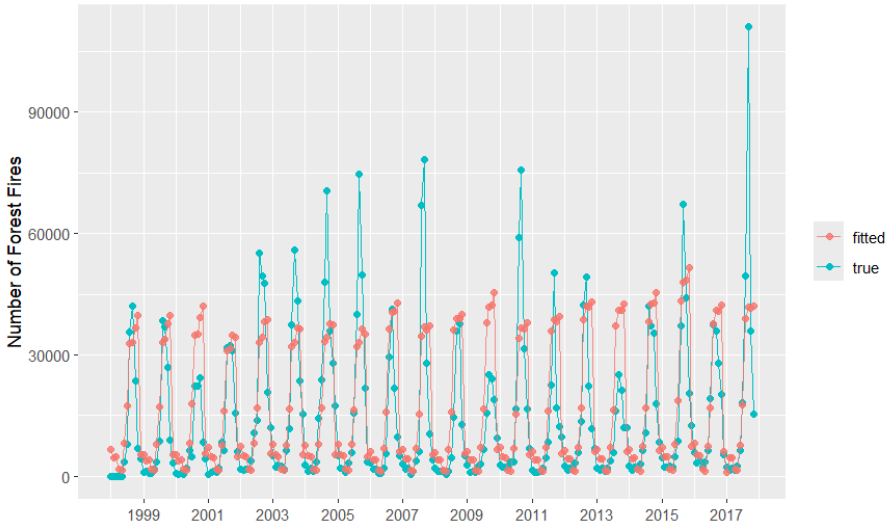

Figure 7: Forest Fires in Brazil: Fitted vs. True values

2.5.2 CAR

CAR, also referred as a Conditional Autoregressive prior, is a technique used to take into account possible unknown and unmeasured important spatially autocorrelated covariates which cannot be accounted for in a regression model, in order to augment it. To do so, they take into account the adjacency structure of the areal units modelled via a set of spatially autocorrelated random effects.

To introduce these random effects we used the library **CARBayes**, which offers a series of functionalities to model both single and multi-valued models as we will show.

2.5.3 Univariate Spatial data model

For this model, the subject study region R is split into K non-overlapping units each with only one response observation, in our case we only focus on the observations of November 2013 in all of the 27 states. Using the function **S.CARleroux()** and choosing a Poisson likelihood model, the given linear mixed model is the following

$$\begin{aligned} Y_k &\sim Pois(\lambda_k) \\ \log(\lambda_k) &= x_k^T\beta + O_k + \Psi_k \\ \beta &\sim N(\mu_\beta, \Sigma_\beta) \\ \psi_k &= \phi_k \\ \phi_k|\phi_{-k}, W, \tau^2, \rho &\sim N\left(\frac{\rho\sum_{i=1}^K w_{ki}\phi_i}{\rho\sum_{i=1}^K w_{ki}+1-\rho}, \frac{\tau^2}{\rho\sum_{i=1}^K w_{ki}+1-\rho}\right) \\ \tau^2 &\sim Inverse-Gamma(a,b) \\ \rho &\sim Uniform(0,1) \end{aligned} \tag{3}$$

Here the vector of regression parameters are denoted by β, O_k is a vector of known spatially lagged covariates and ψ is the spatial structure component that include the set of random effects $\phi = (\phi_1, ..., \phi_k)$ which come from the CAR model. This implementation was proposed by Leroux and allows for the modelling of varying strengths of spatial autocorrelation with only a single set of random effects. The parameter that we are going to focus in our case is ρ, which takes values from 0 to 1 and indicates the level of spatial dependence with 1 being the classic CAR and 0 being total spatial independence.

Since in this dataset in a specific month all the states present identical variables apart from the number of fires itself, we are forced to only take into account random effects for this type of modelling. For this reason we will analyze the impact of the ρ parameter in the modelling: the values we will use are 0,1 and NULL, which lets the model fit it.

In the following table we compare the three models with the WAIC and the MSE once again.

ρ	$WAIC$	MSE
NULL	62401.5	220.6
0	7126.0	193.0
1	21536.3	178.7

As we can see, even though the values are quite high for all three, the model with the lowest WAIC is the model with $\rho = 0$: this stands to indicate that there is probably not a strong spatial dependence between the states. It is important to note that with different runs of the fitting, the WAIC values were highly variable, but the model with $\rho = 0$ had for the most part the lowest WAIC.

The next graphs show the confidence regions of the random effects values for 6 of the states for both the model with $\rho = 0$ and $\rho = 1$. These show how, even though the posterior means are quite similar, the random effects in the second model have higher variance compared to the first one.

