



## MATH501 Coursework Report

10867436 10886338 10760993 10835879

2024-04-20

### 3.1 Machine Learning Task

#### Machine Learning Part (a):

Data Exploration and Visualization:

- On reviewing the provided dataset(earthquake.txt), we see a total of 37 rows in the given data set, with 0 missing values. This small dataset consists of 3 variables or columns, “type”, “body” and “surface”. The goal of our analysis is to create a model to accurately predict the type based on the body and surface values.

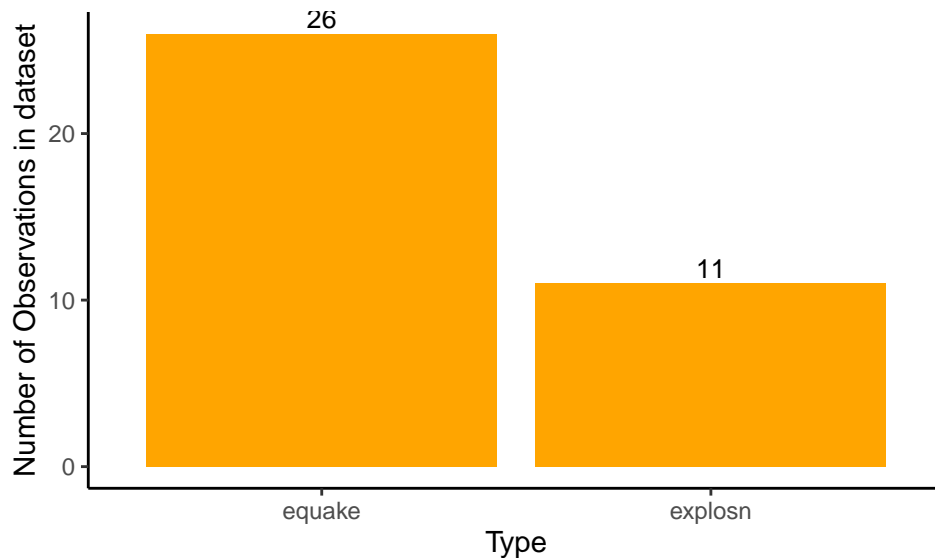


Figure 1: Column Chart of earthquake Data

- From the column chart in Figure 1 we can see that our data is quite imbalanced with 26 observations for ‘earthquakes’ and 11 observations for ‘explosions’.

Table 1: Summary statistics of the earthquake Data

type	mean_body	median_body	sd_body	mean_surface	median_surface	sd_surface
equake	5.4	5.3	0.4	4.9	4.8	0.7
explosn	6.0	5.9	0.2	4.2	4.2	0.2

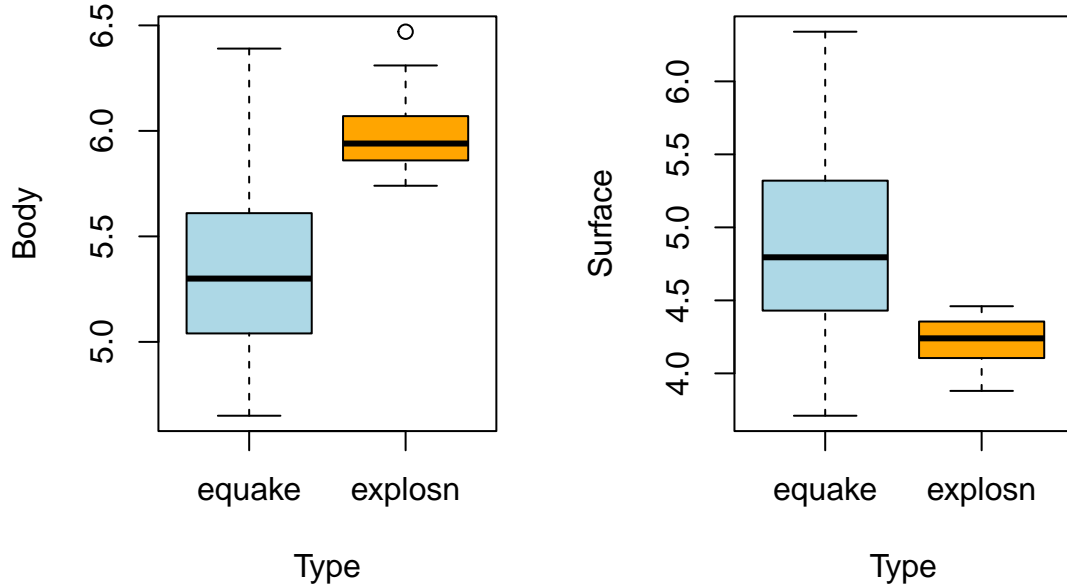


Figure 2: Boxplot of earthquake data

We can see a clear visual distinction between earthquakes and explosions from the distribution of their body values in the boxplots in Figure 2. The boxplots and the summary statistics in Table 1 reveal that explosions tend to have a higher body measures than earthquakes on average. Nonetheless, earthquakes have a wider variability of body values and there is some over lap between the distributions of body measures of earthquakes and explosions. There is also a clear visual distinction between earthquakes and explosions from the distribution of their surface values in the boxplots with explosions having a lower average and median surface value than earthquakes. A similar overlap exists like with the body values.

The scatterplot in Figure 3 highlights these same observations that explosions tend to have higher bodies and lower surface values when compared to earthquakes with some overlap and greater variability with earthquake measures for body and surface. There is a moderate positive correlation between the Surface and Body values of the earthquakes and explosions.

From our sample data, we can infer that both body and surface are properties that can be used effectively to distinguish earthquakes from explosions and will be valuable in creating a classification model.

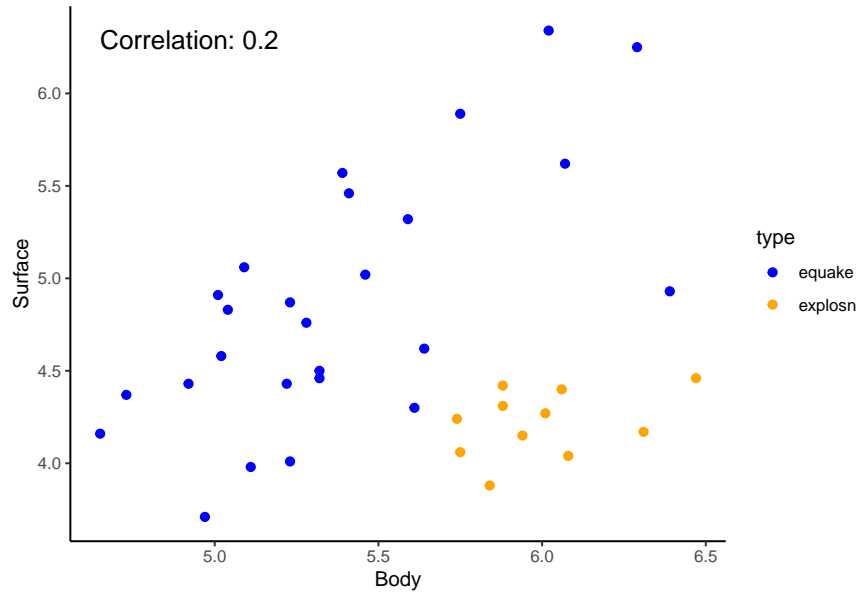


Figure 3: Scatterplot of Body vs Surface

## Machine Learning Part (b):

### Selection of Supervised learning methods

*Justification:* A potential class boundary is revealed by scatter plot which is distinct and potentially distinguishable by both radial and linear class boundaries. Thus a range of classification models could reliably classify these occurrences based on the data provided. Due to the distinct class boundary and considering the mission-critical nature of this context (earthquakes and explosions), a model with interpretability such as decision trees could be beneficial. Hence we decided to primarily explore a simple decision tree classifier and compare it with a KNN classifier. Chosen supervised learning models:

- Decision Trees
- KNN

We split our data in the ratio 75:25, with 75% of our data used for the training of the model.

```
# Preprocessing the Data

n <- nrow(earthquake)
set.seed(1)
train <- sample(n, size = 27) # Indices of observations for training set

## Prepare the two predictors
X <- cbind(earthquake$body, earthquake$surface)

X.train <- X[train, ] # select the data points for the training set
X.test <- X[-train,] # select the data points for the test set

cl.train <- earthquake$type[train] # Class labels for training data
cl.test <- earthquake$type[-train] # Class labels for test data
```

## KNN Classifier:

**KNN Model Tuning:** The minimized validation error after tuning is 0 for a best k value of 3 (See Figure 4)

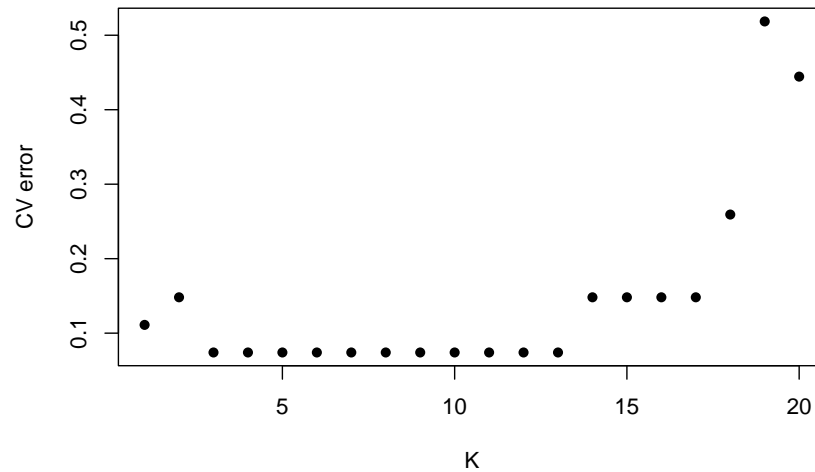


Figure 4: Tuning the K Hyperparameter with Cross validation

**KNN Model Visualisation:** Figure 5 is a visualization of the trained KNN model on an xygrid and highlights the class boundary created by the model.

```
# Evaluating the KNN Model using LOOCV
set.seed(1)

ctrl <- trainControl(method="LOOCV")
knnFit <- train(type ~ body + surface,
  data = earthquake,
  method = "knn",
  #preProcess = c("center", "scale"),
  trControl = ctrl,
  tuneGrid = expand.grid(k = 1:10),
  tuneLength = 10)

knnPredict <- predict(knnFit, newdata = earthquake)

knn.cv.tab <- table(knnPredict, type)
knn_classification_error <- (sum(knn.cv.tab) - sum(diag(knn.cv.tab))) / sum(knn.cv.tab)
```

