Forest Fire Assignment

Giampino Alice - 790347

University of Milano-Bicocca

1 Description of the problem

Forest fires are a major environmental issue, creating economical and ecological damage while endangering human lives. Fast detection is a key element for controlling such phenomenon. To achieve this, one alternative is to use automatic tools based on local sensors, such as provided by meteorological stations. In effect, meteorological conditions (e.g. temperature, wind, relative humidity) are known to influence forest fires and several fire indexes, such as the forest Fire Weather Index (FWI) [2]. Such knowledge is particularly useful for improving firefighting resource management. Forest fire are due to lightnings, human negligence and other causes. Every years millions of forest hectares (ha) are destroyed all around the world, only few months ago dozens of fires erupted in New South Wales, Australia, prompting the government to declare a state of emergency in November 2019; they rapidly spread across all states to become some of the most devastating on record.

We aim to find a model useful to evaluate the probability of fires in different areas and with different climate conditions.

This analysis considers forest fire data from the Montesinho natural park, from the Trás-os-Montes northeast region of Portugal. There are 517 observations, 13 columns and no missing values. The dataset presents the following variables:

1. X: x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y: y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month: month of the year: "jan" to "dec"
4. day: day of the week: "mon" to "sun"
5. FFMC: FFMC index from the FWI system: 18.7 to 96.20
6. DMC: DMC index from the FWI system: 1.1 to 291.3
7. DC: DC index from the FWI system: 7.9 to 860.6
8. ISI: ISI index from the FWI system: 0.0 to 56.10
9. temp: temperature in Celsius degrees: 2.2 to 33.30
10. RH: relative humidity in
11. wind: wind speed in km/h: 0.40 to 9.40
12. rain: outside rain in mm/m2: 0.0 to 6.4
13. area: the burned area of the forest (in ha): 0.00 to 1090.84.

The forest Fire Weather Index (FWI) is the Canadian system for rating fire danger, i.e. it is an indicator of fire intensity. Its components are Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI). FFMC denotes the moisture content surface litter and influences ignition and fire spread, while the DMC and DC represent the moisture content of shallow and deep organic layers, which affect fire intensity. The ISI is a score that correlates with fire velocity spread. Since the response variable (area) is skewed towards 0.0, we have to transform the area distribution in a log scale in order to perform a normal multivariate linear regression.

2 Model specification

The implemented model presents a Normal distribution for the likelihood, with mean equal to the product between β and the model matrix, and the precision distributed as a low informative Gamma (with parameters equal to 0.01). The prior of β is normally distributed with a vector of means set to 0 and covariance-variance matrix that involves the matricial product between the model matrix and precision parameters and the chosen prior; we choose $c2 = g = n$, unit information g-prior that is usually adopted since it has an interpretation of adding prior information equivalent to one data point [8].

```
model{
    # prior
    c2 <- n
    # prior means
    for (j in 1:P){ mu.beta[j] <- 0.0 }
    # calculation of xtx
    for (i in 1:P){ for (j in 1:P){
        inverse.V[i,j] <- inprod( x[,i] , x[,j] )
    }}
    for(i in 1:P){ for (j in 1:P){
        prior.T[i,j] <- inverse.V[i,j] * tau /c2
    }}

    # likelihood

        for (i in 1:n){

                y[i] ~ dnorm( mu[i], tau ) # stochastic componenent
                mu[i] <- inprod( beta[], x[i,] )
        }

    # prior distributions
    beta[1:P] ~ dmnorm( mu.beta[], prior.T[,] )
    tau     ~ dgamma( 0.01, 0.01 )
    s2    <- 1/tau
    s <-sqrt(s2)
}
```

More in details, the model in Table 1 takes $a, b = 0.01$ and $g = n$:

Fitting the full model, we find that only the credible set of two variables do not overlap 0, DMC and the month of December. The convergence is reached for every coefficient looking at $\hat{R}$ (< 1.1). Heidelberger and Welch's convergence diagnostic test confirms that all the chain are stationary, as it can be seen in the traceplot (Appendix), and show that

Likelihood	Prior
$Y_i \sim N(\mu, \sigma^{-2})$	$\mu \vert \sigma^2 \sim N(\mu_0, g\sigma^2)$ $\sigma^2 \sim InvGamma(a, b)$

Table 1: Model Normal-Inverse Gamma (NIG).

some coefficient do not passed the Halfwidth Mean test that removes up to half the chain in order to ensure that the means are estimated from a chain that has converged, but they are coefficients not significant in the model. We set 2 chains, 5000 iterations, 2000 as burning period and adaptation equal to 1000. The DIC (*Deviance Information Criterion*) is a hierarchical modelling generalization of the AIC and the BIC, based on deviance, is equal to 1830 (estimate of expected predictive error), the deviance is 1803.

3 Model selection and model averaging

In order to improve the performance of the model, we compare different methods for variable selection with the aim to find the covariates which are more significant to explain the response variable. The results of the full model suggests to use only the indicator DMC and the month of December as regressors.

One of the methods is based on DIC. First of all, we compare different distributional assumptions to find the optimal set for the model. After this step, we proceed with a stepwise variable selection and at the end we check for other structural components.

Trying different value for the precision and for the prior means, we obtain:

prior means	tau	DIC
0.0	Gamma(0.01, 0.01)	1830
0.0	Gamma(0.001, 0.001)	1830
0.0	Gamma(0.1, 0.1)	1830
0.0	Gamma(1, 1)	1830
0.0	Gamma(10, 10)	1830
empirical prior	Gamma(0.1, 0.1)	1830
empirical prior	Gamma(1, 1)	1831
empirical prior	Gamma(10, 10)	1830
0.0	flat prior	1830

Table 2: DIC with different distributional assumptions.