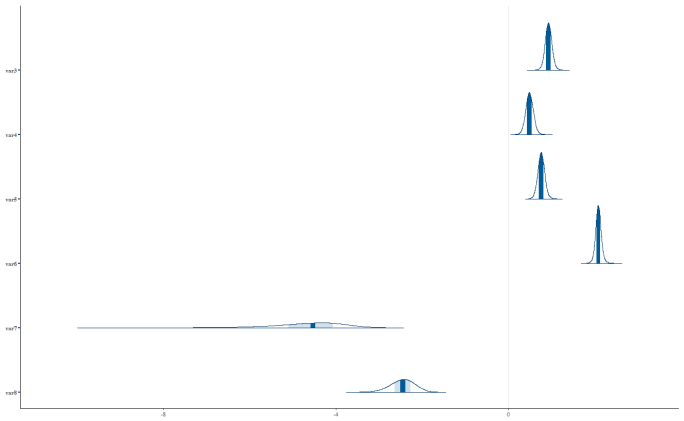

Figure 8: Confidence Intervals of random effects with $\rho = 0$

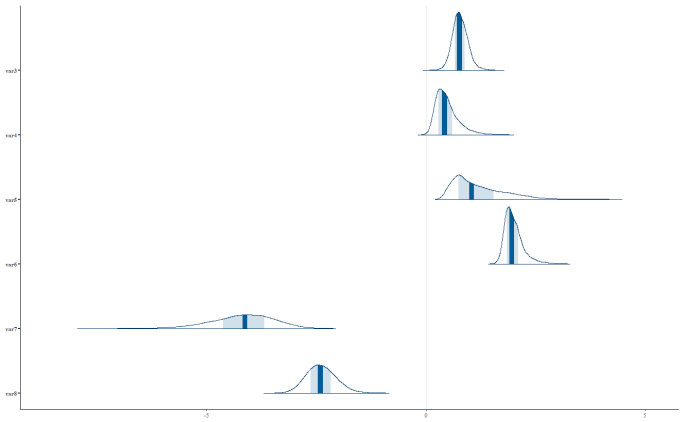

Figure 9: Confidence Intervals of random effects with $\rho = 1$

2.5.4 Multilevel Spatial data model

With this model, like the previous one, the data is still split into K non-overlapping areal units, but now there are m_k observations within area k. That means that there are m_k different response variables being modelled: thanks to this, now the variation is not only available at a spatial level, but also at an individual level. The **S.CARmultilevel** function gives the following linear mixed model:

$$
\begin{aligned}
Y_{kj} &\sim Pois(\lambda_{kj}) \\
\log(\lambda_{kj}) &= x_{kj}^T\beta + O_{kj} + \Psi_{kj} \\
\beta &\sim N(\mu_\beta, \Sigma_\beta) \\
\psi_{kj} &= \phi_k + \zeta_{\lambda(kj)} \\
\phi_k|\phi_{-k} &\sim N\left(\frac{\rho\sum_{j=1}^K w_{kj}\phi_i}{\rho\sum_{j=1}^K w_{kj} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{i=1}^K w_{kj} + 1 - \rho}\right) \\
\zeta_r &\sim N(0, \sigma^2) \quad \text{for all r,} \\
\tau^2, \sigma^2 &\sim InverseGamma(a, b) \\
\rho &\sim Uniform(0, 1)
\end{aligned} \tag{4}
$$

The main difference here is that now the response and each covariate vector is of length $m = \sum_{k=1}^K m_k$. Also, the second term $\zeta_{\lambda(kj)}$ is the random effect that allows for individual level variation.

Since in this case we are able to take into consideration more variables compared to the previous model, we will analyze three different models with different covariates each:

- **CARMultimodel1:** Season + Year Trend + Nino: ClimAdjust + Nino:ANOM
- **CARMultimodel2:** Season
- **CARMultimodel3:** No extra covariates

As was previously mentioned, for this task we will only focus on data from 2005 onward to reduce the time effort need and to reduce the impact of the year trend. We fit these model using 5000 burn-in samples and 45000 actual samples.

To compare these 3 models we will use WAIC, MSE and R2 score. The last one will give us a measure of how well the regression performs.

	$WAIC$	MSE	$R2$
CARMultimodel1	1178787	1036998	0.567
CARMultimodel2	2650328	1137881	0.525
CARMultimodel3	4068402	1967964	0.178

The previous table shows how the model with all the covariates obtains the best values for all three metrics. We will now plot the fitted points against the true values and the residuals in the state of Tocantins as a compraison to the previous models.