**KNN Model Evaluation with LOOCV:** The classification error estimate for the KNN model after using LOOCV is 0.054

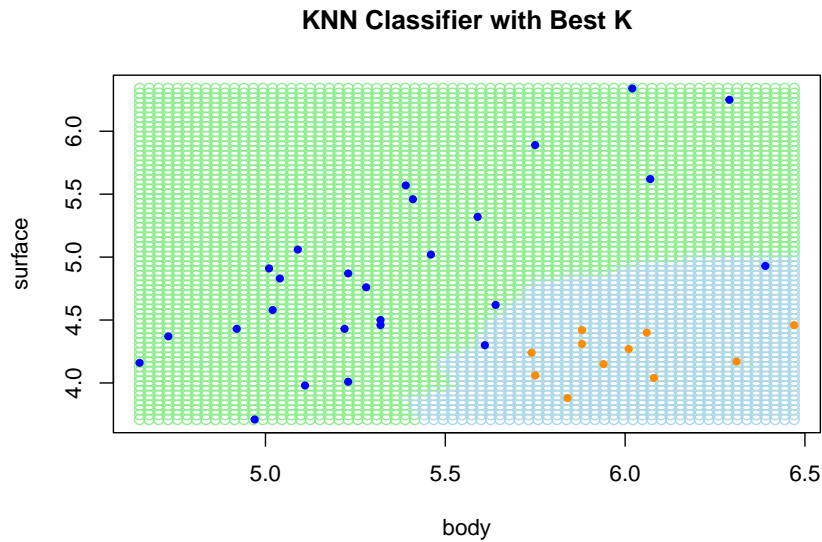


Figure 5: Visualizing the KNN classifier

### Decision Tree Classifier:

#### Decision Tree Definition & Model Tuning:

```
##
## Classification tree:
## tree(formula = type ~ ., data = earthquake, subset = train)
## Number of terminal nodes: 3
## Residual mean deviance: 0.2085 = 5.004 / 24
## Misclassification error rate: 0.03704 = 1 / 27
```

We tuned the tree (See Figure 7) to determine the best size which was 3. Considering our original tree is already of same size, we maintain the original tree. Figure 6 is a depiction of the decision tree rules.

```
##          cl.test
## tree.pred equake explosn
##   equake      9      0
##   explosn      0      1
```

The resulting validation error for our decision tree model is 0

**Decision Tree Visualisation:** Figure 8 is a visualization of the trained Decision Tree model on an xygrid and highlights the class boundary created by the model.

```
#Applying LOOCV to decision tree model
```

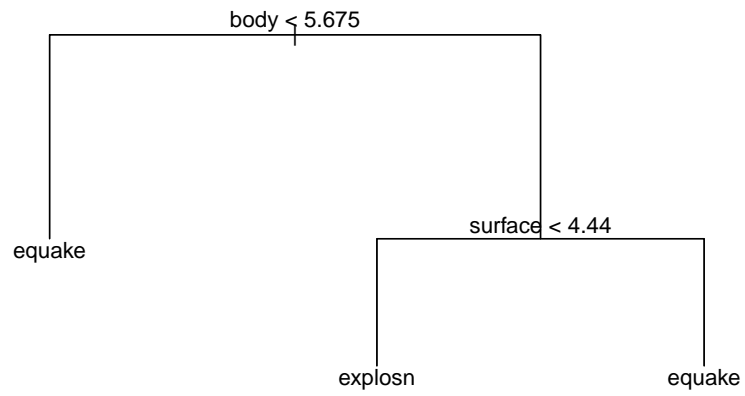


Figure 6: Visualizing the Decision Tree Rules

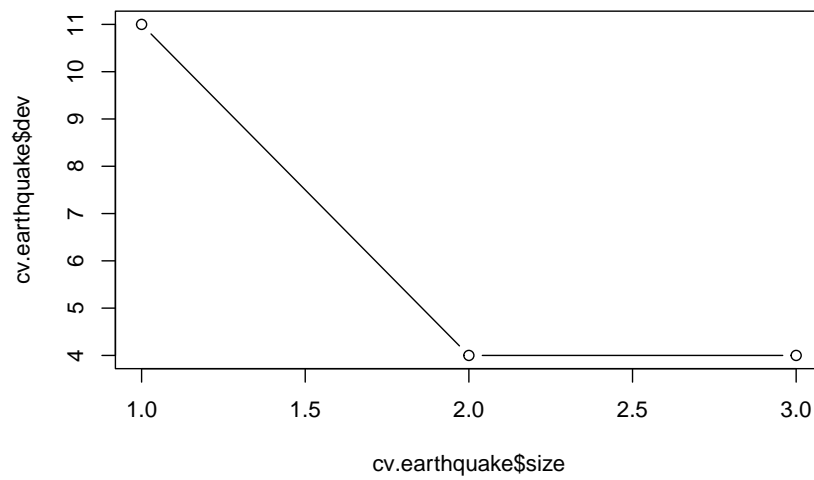


Figure 7: Tuning the Decision Tree using cross-validation

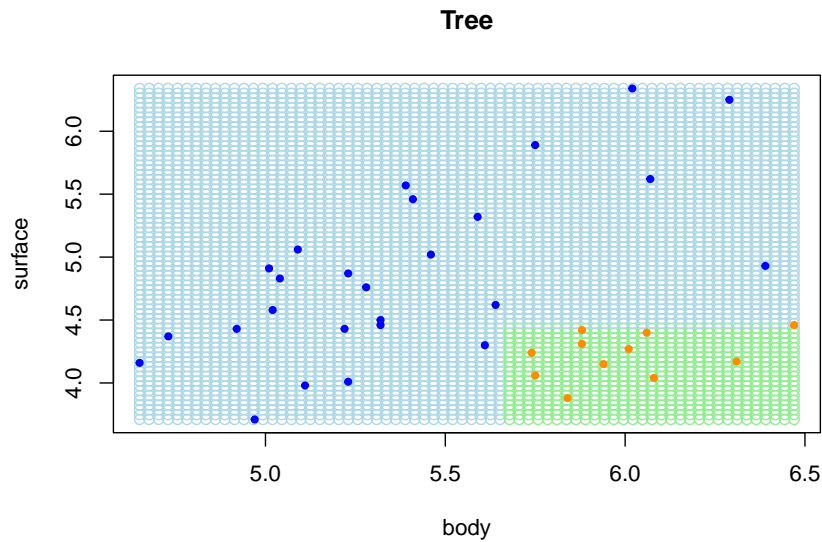


Figure 8: Visualization of Decision Tree grid classification

```
n <- nrow(earthquake) # the number of data points in the data
cv.predictions <- rep('equake', n)

for(i in 1:n) { # start a loop over all data points
  # Fit a classification tree using all data except one data point:
  tree.fit <- tree(type ~ ., data = earthquake[-i, ])

  # Make a prediction for the excluded data point:
  cv.predictions[i] <- predict(tree.fit, newdata = earthquake[i,], type = "class")
}
tree.cv.tab <- table(cv.predictions, type)
cv.error = (tree.cv.tab[1,2] + tree.cv.tab[2,1]) / sum(tree.cv.tab)
```

Decision Tree Model Evaluation with LOOCV:

```
##           type
## cv.predictions equake explosn
##           1      26        0
##           2       0       11
```

The classification error for the decision tree after applying leave-one-out cross validation is 0

Machine Learning Part (c):

Comparing the performance of selected methods

Table 2: Comparing Performance metrics for the Decision Tree and KNN classifiers

Classifier	FP Rate	FN Rate	TP Rate	Test error
Decision Tree	0.0000000	0	1	0.0000000
KNN	0.0769231	0	1	0.0540541

The decision tree has a lower test error rate than the KNN model which is a good indicator of a better model. In addition to this, the critical nature of the potential application of this model makes it very important to minimize false positives as much as possible. An earthquake wrongly detected as an explosion could result in an inadequate response. Lastly the simple interpretability of the decision tree model is also beneficial to this application where quick responses are needed. Hence we will select the decision tree as our preferred model from our analysis.

#### Machine Learning Part (d):

We defined and tuned a k-means model to determine the value of K (number of clusters) that will result in the best within-cluster variability.

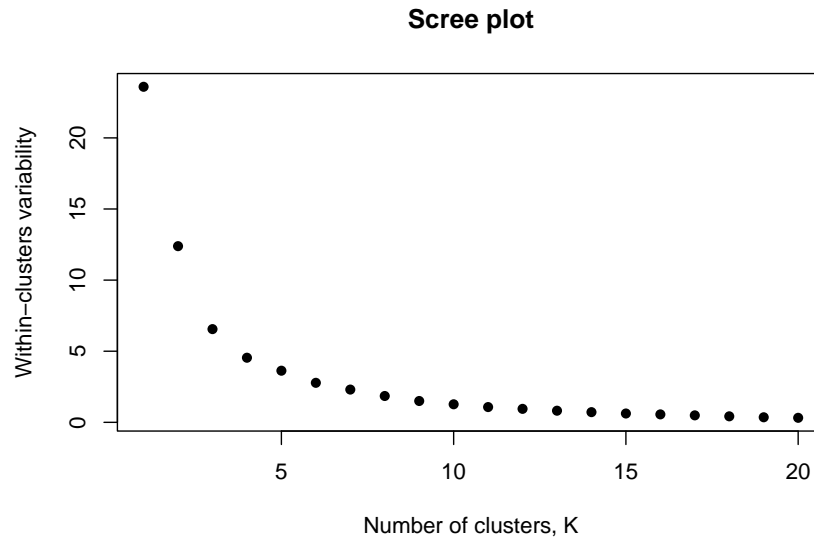


Figure 9: Scree Plot of Within Cluster variability vs K

From the scree plot we see that the number after which there is no longer a significant decrease in within cluster variation is 3. Below is a summary of the classifications of the observations based on the selected 3 clusters model:

```
##
##  1  2  3
## 16 14  7
```

Visualizing the 3 clusters:

The clusters in Figure 10 reveal a similar breakdown to the earlier supervised learning classification models used. Group 3 shows similar characteristics to the explosion class and the original earthquake class can



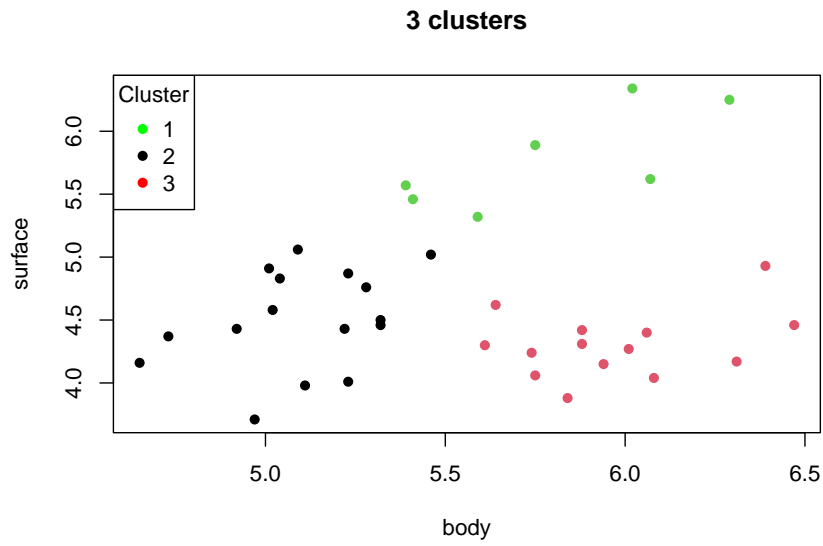


Figure 10: Visualization of the resulting clusters

be likened to a combination of groups 1 and 2. Group 1 seems to be a class of earthquakes characterized by higher values of ‘body’ and ‘surface’ and group “2”, the largest group of all 3, seems to be a class of earthquakes characterized by lower values of ‘body’ and ‘surface’.