The choice of the values for the distributions does not affect the results. For this reason we decide to set the model as presented in the previous paragraph.

The stepwise procedure saves the results of DIC for different subset of variables, the results can be seen in Figure 1. As the number of variables increases, the DIC increases. Better results are obtained with few variables where DIC assumes close values for each group. For the parsimony criterion, a model with a maximum of 4 variables perform better.

The last step consists in fitting a model with a random effects due to the fact that our data could be better explained with the introduction of the spatial effects. The DIC for this model is 11546, higher than the fixed effect one.

We perform variable selection also using R package BAS [1] which includes function to accomplish Bayesian Model Averaging in linear models. We compare three methods: Bayesian Information Criteria (BIC), Zellner's g prior where g = n and hyper-g with alpha = 3 combined with a Uniform and a Beta-Binomial model on the model space, for a total

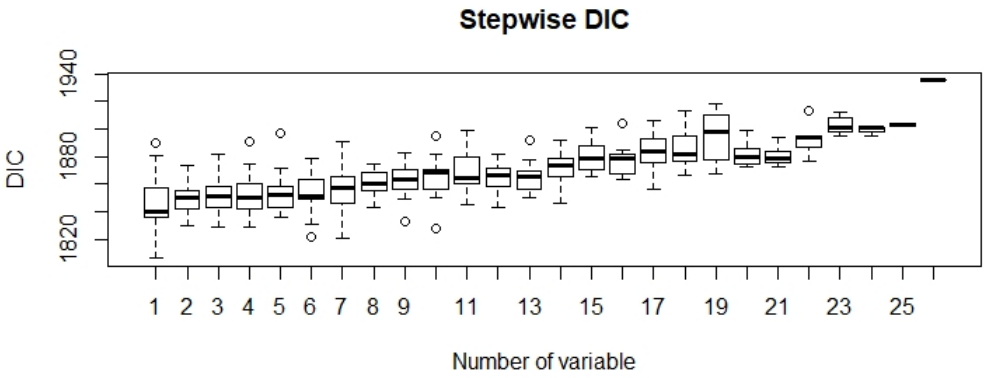

Figure 1: Stepwise DIC.

of 6 different combination. The Tables 3-4-5 resume the results of the selected variables (the intercept is always selected and no reported into the tables). To understand better how the methods select the variables we can look at Figure 2. Only the variables that take the higher marginal posterior probability and that compare in almost all the model are selected. We choose to report variable with a probability of inclusion greater than 0.2 for Bic and g-prior, and 0.5 for hyper-g.

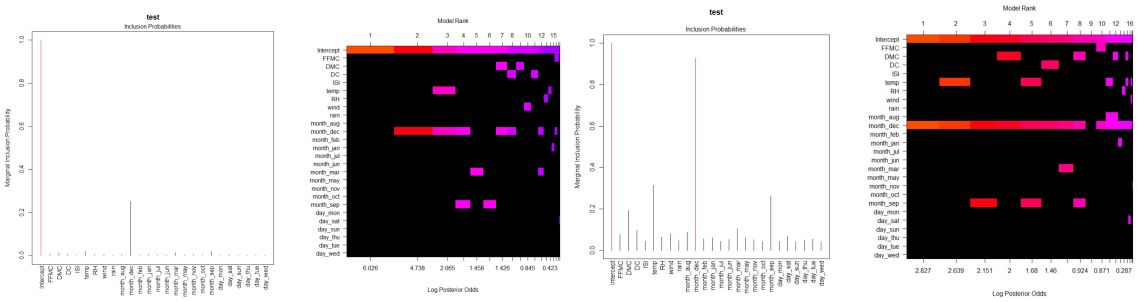

Figure 2: Inclusion probability and image of the variables included in the different model with method BIC and Beta-Binomial model (left) and Uniform model (right) on the model space.

It is clear that for all the combination the month of December has to be included in the final model. The hyper-g prior selects more variables including the indicator DMC, like for the jags model. The MPM is defined as the model consisting of those variables whose marginal posterior probability of inclusion is at least 0.5 and we do not obtain results for all the combination. Moreover, Barbieri and Berger consider a particular class of models for which the MPM is guaranteed to exist [3], but it can happen that it does not select any covariates.

However, due to the correspondence between the selected variables we can assume that our model is not much dependent from the assumption of the used prior. The only exception is the hyper-g prior.

Prior	Uniform	Beta Binomial
BIC	Temperature, December and September	December
g-prior	Temperature, December and September	December
hyper-g	DMC, Temperature, Wind, December and September	DMC, Temperature, December and September

Table 3: BAS variable selection using BMA (*Bayesian Model Averaging*).

Prior	Uniform	Beta Binomial
BIC	December	December
g-prior	Temperature, December and September	December
hyper-g	DMC, Temperature, Wind, December and September	DMC, Temperature, December and September

Table 4: BAS variable selection using MAP (*Maximum A Posteriori*) model.

Prior	Uniform	Beta Binomial
BIC	None	December
g-prior	None	December
hyper-g	None	None

Table 5: BAS variable selection using MPM (*Median Probability Model*).

Another way to conduct variable selection is using the model written in BUGS. To do that we compare 6 different combination of prior and model prior on the model space: unit-information empirical prior, g-prior, hyper-g prior with a Uniform and a Beta Binomial model on the model space. We consider as selected the only variable that do not overlap the 0. As it can be seen in Table 6, with these methods we do not find a high correspondence between the selected variables, even if the temperature and DC compare in 4 different combinations.