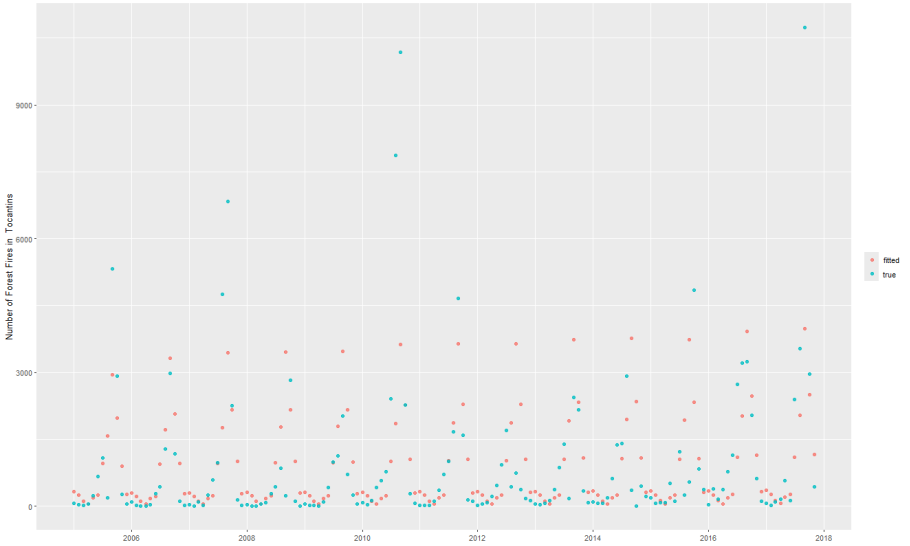

Figure 10: Forest Fires in Tocantins: Fitted vs. True values

The graphs show clearly how, even if the performance of the model is quite strong, it still struggles in modelling the months with very high variance. Figure 12, instead, shows the confidence

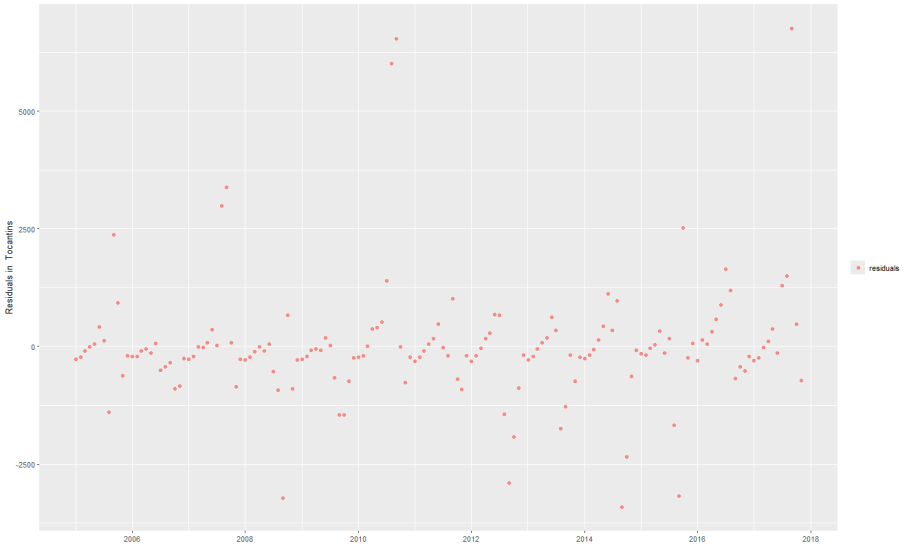

Figure 11: Forest Fires in Brazil: Residuals

intervals for the different covariates: from it we can gather that the variance of the different coefficients is quite low.