## 3.2 Bayesian Statistics Task (with some frequentist analysis)

### 3.2.1 First Sub-Task: Frequentist One-way Analysis of Variance

Four different airlines asked their customers how satisfied they were with the service provided by the airline. For the purpose of our analysis, airline A will be our control group.

Table 3: Summary statistics table for the airline satisfaction scores

airline	Count	Mean	Median	IQR	SD
A	15	4.33	4	1.5	1.35
B	15	5.67	6	2.0	1.63
C	15	4.47	4	2.5	1.68
D	15	6.33	6	3.0	1.80

**Bayesian Statistics Part (a)\*:** Using ggplots to insightfully visualize the data :

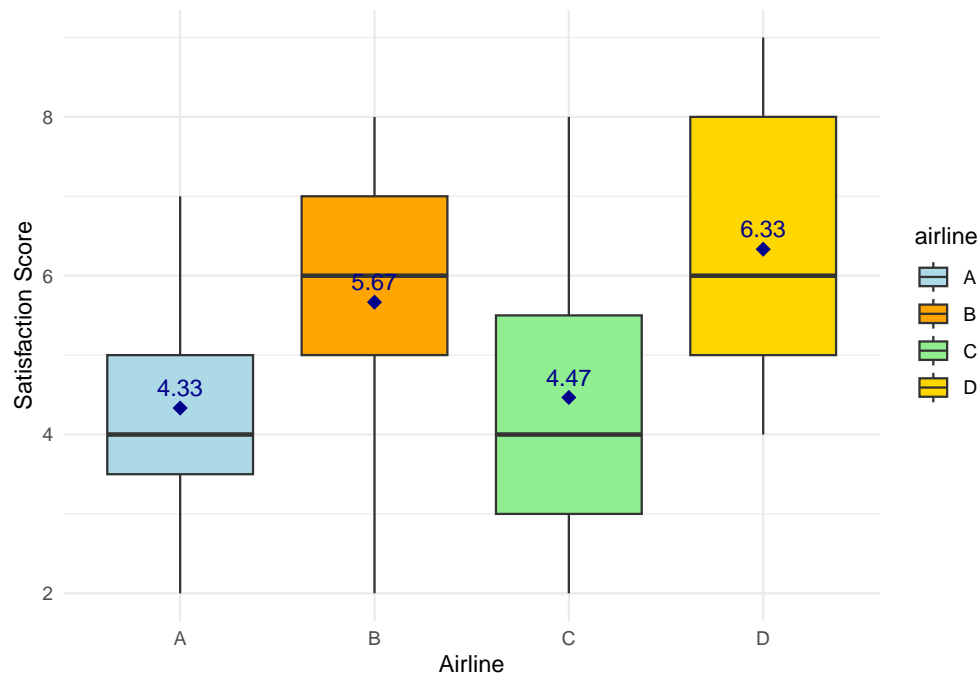


Figure 11: Boxplot of satisfaction scores by airline

#### Discussion:

The number of observations is consistent at 15 satisfaction scores per airline meaning that the comparisons between airlines is not affected by sample size. There are no outliers in the dataset. From the boxplots we see that Airline D has the highest mean satisfaction score (approximately 6.33), indicating that customers tend to rate it more favorably on average. It also has the highest variability as seen in its standard deviation and inter-quartile range. Airline B has the same median value as D, with a lower mean satisfaction score and smaller variability. Airlines A and C have the same median value and similar mean satisfaction scores but airline C has a larger variability.

#### Conclusion

Our preliminary inference is that there is likely no statistically significant difference in satisfaction scores between airlines B and D and also between A and C. A statistical test can further confirm or refute these preliminary inferences.

**Bayesian Statistics Part (b)\*:** Given  $y_{i,j}$  be the score given by the  $j$ -th customer using the  $i$ -th airline, with  $i = 1, \dots, 4$  and  $j = 1, \dots, 15$ . The following one-way Analysis of Variance model has been suggested for these data:

$$y_{i,j} \sim N(\mu_{i,j}, \sigma^2), \quad i = 1, \dots, 4, \quad j = 1, \dots, 15$$

where

$$\begin{aligned} \mu_{1,j} &= \mu_1, & j &= 1, \dots, 15 \\ \mu_{2,j} &= \mu_1 + \alpha_2, & j &= 1, \dots, 15 \\ \mu_{3,j} &= \mu_1 + \alpha_3, & j &= 1, \dots, 15 \\ \mu_{4,j} &= \mu_1 + \alpha_4, & j &= 1, \dots, 15 \end{aligned}$$

#### Description of $\alpha_4$ :

This One-Way ANOVA model uses  $\mu_1$  the average satisfaction score for airline A as a “base” group and the mean satisfaction scores of airlines B, C and D ( $\mu_2$ ,  $\mu_3$  and  $\mu_4$  respectively) are described in terms of the mean score A plus or minus constants  $\alpha_2$ ,  $\alpha_3$  and  $\alpha_4$ .  $\alpha_4$  is thus an estimate of the difference between the average satisfaction score of airline D and airline A. Airline A is selected arbitrarily as the base group for comparisons, but in a more general sense,  $\mu_1$  could represent the mean satisfaction score of any of the airlines depending on the goal of the analysis, the questions we seek to answer or the statistical inferences we seek.

**Bayesian Statistics Part (c):** Fitting the model in the frequentist framework we get the following outcome:

```
##
## Call:
## lm(formula = satisfactionscore ~ airline, data = airline_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.6667 -1.3333 -0.3333  1.3333  3.5333
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.3333     0.4194  10.333 1.38e-14 ***
## airlineB      1.3333     0.5931   2.248  0.02852 *
## airlineC      0.1333     0.5931   0.225  0.82294
## airlineD      2.0000     0.5931   3.372  0.00136 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.624 on 56 degrees of freedom
## Multiple R-squared:  0.2208, Adjusted R-squared:  0.1791
## F-statistic:  5.29 on 3 and 56 DF,  p-value: 0.00278
```

Inferences on  $\mu_1$ ,  $\alpha_2$ ,  $\alpha_3$  and  $\alpha_4$ :

- $\mu_1 = 4.3333$ : This is the estimated average satisfaction score for Airline A. It represents the mean of the reference category and is the baseline mean satisfaction score against which the other airlines are compared.
- $\alpha_2 = 1.3333$ : This is the estimated difference in the average satisfaction score between Airline B and the reference Airline A. It suggests that on average, customers rate Airline B 1.3333 points higher than Airline A. The p-value suggests that this difference is statistically significant making it a probable estimate of true difference in reality.
- $\alpha_3 = 0.1333$ : This coefficient represents the difference in average satisfaction scores between Airline C and Airline A. Since this coefficient is not statistically significant (p-value  $0.82294 > 0.05$ ), we do not have evidence to conclude that there is a true difference in average satisfaction scores between these airlines.
- $\alpha_4 = 2.0000$ : This is the estimated difference in average satisfaction scores between Airline D and Airline A. Airline D scores, on average, 2 points higher than Airline A. This effect is statistically significant at the 0.01 level which suggests that Airline D is rated significantly higher in satisfaction than Airline A.
- The model's F-statistic and its p-value indicate that there is a significant effect of airline on satisfaction scores. The adjusted R-squared value tells us that around 17.91% of the variability in satisfaction scores is explained by which airline the customers flew with.

Given  $\mu_1, \mu_2, \mu_3$  and  $\mu_4$  be the average satisfaction scores for airlines A,B,C and D from our data, we will use an ANOVA test to test possible differences between the average satisfaction scores for the 4 airlines.

Hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

Performing a frequentist hypothesis test of size 0.05 of whether the satisfaction score is different for each of the airlines, we get the following results:

```
## Analysis of Variance Table
##
## Response: satisfactionscore
##           Df Sum Sq Mean Sq F value Pr(>F)
## airline    3  41.867  13.9556    5.29 0.00278 **
## Residuals 56  147.733   2.6381
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test conducted to compare the satisfaction scores across the four airlines resulted in a significant F-statistic (F value = 5.29) and a p-value of 0.00278, which is below the alpha level/test size of 0.05. This affirms that the satisfaction scores are statistically different across at least one pair of the airlines. Hence, we reject the null hypothesis that all airlines have the same average satisfaction score. It is sufficient evidence to suggest that not all airlines have equal levels of customer satisfaction.

**Bayesian Statistics Part (d):** As a follow up to the ANOVA test, the Tukey test will estimate the pairwise differences in average satisfaction score between the groups. We will test 6 pairs of hypotheses together simultaneously. The test is modelled as follows:

$$\begin{aligned}
H_0 : \mu_2 - \mu_1 &= 0 & (\text{or, } \mu_1 = \mu_2) \\
H_1 : \mu_2 - \mu_1 &\neq 0 & (\text{or, } \mu_1 \neq \mu_2) \\
H_0 : \mu_3 - \mu_1 &= 0 & (\text{or, } \mu_1 = \mu_3) \\
H_1 : \mu_3 - \mu_1 &\neq 0 & (\text{or, } \mu_1 \neq \mu_3) \\
H_0 : \mu_4 - \mu_1 &= 0 & (\text{or, } \mu_1 = \mu_4) \\
H_1 : \mu_4 - \mu_1 &\neq 0 & (\text{or, } \mu_1 \neq \mu_4) \\
H_0 : \mu_3 - \mu_2 &= 0 & (\text{or, } \mu_2 = \mu_3) \\
H_1 : \mu_3 - \mu_2 &\neq 0 & (\text{or, } \mu_2 \neq \mu_3) \\
H_0 : \mu_4 - \mu_2 &= 0 & (\text{or, } \mu_2 = \mu_4) \\
H_1 : \mu_4 - \mu_2 &\neq 0 & (\text{or, } \mu_2 \neq \mu_4) \\
H_0 : \mu_4 - \mu_3 &= 0 & (\text{or, } \mu_3 = \mu_4) \\
H_1 : \mu_4 - \mu_3 &\neq 0 & (\text{or, } \mu_3 \neq \mu_4)
\end{aligned}$$

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = satisfactionscore ~ airline, data = airline_data)
##
## $airline
##          diff          lwr          upr          p adj
## B-A  1.3333333 -0.2370805  2.9037471  0.1229720
## C-A  0.1333333 -1.4370805  1.7037471  0.9959461
## D-A  2.0000000  0.4295862  3.5704138  0.0072205
## C-B -1.2000000 -2.7704138  0.3704138  0.1917762
## D-B  0.6666667 -0.9037471  2.2370805  0.6763548
## D-C  1.8666667  0.2962529  3.4370805  0.0136641
```

- B vs. A: Difference of 1.33 points ( $p = 0.122972$ ). No significant difference in satisfaction scores between Airline B and A.
- C vs. A: Difference of 0.13 points ( $p = 0.995946$ ). No significant difference in satisfaction scores between Airline C and A.
- D vs. A: Difference of 2.00 points ( $p = 0.007221$ ). Airline D has a significantly higher satisfaction score compared to Airline A.
- C vs. B: Difference of -1.20 points ( $p = 0.191776$ ). No significant difference in satisfaction scores between Airline C and B.
- D vs. B: Difference of 0.67 points ( $p = 0.67635$ ). No significant difference in satisfaction scores between Airline D and B.
- D vs. C: Difference of 1.87 points ( $p = 0.013664$ ). Airline D has a significantly higher satisfaction score compared to Airline C.