Prior	Uniform	Beta Binomial
unit-information empirical	DMC, DC, Temperature, Wind, December, Saturday	DMC, DC, Temperature, Wind, December, September, Saturday
g-prior	DMC, December,	DMC, December
hyper-g	Intercept, December	Intercept

Table 6: BUGS variable selection.

For sake of completeness in Table 7, we report the DIC to show better how these implementation of the model are not the optimal ones. Only the hyper-g prior obtain a result just a little better than the full model.

Prior	Uniform	Beta Binomial
unit-information empirical	1889	1888
g-prior	2071	2069
hyper-g	1818	1828

Table 7: BUGS variable selection, DIC.

Comparing the obtained results with BUGS and BAS, we can assert that the different methodologies capture in a different way the relationship between the response

variable and the covariates. The only corrispondence is for the variable month of December that is chosen in almost all the methods. Other variables that seems to be relevant are Temperature, the month of September and the indicator DMC. For this reason we use one more method for variable selection: in the library BMA [**BMA**] we fit the model and look at the probability of inclusion. We find that the most important variable is the month of December (probne0=97%) and with a lower probability we find the Temperature (probne0=31.1%), month of September (probne0=19.3%) and DMC (probne0=14.2%).

According to the different methodologies, we test models that only includes these variables.

4 Posterior analysis and interpretation of the results

The final model takes into account the variables: Temperature, months of December and September. We choose to include this variable testing different combination of variables and select the combination with the lowest DIC. We use 2 chain and 5000 iterations. The burn-in period is set equal to 2000 and the thin rate is 1.

	mean	sd	2.5%	50%	97.5%	overlap0	f	Rhat	n.eff
Intercept	0.502	0.225	0.062	0.499	0.945	FALSE	0.986	1	6000
Temperature	0.026	0.011	0.003	0.026	0.047	FALSE	0.988	1	6000
December	1.947	0.490	0.981	1.945	2.900	FALSE	1.000	1	5443
September	0.268	0.127	0.027	0.268	0.515	FALSE	0.983	1	6000
Deviance	1797.624	3.129	1793.486	1796.946	1805.370	FALSE	1.000	1	6000

Table 8: Estimates of the final model.

Looking at the output in the Table 8, we can see the point estimate for the posterior mean of every coefficients, the posterior standard deviation and the quantiles for the credible intervals. All the chains are convergent based on Rhat values (all < 1.1). Every variable do not overlap the 0. The variable overlap0 checks if 0 falls in the parameter's 95% credible interval. n.eff is related to the effective sample size.

The variable that contributes more to the response is the month of December, due to the fact that in December the vegetation is dry and it could be possible to have frequent snowfalls. The month of September affects the response because after the summer period we have high temperature and dry vegetation, for these reason the chance of fires increases. The variable temperature is intuitive.

Information about DIC:

```
DIC info: (pD = var(deviance)/2)
pD = 4.9 and DIC = 1802.521
DIC is an estimate of expected predictive error (lower is better).
```

The probabilities of inclusion using BMA (*Bayesian Model Averaging*), MAP (*Maximum-A-Posteriori*) and MPM (*Median Probability Model*) are reported in Table 9.

	BMA	MAP	MPM
Temperature	55%	55%	100%
December	81%	81%	100%
September	48%	48%	100%

Table 9: Probabilities of inclusion.

The results confirm that the most important variable is the month of December. The MAP model and the MPM using jags select only the intercept.

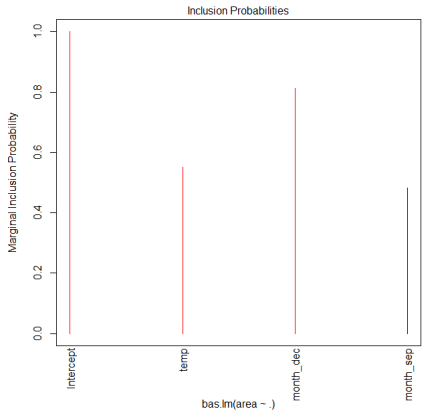

Figure 3: Inclusion probabilities.

However, looking at the representation of the credible intervals (Figure 4) we can understand better that the variable temperature has a coefficient really close to zero, even if it contributes to explaining the response variable. Moreover, all the coefficients have the inferior extreme near to zero, except than for the coefficient related to the month of December.

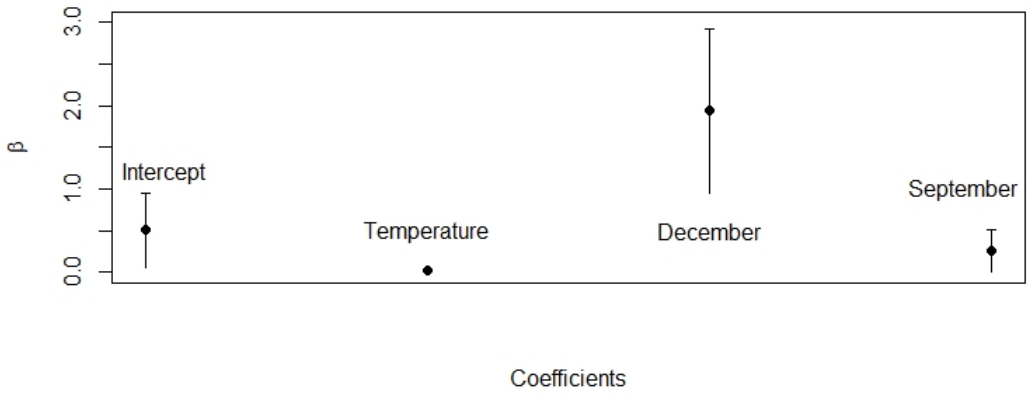

Figure 4: Credible intervals of the coefficients.