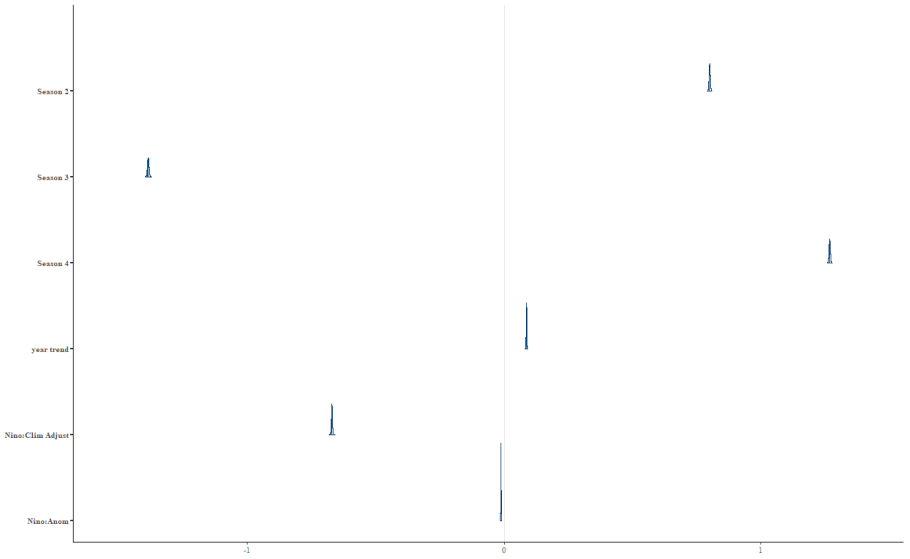

Figure 12: Confidence Intervals for CARMultimodel1

2.5.5 Aggregate CARMultiModel

In conclusion, we aggregated the fitted values of CARMultimodel1 by summing the values of all the states in order to have a model for the whole of Brazil and to be able to compare it to the bayesian version of it.

In this case, the performance is much better compared to the traditional bayesian one as we can see from the next table:

	MSE	$R2$
CARMultimodel1 - aggregated	93796204	0.74
Traditional Bayesian	154773325	0.56

We have to note that, even though both models try to predict the same response variable, the CARMultimodel1 is only fitted on data from 2005 onward; meanwhile the traditional Bayesian one uses the whole dataset. Future analysis should focus on training the CARMultimodel over all of

the data to have a better comparison of the performance.

The following graph shows the fitted data from the aggregated CAR model against the real data.

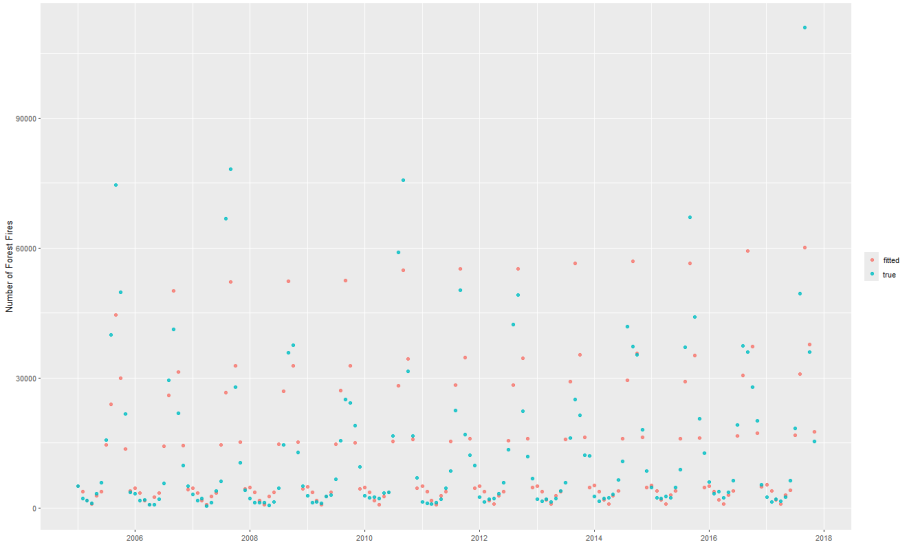

Figure 13: Number of fires in Brazil CARMultimodel1-aggregated: Fitted vs. True Values

3 Conclusion

This dataset presented a very multifaceted challenge even though it had very few variables. In fact, we were able to apply many different modelling techniques with varying characteristics and performances.

Regarding the traditional Bayesian approach, we tried using an Autoregressive model to capture the time dependence in the data and, even though the performances weren't terrible, they also weren't too satisfying. In this case the best model proved to be a Poisson AR(2); at the same time we have seen that the model with normal likelihood performs better on our data then the ones with poisson distribution. It seems that the poisson distribution might not be a perfect fit, and another more flexible distribution that only draws from the realm of positive numbers and preferably models count data should tried. Also we did not perform out of sample prediction. The inflexibility of the poisson distribution might lead to higher errors on in sample prediction but might be better in generalising to out of sample predictions.

Instead, the modelling with spatial autocorrelation proved very adapt for the dataset: even if the single valued model showed a high spatial independence between the states in the random effects, the CARMultimodel1 obtained satisfying performances especially when aggregated to predict the number of fires in Brazil. In this case as well, out of sample prediction was not performed and it could help us better test the actual performance of the model.

Future interesting experiments on the dataset could use a discretization of the Nino:ClimAdjust variable, which could help in modelling the high variance of the data as we saw in the exploration phase. Also, the use of Spatio-Temporal models with CAR priors could prove to also be really effective in order to capture both space and time dependences.

4

References

[ON] National Oceanic and Atmospheric Administration (NOAA). Noaa extended reconstructed sst v5.