**Bayesian Statistics Part (e):** Is the satisfaction score for airline D more than 3 points higher than the average satisfaction score for airline B and C?

To answer this question, we formulated our hypothesis as follows:

$$\begin{aligned}
H_0 : \mu_4 - \mu_2 &\leq 3 & : \mu_2(\text{airline B}) \text{ is not more than 3 points lower than } \mu_4(\text{airline D}) \\
H_1 : \mu_4 - \mu_2 &> 3 & : \mu_2(\text{airline B}) \text{ is lower than } \mu_4(\text{airline D}) \text{ by more than 3 points} \\
H_0 : \mu_4 - \mu_3 &\leq 3 & : \mu_3 \text{ is not more than 3 points lower than } \mu_4 \\
H_1 : \mu_4 - \mu_3 &> 3 & : \mu_3(\text{airline C}) \text{ is lower than } \mu_4(\text{airline D}) \text{ by more than 3 points}
\end{aligned}$$

```
##
## Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = satisfactionscore ~ airline - 1, data = airline_data)
##
## Linear Hypotheses:
##               Estimate Std. Error t value Pr(>t)
## airlineD - airlineB <= 3    0.6667     0.5931  -3.934  1.000
## airlineD - airlineC <= 3    1.8667     0.5931  -1.911  0.994
## (Adjusted p values reported -- single-step method)
```

Based on these results, we conclude that :

- Airline D vs. Airline B: The satisfaction score for Airline D is estimated to be 0.6667 points higher than for Airline B, with a p-value of 1.000, indicating no significant evidence that Airline D's score is more than 3 points higher.
- Airline D vs. Airline C: The satisfaction score for Airline D is estimated to be 1.8667 points higher than for Airline C, with a p-value of 0.994, also indicating no significant evidence that Airline D's score is more than 3 points higher.

### 3.2.2 Second Sub-Task: Bayesian Two-ways Analysis of Variance

**Bayesian Statistics Part (f):** A farmer wants to test the level of carbon sequestration in their fields. There are five possible techniques to capture carbon: T1, T2, T3, T4 and T5. The farmer suspected that there may be variation because of slight differences in the locations of the fields. To allow for this the five possible types of treatment were used on each of three different fields.

**Data Provided:**

Field	Type of treatment				
	T1	T2	T3	T4	T5
1	208	216	220	226	209
2	194	212	218	239	224
3	199	211	227	227	221

Following from the suggested Bayesian two-way ANOVA model, we will consider Field 1  $\alpha_1$  and Treatment 1  $\beta_1$  as our baselines for comparison.

```
## Compiling model graph
##   Resolving undeclared variables
##   Allocating nodes
## Graph information:
##   Observed stochastic nodes: 15
##   Unobserved stochastic nodes: 8
##   Total graph size: 46
##
## Initializing model

## Inference for Bugs model at "C:/Users/OGHENE~1/AppData/Local/Temp/RtmpWI5mzu/model34c0773814c3", fit
## 3 chains, each with 1e+05 iterations (first 50000 discarded), n.thin = 50
## n.sims = 3000 iterations saved
```

```
##          mu.vect sd.vect  2.5%    50%   97.5%  Rhat n.eff
## alpha[1]   0.000  0.000   0.000   0.000   0.000 1.000    1
## alpha[2]   1.733  5.039  -8.490   1.729  11.315 1.001   3000
## alpha[3]   1.322  5.104  -9.021   1.287  10.972 1.001   2600
## beta[1]    0.000  0.000   0.000   0.000   0.000 1.000    1
## beta[2]   12.832  6.528  -0.360  12.950  25.443 1.002   1200
## beta[3]   21.662  6.599   8.345  21.627  34.803 1.002   1900
## beta[4]   30.670  6.391  18.319  30.500  43.697 1.002   1100
## beta[5]   18.045  6.435   5.276  18.185  30.957 1.002   1600
## m         198.975  5.327 188.642 199.003 209.304 1.003   1600
## sigma      7.650  2.232   4.642   7.249  12.945 1.001   2100
## tau        0.021  0.011   0.006   0.019   0.046 1.001   2100
## deviance 102.523   5.831  94.006 101.677 117.034 1.001   3000
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = var(deviance)/2)
## pD = 17.0 and DIC = 119.5
## DIC is an estimate of expected predictive error (lower deviance is better).
```

#### Model Discussions:

- $\alpha$  (Effects of Fields on Total Carbon Content):
- $\alpha_1$  (Field 1): Set as the baseline with no deviation from the global mean ( $\mu$ ), effectively -  $\alpha_1 = 0.000$ .
- $\alpha_2$  (Field 2): Estimate of 1.733 with a credible interval of (-8.490, 11.315). This interval includes zero, suggesting uncertainty about Field 2's impact on carbon content relative to Field 1.
- $\alpha_3$  (Field 3): Estimate of 1.322 with a credible interval of (-9.021, 10.972), also including zero. This indicates similar uncertainty as Field 2 in comparison to Field 1.
- $\beta$  (Effects of Treatments on Total Carbon Content):
- $\beta_1$  (Treatment 1): Set as the baseline,  $\beta_1 = 0.000$ .
- $\beta_2$  (Treatment 2): Estimate of 12.832 with a credible interval (-0.360, 25.443), narrowly excluding zero, suggesting a potentially significant positive effect.
- $\beta_3$  (Treatment 3): Estimate of 21.662 with a credible interval (8.345, 34.803), indicating a significant positive effect.
- $\beta_4$  (Treatment 4): Estimate of 30.670 with a credible interval (18.319, 43.697), confirming a significant increase in carbon content.
- $\beta_5$  (Treatment 5): Estimate of 18.045 with a credible interval (5.276, 30.957), showing a significant positive effect, albeit with some uncertainty.
- $\mu$  (Global Mean Carbon Content): The global mean carbon content across all treatments and fields is estimated at 198.975, with a narrow credible interval (188.642, 209.304), reflecting high certainty about the overall average.
- $\sigma$  and  $\tau$ :
- $\sigma$ : Posterior mean standard deviation of 7.650, reflecting the variability in carbon content among the treatments and fields.

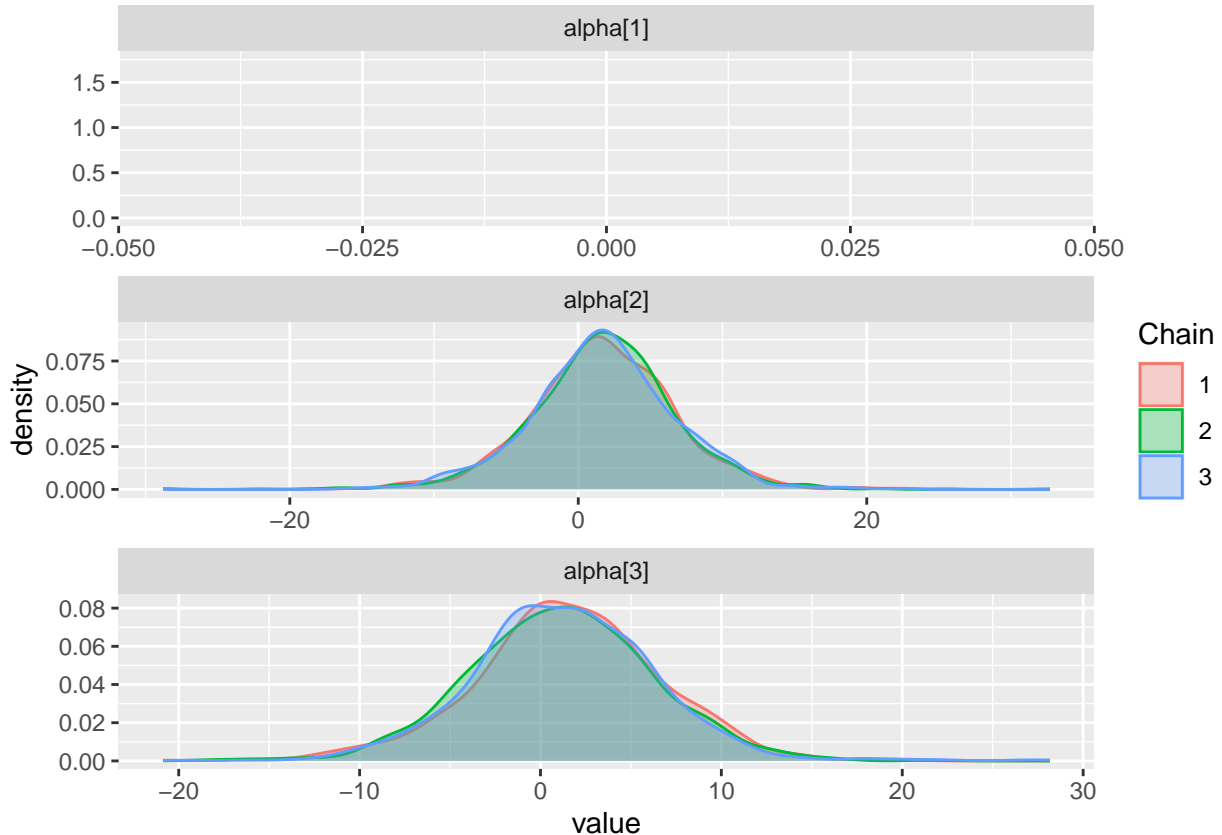
- $\tau$ : The precision (inverse of variance) of the measurements is estimated at 0.021, with a credible interval (0.006, 0.046).
- Deviance: Provides a measure of model fit; the mean deviance is 102.523.
- DIC: The Deviance Information Criterion is computed as 119.5, aiding in model comparison; lower values indicate a better model fit.