Since the final model reaches the optimal result in term of DIC, we check the convergence with diagnostic test and the fit of the model. Heidelberger and Welch's convergence diagnostic test confirms that all the chain are stationary, and show that all coefficient passed the Halfwidth Mean test that removes up to half the chain in order to ensure that the means are estimated from a chain that has converged. The traceplot shows that all the chains are stationary and the distribution quite symmetric (all the plot are in the Appendix). The Geweke test is passed for every parameters. The acf confirms that there is no residual correlation in our model.

Fitting again the model with random effects to capture the spatial number of the data brings to very worse results.

5 Final comments and conclusions

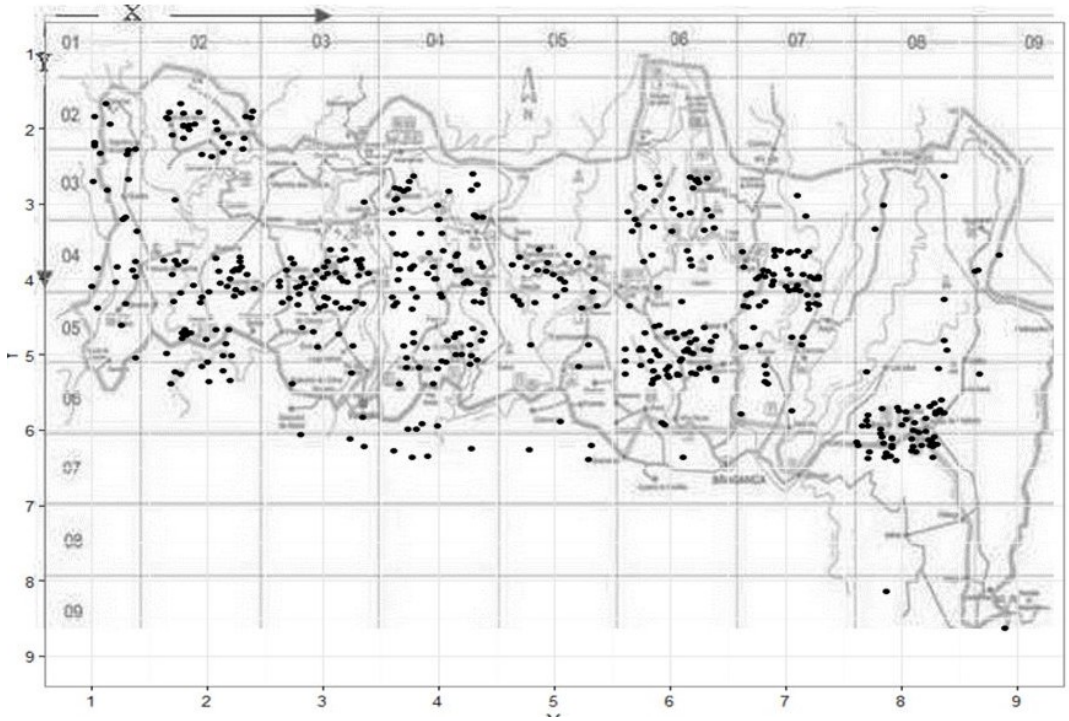

Figure 5: Montesinho Natural Park location of fires [10].

Forest fires cause a significant environmental damage while threatening human lives. In the last two decades, a substantial effort was made to build automatic detection tools that could assist Fire Management Systems (FMS). The three major trends are the use of satellite data, infrared/smoke scanners and local sensors (e.g. meteorological) [2]. In this work we propose a Bayesian approach with the aim of evaluating the area of fires in different places and with different climate conditions. The final model reaches the best results in term of smaller deviance. The variables that emerge to be useful to explain the logarithm of the burned area are the temperature, and if the fire occurs in the month of December or September. The drawback is the lower predictive accuracy for large fires (*Mean Square Error*, MSE=1453).

To improve the proposed model other variables could be insert, like type of vegetation, distance from the nearest fire station, time elapsed before the intervention, firefight strategy (e.g. use of air tankers). Since this model has a high predictive power for small fires (MSE = 6.5), it could be a support for the firefighting resource management.

Seeing that large fires are rare events, an idea is to use a model that classifies the fires, e.g. with the techniques presented here [4], and after this step we could fit the Bayesian model with several adjustment.

Due to the fact that the FWI system is widely used around the world, further research is need to suggest other variables to be inserted in the model and if direct weather conditions are preferable than accumulated values.

A Traceplots and diagnostic tests

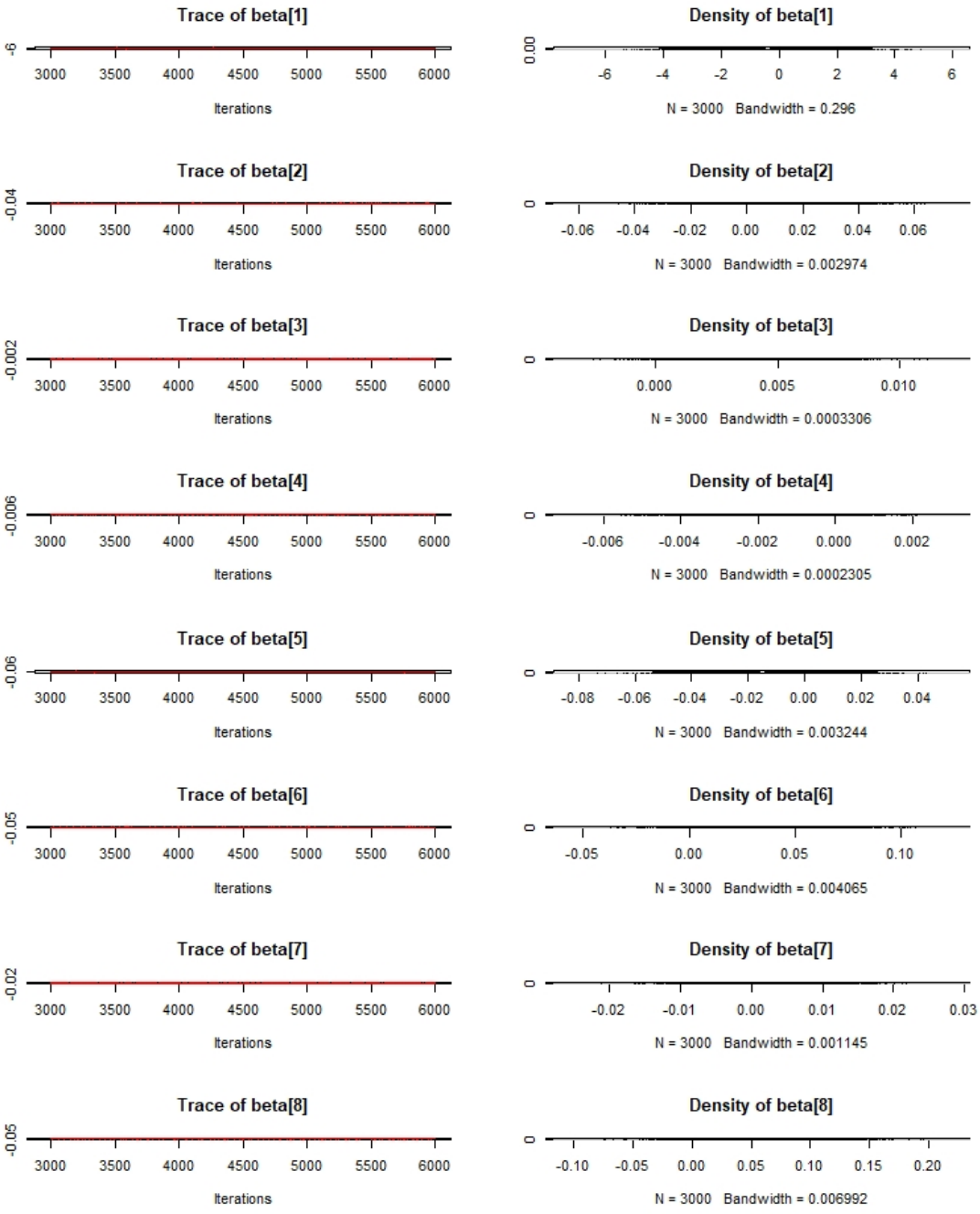

Figure 6: Traceplots and densities of the full model.