### Conclusions:

There is significant uncertainty regarding whether Fields 2 and 3 differ from Field 1 in terms of carbon content. Their credible intervals contain zero, and there's high variability in their estimates. All the treatments alternative treatments significantly increase the carbon content compared to the baseline, with Treatment 4 showing the strongest effect. The credible intervals for these treatments do not contain zero except for treatment 2 for which a minor part of it's credible interval falls below 0. This reinforces the significant positive effects of the treatments on total carbon content. The model effectively captures the variability in carbon content across different fields and treatments, as indicated by the specific estimates and credible intervals. The DIC value provides a basis for model comparison if alternative models are considered (like the model in the third sub-task).

**Bayesian Statistics Part (g)\*\*:** Including a graphical representation of the trace plots and posterior densities of  $\alpha_i$  and  $\beta_j$ :

$\alpha_2$  and  $\alpha_3$  both have broad, bell-shaped distributions, which suggests some uncertainty in these estimates. Both distributions are centered around zero, indicating that Fields 2 and 3 are estimated to have a minimal effect on carbon sequestration compared to the baseline Field 1. The spread of the distributions is worth highlighting, especially for  $\alpha_3$  which could indicate a larger variance in the effect or less certainty in the estimate.

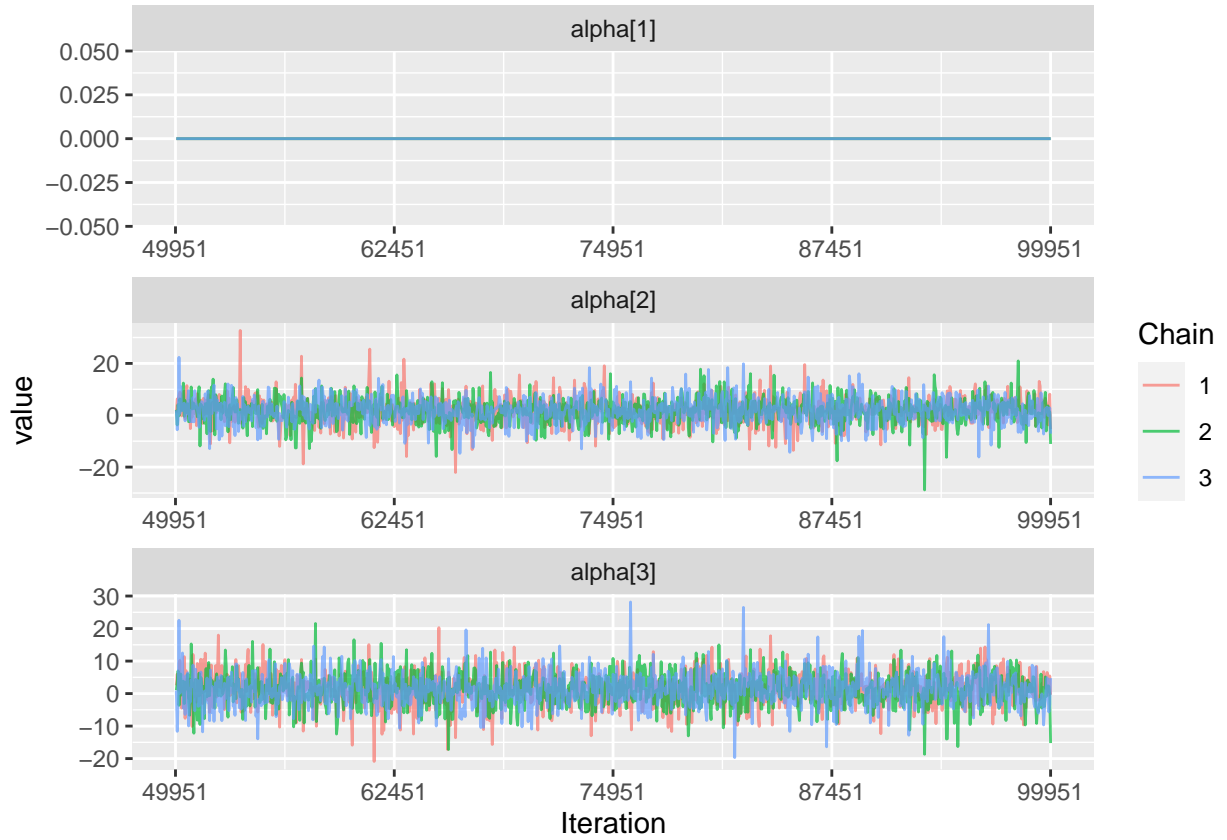




- $\alpha_1$ : The traceplot shows all three chains fluctuating closely around zero without trending away, which suggests that the sampler has converged for this parameter.
- $\alpha_2$  and  $\alpha_3$ : The traceplots of these parameters display stable, horizontal bands, with the chains mixing well and showing no signs of drift or “walking” behavior. This indicates good convergence and suggests that the sampler is reliably exploring the posterior space of these parameters. There are no divergent patterns, and the chains overlap considerably, which is another sign of convergence

For all parameters, the fact that the three chains are largely indistinguishable from one another is a good indication that the chains have converged to a common distribution. The traceplots demonstrate that the sampling is stationary, which means the chains do not exhibit trends and the mixing is good (chains are exploring the parameter space efficiently).

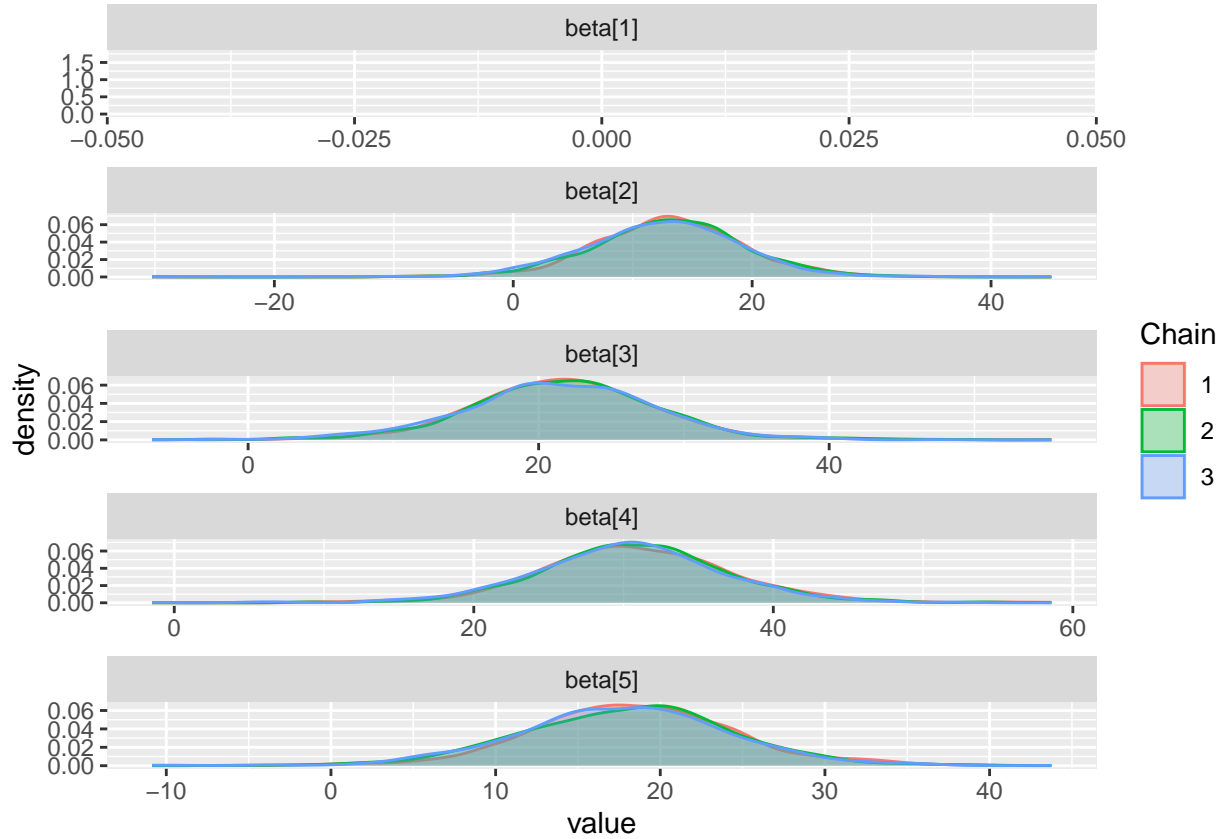
Overall, the MCMC diagnostics suggest that the model is well-specified and that the posterior distributions for the alpha parameters are being well-estimated by the sampler. Fields 2 and 3 show near zero effect on total carbon relative to Field 1, with the caveat that there is more uncertainty or variability associated with these estimates. The exact quantitative impact and the credibility of these effects can be further analyzed using credible intervals from the posterior distributions.



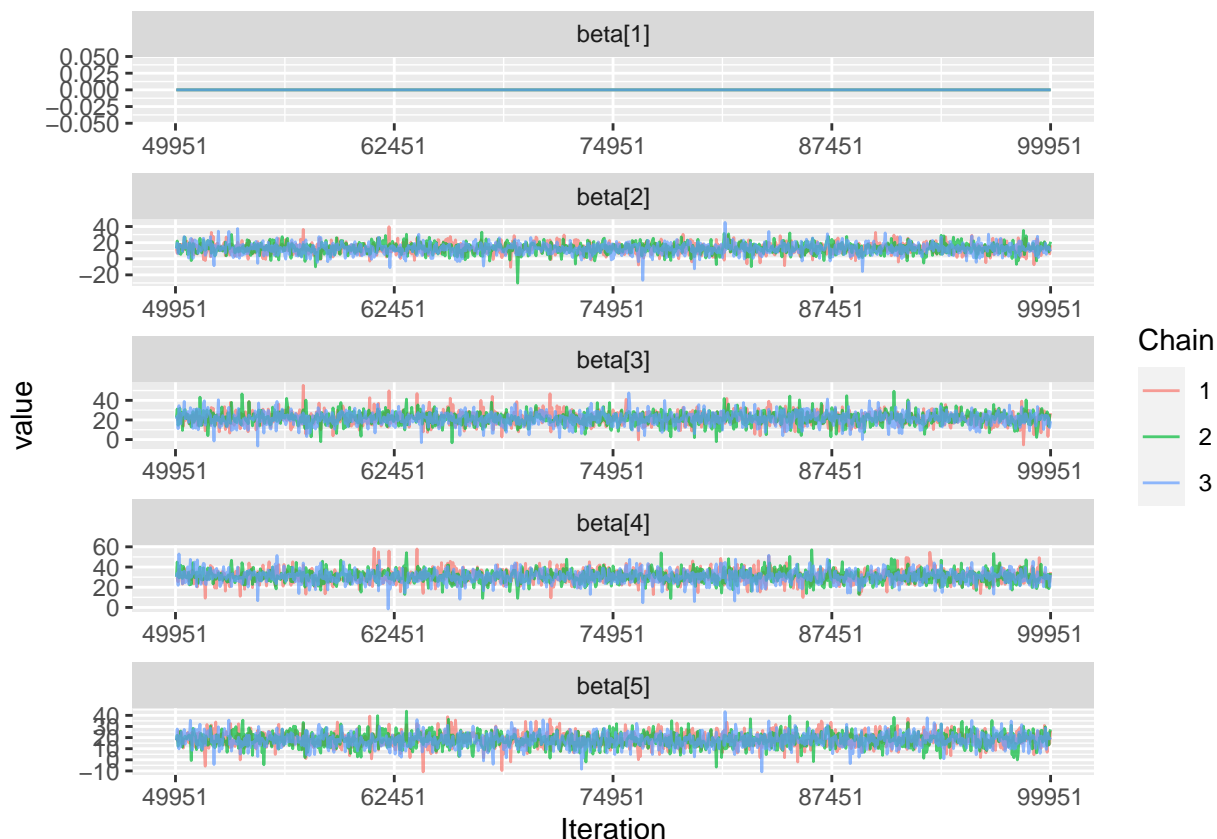
The density plots for the  $\beta$  parameters represent the posterior distributions of the effects of the carbon sequestration treatments (T2–T5) as compared to a baseline (T1, represented by  $\beta_1$ ).