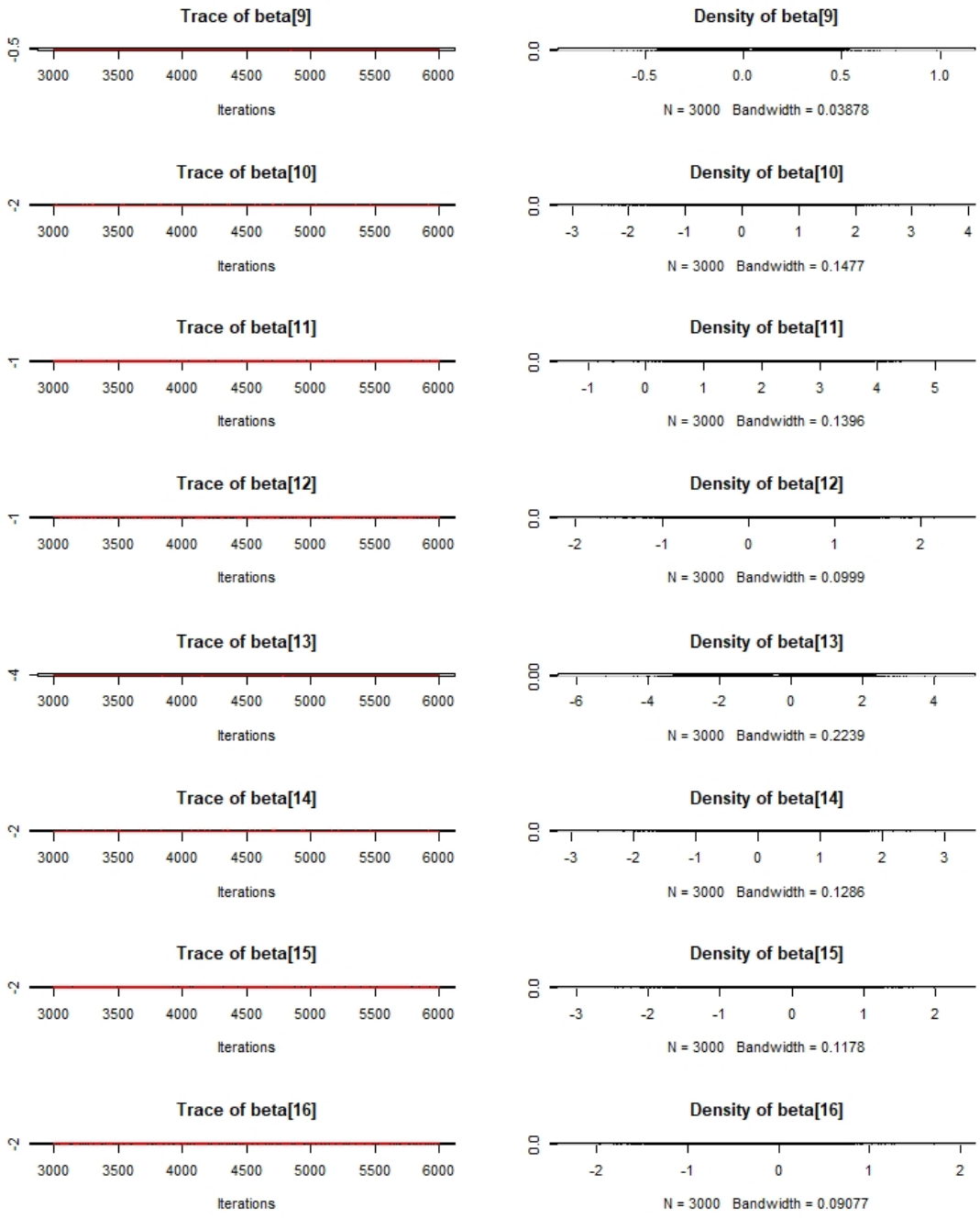

Figure 7: Traceplots and densities of the full model.

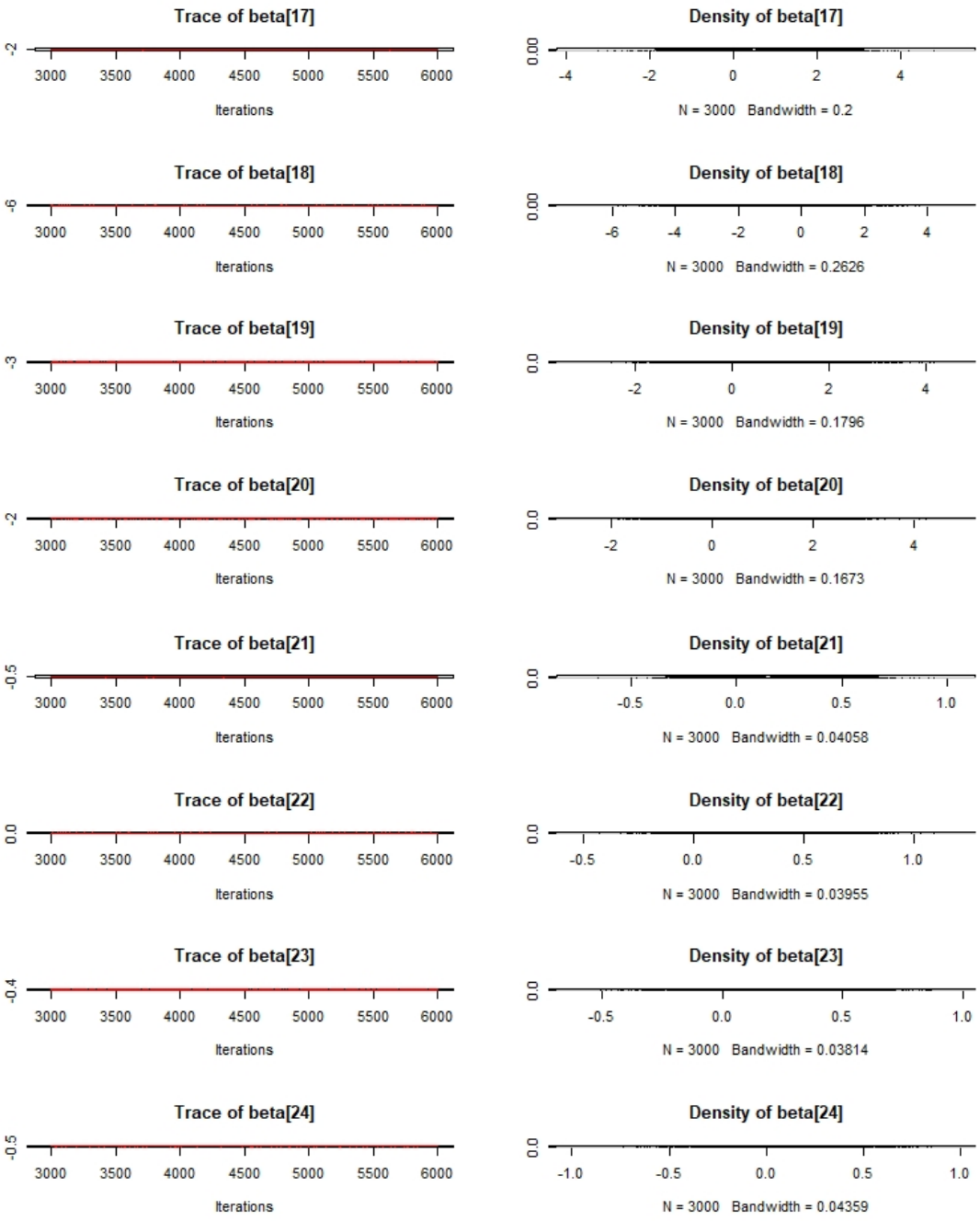

Figure 8: Traceplots and densities of the full model.