- $\beta_3$ : The narrow distribution centered around zero suggests that this parameter, representing the baseline treatment effect, has little uncertainty around it. It is effectively serving as the reference category against which the other treatments are compared.
- $\beta_2$ ,  $\beta_3$ ,  $\beta_4$ , and  $\beta_5$ : These show the estimated effects of the treatments T2–T5, respectively, relative to T1 ( $\beta_1$ ). The wider distributions indicate more uncertainty compared to the baseline, but each one is

centered around positive values. This indicates that these treatments are estimated to have a positive effect on carbon sequestration relative to the baseline. Notably,  $\beta_5$  shows a particularly large effect, with a distribution that extends significantly further to the right, although it also has a long left tail suggesting noteworthy variance or uncertainty.



- $\beta_1$ : The trace is stable and hovers around the zero line, indicating that the chains are consistent in estimating this parameter as having a negligible effect.
- $\beta_2$ ,  $\beta_3$ ,  $\beta_4$ , and  $\beta_5$ : The traces for these parameters display a healthy “fuzziness” without any clear trends or cycles, suggesting that the chains are mixing well and exploring the parameter space appropriately. There’s no indication of non-convergence such as diverging paths or a lack of overlap between the chains. The values are stable throughout the iterations, showing no systematic drift, which is a good sign of convergence. The vertical spread of the traces for  $\beta_2$  to  $\beta_5$  is consistent with the wider posterior distributions seen in the density plots, reflecting the greater uncertainty in the estimation of these treatment effects as compared to the baseline.



In conclusion, the density and trace plots indicate that the MCMC algorithm has sampled effectively from the posterior distributions, and the chains appear to have converged. The treatments T2–T5 all have distributions suggesting a positive effect on carbon sequestration relative to the baseline treatment T1, with T5 showing potentially the largest effect. However, the actual effect sizes and their credible intervals would need to be examined to make quantitative conclusions about the magnitude and certainty of these effects.

## Bayesian Statistics Part (h)

```
## Compiling model graph
##   Resolving undeclared variables
##   Allocating nodes
## Graph information:
##   Observed stochastic nodes: 15
##   Unobserved stochastic nodes: 8
##   Total graph size: 46
##
## Initializing model

## Inference for Bugs model at "C:/Users/OGHENE~1/AppData/Local/Temp/RtmpWI5mzu/model134c067e97a26", fit
##   3 chains, each with 1e+05 iterations (first 50000 discarded), n.thin = 50
##   n.sims = 3000 iterations saved
##           mu.vect sd.vect   2.5%   25%   50%   75%   97.5% Rhat n.eff
## alpha[1]  0.000  0.000   0.000   0.000   0.000   0.000   0.000 1.000    1
## alpha[2]  1.988  5.014  -7.743  -1.212   1.991   5.014  12.384 1.001  3000
## alpha[3]  1.468  5.044  -8.203  -1.738   1.553   4.622  11.378 1.001  3000
```

```

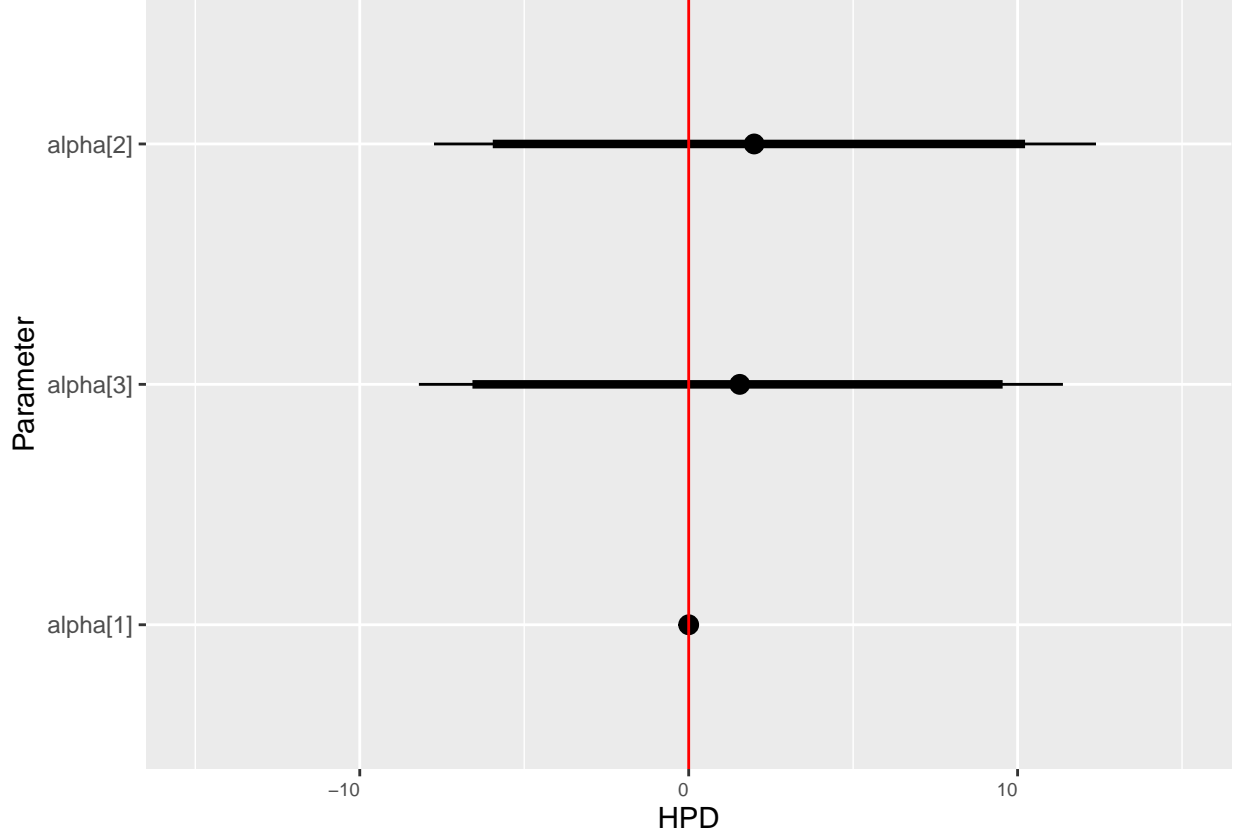
## beta[1]      0.000    0.000    0.000    0.000    0.000    0.000    0.000    1.000      1
## beta[2]     12.828    6.611    0.365    8.675    12.788    16.821    26.211    1.001    3000
## beta[3]     21.485    6.305    9.062    17.405    21.547    25.396    34.079    1.002    1500
## beta[4]     30.539    6.458    18.208    26.478    30.401    34.457    43.231    1.002    1400
## beta[5]     17.767    6.529    4.571    13.604    17.923    21.795    30.966    1.002    1100
## mu[1,1]    198.923    5.318   188.744   195.648   198.942   202.313   209.403    1.001    2000
## mu[2,1]    200.911    5.398   189.946   197.614   201.016   204.207   211.473    1.001    3000
## mu[3,1]    200.391    5.571   189.294   197.072   200.468   203.817   211.161    1.003    2400
## mu[1,2]    211.752    5.560   200.797   208.302   211.803   215.212   222.556    1.001    2600
## mu[2,2]    213.740    5.609   202.544   210.201   213.664   217.224   224.817    1.001    3000
## mu[3,2]    213.220    5.455   202.684   209.872   213.304   216.525   223.945    1.001    2200
## mu[1,3]    220.408    5.302   209.477   217.219   220.546   223.752   230.741    1.002    1600
## mu[2,3]    222.396    5.434   211.291   219.164   222.412   225.612   233.286    1.002    1100
## mu[3,3]    221.876    5.395   210.930   218.556   221.945   225.250   232.490    1.001    3000
## mu[1,4]    229.463    5.541   218.069   226.108   229.534   232.777   240.121    1.001    3000
## mu[2,4]    231.451    5.622   219.786   227.986   231.391   234.788   242.913    1.002    1700
## mu[3,4]    230.931    5.476   220.680   227.421   230.938   234.366   241.939    1.001    3000
## mu[1,5]    216.690    5.429   205.694   213.504   216.644   220.077   227.035    1.001    3000
## mu[2,5]    218.678    5.487   207.794   215.501   218.656   222.009   229.427    1.001    3000
## mu[3,5]    218.158    5.429   207.293   214.794   218.236   221.463   228.952    1.003     960
## sigma       7.628    2.224    4.603    6.076    7.215    8.660   12.851    1.001    3000
## tau         0.021    0.011    0.006    0.013    0.019    0.027    0.047    1.001    3000
## deviance   102.605    5.749   94.297   98.517  101.612  105.623  116.506    1.001    3000
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = var(deviance)/2)
## pD = 16.5 and DIC = 119.1
## DIC is an estimate of expected predictive error (lower deviance is better).