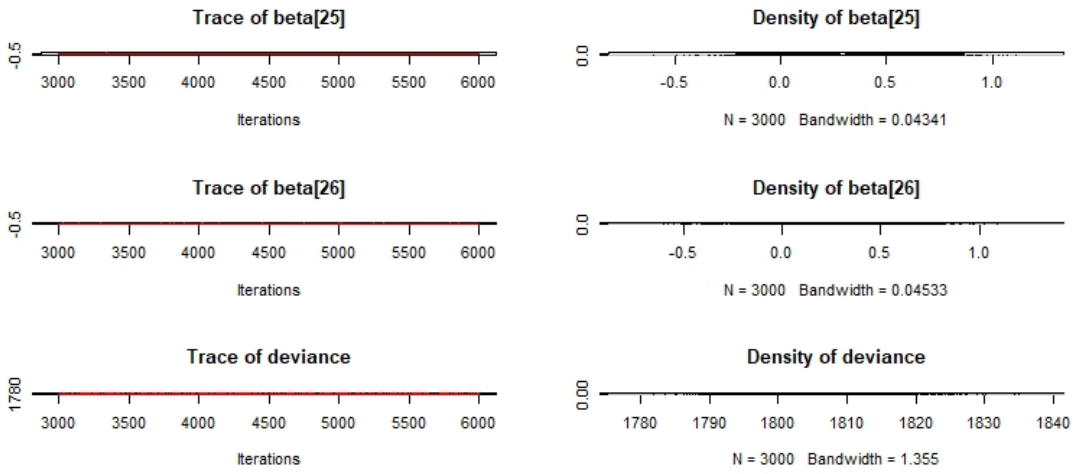

Figure 9: Traceplots and densities of the full model.

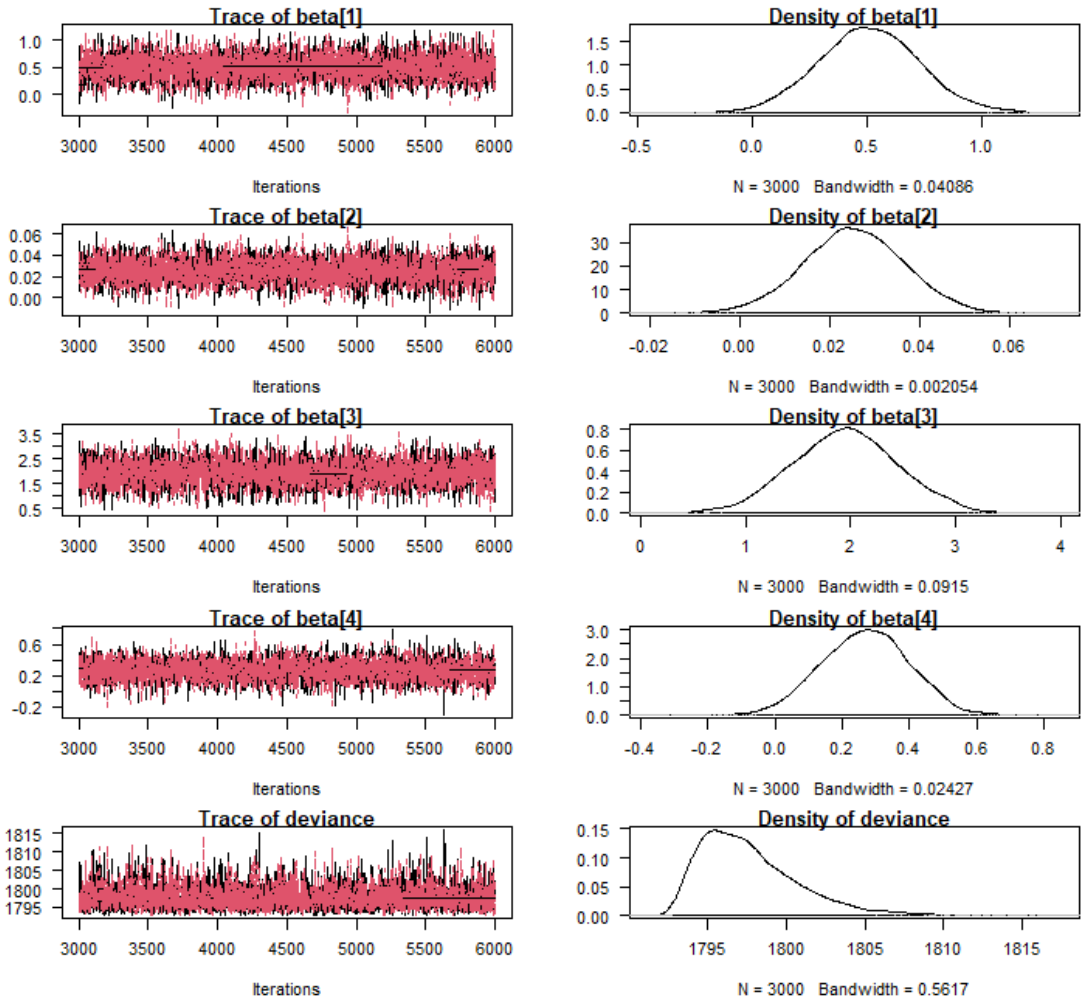

Figure 10: Traceplots and densities of the final model.

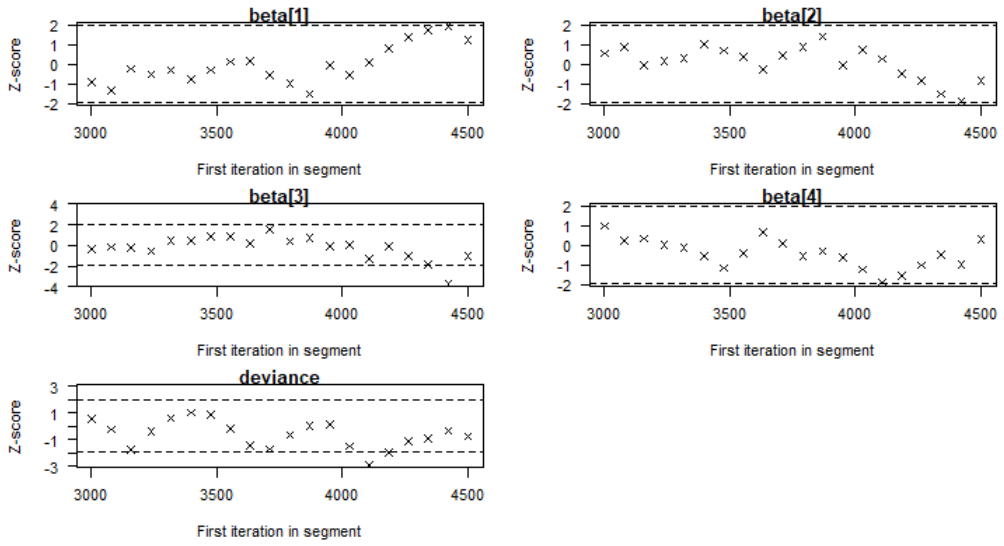

Figure 11: Geweke diagnostic test for the final model.

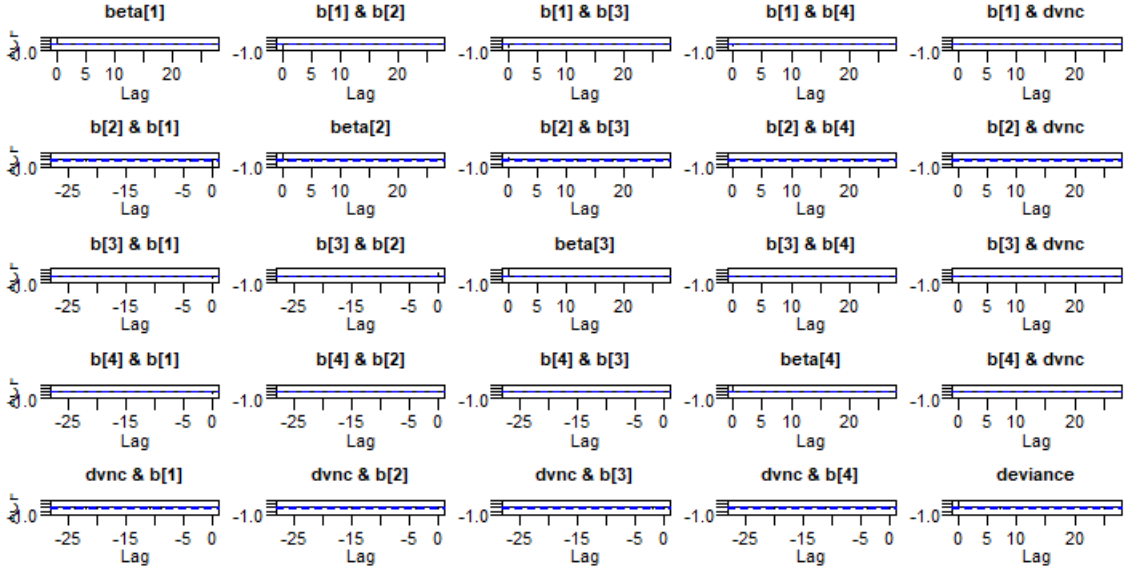

Figure 12: Autocorellation function plot of the final model.

References

[1] Merlise Clyde. *BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling*. R package version 1.5.5. 2020.

[2] Paulo Cortez and Anıbal de Jesus Raimundo Morais. "A data mining approach to predict forest fires using meteorological data". In: (2007).

[3] Joyee Ghosh. "Bayesian model selection using the median probability model". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 7.3 (2015), pp. 185–193.

[4] Somaschini Beatrice Giampino Alice Melograna Federico. *Outlier Detection*. 2019. URL: `https://alicegiampino.github.io/SL/`.

[5] Peter D Hoff. *A first course in Bayesian statistical methods*. Vol. 580. Springer, 2009.

[6] Ken Kellner. *jagsUI: A Wrapper Around 'rjags' to Streamline 'JAGS' Analyses*. R package version 1.5.1. 2019. URL: https://cran.r-project.org/web/packages/jagsUI/index.html.

[7] Matt Denwood Martyn Plummer Alexey Stukalov. *rjags: Bayesian Graphical Models using MCMC*. R package version 4-10. 2019. URL: https://cran.r-project.org/web/packages/rjags/rjags.pdf.

[8] Ioannis Ntzoufras. *Bayesian modeling using WinBUGS*. Vol. 698. John Wiley & Sons, 2011.

[9] Hadley Wickham et al. *dplyr: A Grammar of Data Manipulation*. R package version 0.7.6. 2018. URL: https://CRAN.R-project.org/package=dplyr.

[10] Miaozhi Yu. *Analysis of forest fires*. 2016. URL: https://nycdatascience.com/blog/student-works/analysis-forest-fire-predictors-montesinho-natural-park/.