```

### Comparative Effects of Fields 2 and 3 vs. Field 1:

- The point estimates for  $\alpha_2$  and  $\alpha_3$  suggest that Fields 2 and 3 may enhance carbon sequestration compared to Field 1. However, the large credible intervals for both  $\alpha_2$  and  $\alpha_3$  contain 0, which indicates that the data do not provide strong evidence that the carbon sequestration effects of these fields are different from Field 1. In other words, while there is an indication of a positive effect, the uncertainty is too great to conclusively say that the total carbon values are different when using Fields 2 or 3 instead of Field 1.
- **Statistical Significance and Uncertainty:** There is a significant overlap in the credible intervals for  $\alpha_2$  and  $\alpha_3$ , and both contain the null value of zero, which further emphasizes the point that there is no clear evidence of a difference. This suggests that any apparent differences in carbon sequestration effects could be due to chance rather than systematic differences between the fields.
- **Model Convergence and Reliability:** The Rhat values for both  $\alpha_2$  and  $\alpha_3$  being close to 1 suggest that the MCMC chains have converged properly, which lends credibility to the model's estimates. However, even with good convergence and a large effective sample size, the width of the credible intervals reveals the variability in the data and suggests the need for caution in interpreting these results.
- **Justification:** The conclusion is based on both the visual evidence from the caterpillar plot and the numerical output from the JAGS model. The point estimates alone suggest a difference, but the uncertainty (reflected in the wide credible intervals) prevents us from making a definitive statement

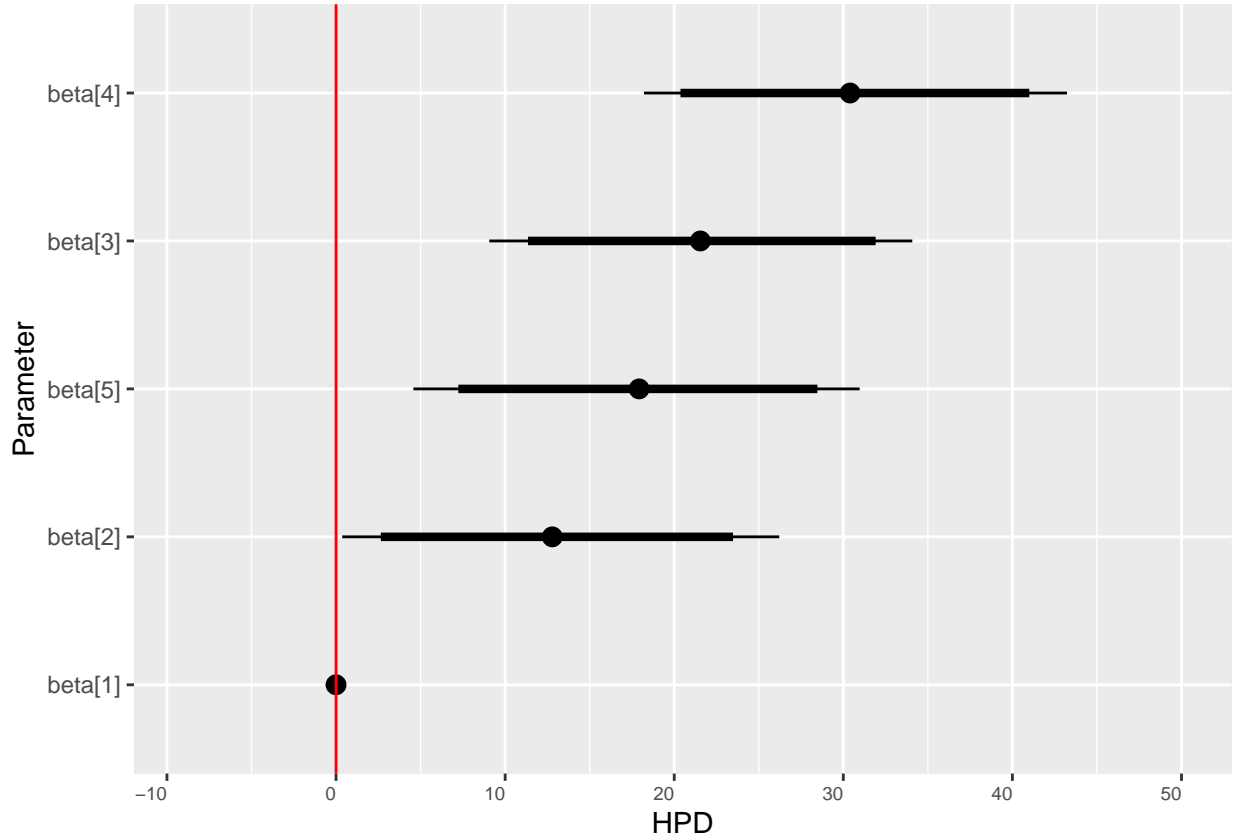
about the effects of different field locations on carbon sequestration. It is essential to consider the entire credible interval when making inferences from Bayesian analyses, not just the point estimate. In the context of making policy or management decisions regarding carbon sequestration strategies, it would be prudent to gather more data to reduce the uncertainty of these estimates before concluding that one field location results in a significantly different total carbon value than another.



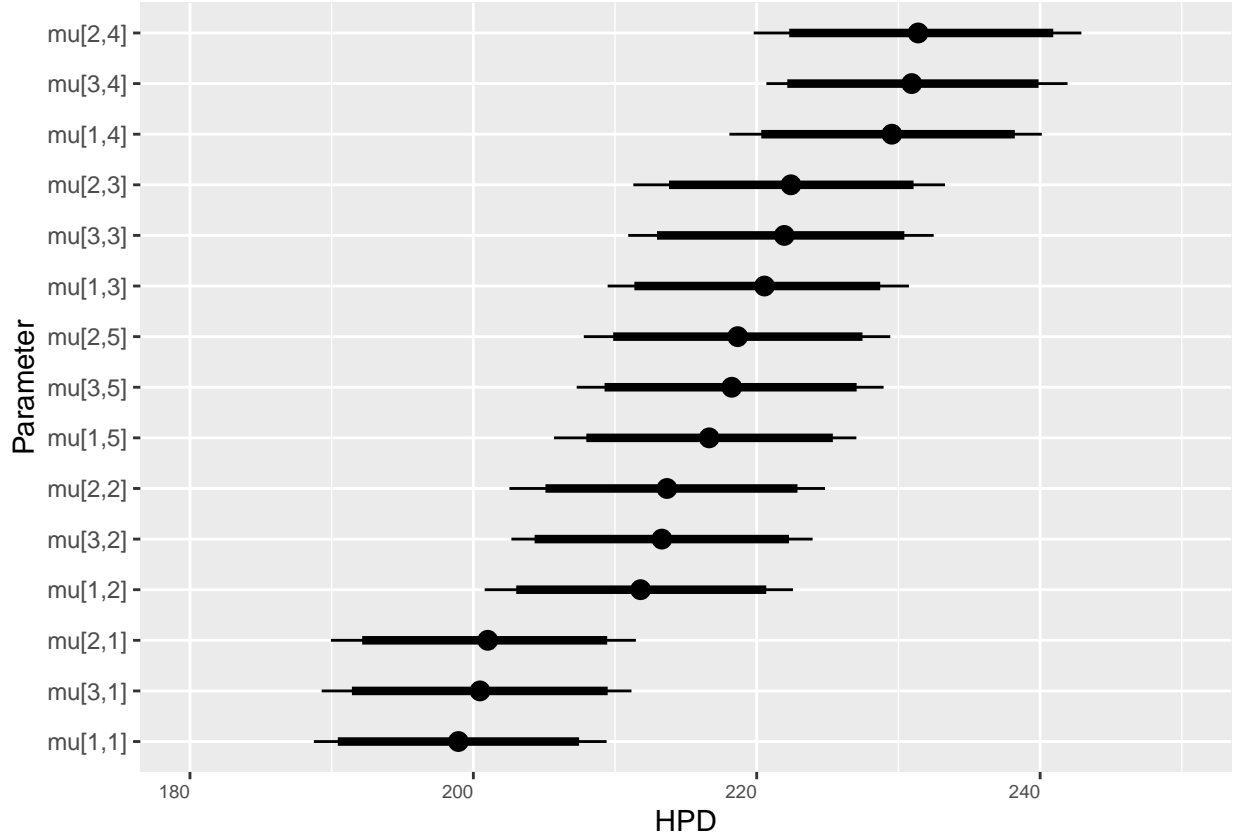
**Effectiveness of Different Treatments:** Treatments 2 through 5 all show positive effects on carbon sequestration relative to Treatment 1, with their credible intervals lying mostly or entirely above zero. This suggests that these treatments are more effective than the baseline treatment.

- **Statistical Significance:** For Treatments 2, 3, 4, and 5, the credible intervals do not include zero (with the possible exception of the lower bound of Treatment 2), indicating that their effects are statistically significant.
- **Comparison Between Treatments:** The effects of the treatments increase from Treatment 2 to Treatment 4, with Treatment 4 showing the most substantial impact. Treatment 5, while still effective, shows a smaller effect than Treatment 4 but is more effective than Treatments 2 and possibly 3.
- **Model Reliability:** The Rhat values are all close to or at 1, and the effective sample sizes are adequate (3000 for Treatments 2, 3, and 5; 1800 for Treatment 4), indicating that the chains have likely converged, and the model estimates can be considered reliable.
- **Justification:** These conclusions are justified by both the numerical output and the caterpillar plot, which visually represents the central tendency and variability of the estimates. The credible intervals provide a Bayesian measure of the precision of the estimates and the uncertainty surrounding them. The fact that all treatments' credible intervals are largely or entirely above zero strongly indicates that

there are real differences in the effects of these treatments on carbon sequestration. Given the clear separation from the baseline treatment and the lack of overlap with zero for most treatments, we can conclude that the different treatments lead to different total carbon values in the soil samples.



- The HDP intervals for the estimated total carbon for different combinations of the 5 treatments and for the 3 fields varies across all combinations from our posterior inferences. The plot reveals no clear variation between combinations based on the 3 fields with little distinguishable effect between combinations on the basis of the fields
- On the other hand, there is a clear distinguishable variance between the combinations on the basis of the treatments. The largest variance can be seen between combinations with the treatment T4 and those with T1. The estimated total carbon for combinations with treatments T2, T3 and T5 show more significant overlap indicative of less variation between these combinations.
- Generally the plot confirms a notable variation in the total carbon when different treatments are applied but little variation between the application of same treatments across different fields.



**Bayesian Statistics Part (i)** Performing posterior inferences about the differences between:

- $\beta_4$  and  $\beta_1$ ;
- $\beta_4$  and  $\beta_2$ ;
- $\beta_4$  and  $\beta_3$ ;
- $\beta_4$  and  $\beta_5$ .

**Explanation of the quantities:**

To test the farmer's expectations, we can consider the average total carbon ( $\mu_4$ ) resulting from T4 as the base treatment we want to compare with in our model. Hence we set the prior distribution for  $\beta_4$  to zero. The difference between  $\beta_4$  and  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$  and  $\beta_5$  quantifies the difference in the effect of the corresponding treatments on carbon sequestration when compared with T4. In this model, positive values for these parameters implies a higher total carbon estimate for these treatments compared with T4 while negative values indicate a lower total carbon estimate versus T4.

We get the following results:

- **Consideration of the Caterpillar Plot:** The caterpillar plot corroborates these findings, showing the 95% HPD intervals for each beta estimate. All beta coefficients are to the left of beta 4, indicating that all treatments are estimated to be less effective than T4. This visual representation allows us to see at a glance that none of the other treatments outperform T4 within the modeled certainty levels.
- **Conclusions:** The model results and caterpillar plot provide evidence supporting the farmer's belief that treatment T4 leads to higher carbon sequestration in the soil compared to the other treatments

tested (T1, T2, T3, and T5). This is evidenced by the fact that all other beta coefficients are negative and their HPD intervals do not include zero (with the slight exception of T3 where zero is barely included), suggesting a lower effect than the baseline T4.

The consistency in the statistical diagnostics (Rhat values around 1, effective sample sizes of 3000) across most parameters also suggests that the model is well-converged and the estimates are reliable. The model's deviance is relatively low, and the DIC score of 99.004 suggests a good model fit to the data. Therefore, in the context of this analysis, the results are in line with the farmer's expectations for the effectiveness of treatment T4.

### 3.2.3 Third Sub-Task: Simpler Bayesian model

**Bayesian Statistics Part (j):** Simpler Bayesian model for the carbon sequestration treatment data:

$$\begin{aligned}
 y_{ij} &\sim \mathcal{N}(\mu_j, \text{precision} = \tau), \quad i = 1, 2, 3, \quad j = 1, \dots, 5 \\
 \mu_j &= \mu + \beta_j, \quad \beta_1 = 0 \\
 \mu &\sim \mathcal{N}(0, \text{precision} = 0.0001) \\
 \beta_j &\sim \mathcal{N}(0, \text{precision} = 0.0001), \quad j = 2, \dots, 5 \\
 \tau &\sim \text{Gamma}(\text{shape} = 0.001, \text{rate} = 0.001) \\
 \text{standard deviation } \sigma &= \sqrt{\frac{1}{\tau}}
 \end{aligned}$$

**Posterior estimates:**

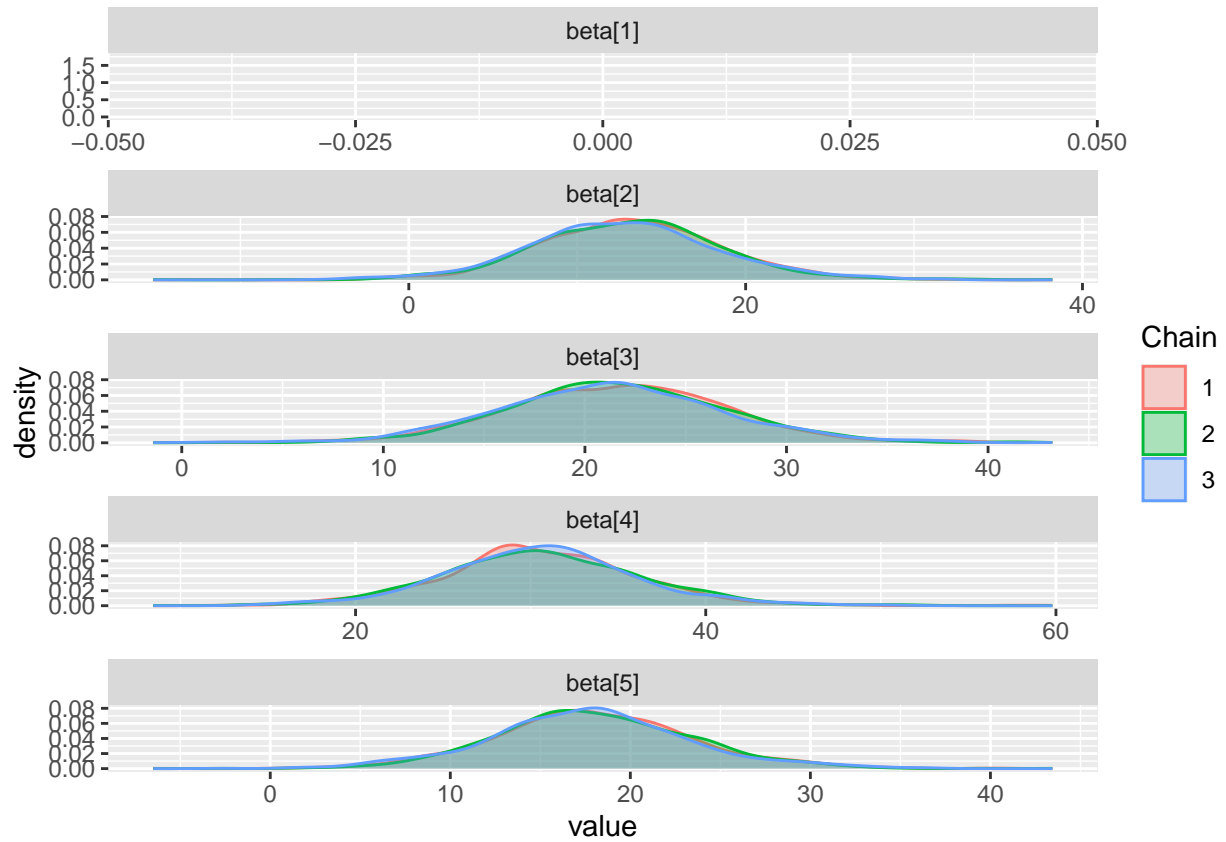
```
## Compiling model graph
##   Resolving undeclared variables
##   Allocating nodes
## Graph information:
##   Observed stochastic nodes: 15
##   Unobserved stochastic nodes: 6
##   Total graph size: 49
##
## Initializing model

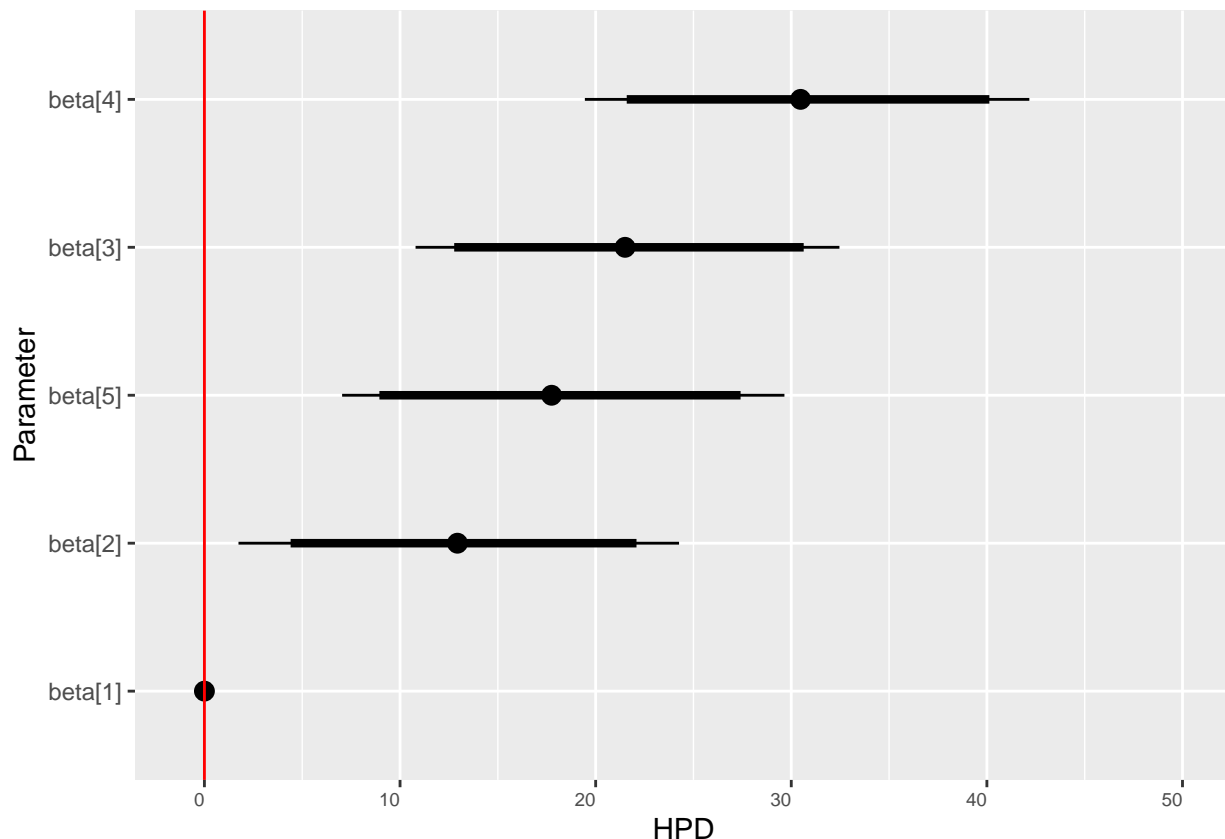
## Inference for Bugs model at "C:/Users/OGHENE~1/AppData/Local/Temp/RtmpWI5mzu/model34c015a34383", fit
##   3 chains, each with 1e+05 iterations (first 50000 discarded), n.thin = 50
##   n.sims = 3000 iterations saved
##           mu.vect sd.vect   2.5%   50%   97.5%  Rhat n.eff
## beta[1]    0.000   0.000   0.000   0.000   0.000 1.000    1
## beta[2]   13.008   5.564   1.742  12.934  24.263 1.001  2100
## beta[3]   21.515   5.531  10.793  21.505  32.462 1.002  2000
## beta[4]   30.603   5.677  19.449  30.477  42.171 1.000  3000
## beta[5]   17.916   5.547   7.039  17.747  29.651 1.001  2600
## m         200.037   3.944 191.952 200.182 207.578 1.001  3000
## sigma      6.759   1.721   4.324   6.468  11.006 1.002  1900
## tau        0.026   0.012   0.008   0.024   0.053 1.002  1900
## deviance  99.034   4.365  92.962  98.250 109.615 1.002  1700
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
```



```
## DIC info (using the rule, pD = var(deviance)/2)
## pD = 9.5 and DIC = 108.6
## DIC is an estimate of expected predictive error (lower deviance is better).
```

### Bayesian Statistics Part (k):





The density plots show the distribution peaks at values well above zero, indicating the posterior inferences of the parameters  $\beta_2, \beta_3, \beta_4$  and  $\beta_5$  are likely greater than zero from our model. This is further reinforced in the caterpillar plot where the entire HPD intervals for these parameters all lie to the right of the 0 mark. The JAGS model inferences demonstrate that treatments T2, T3, T4, and T5 all significantly increase carbon sequestration in the soil compared to the baseline treatment T1. Specifically, treatment T4 shows the largest increase in carbon sequestration among the treatments tested, as indicated by both the caterpillar plot and the summary statistics. All treatments have positive effects, but T4 stands out with the highest difference in total carbon compared to T1.

## Bayesian Statistics Part (I)

### Justification of Preferred Model:

Typically, when evaluating models, those with a lower Deviance Information Criterion (DIC) are favored as the DIC assesses model fit while accounting for model complexity. In comparing the Bayesian ANOVA models, the one noted in section (f) exhibits a DIC of 119.53, whereas the model referenced in section (j) shows a more favorable DIC of 108.56, indicating that the latter model provides a more suitable fit. Furthermore, the model in (f) incorporates an additional complexity with the inclusion of the alpha term, which represents the parametric effect of the field on soil sample results. Therefore, we lean towards the less complex model presented in section (j), which also benefits from a lower DIC value.

## References

- Fernández i Marín, Xavier. 2021. *Ggmcnc: Tools for Analyzing MCMC Simulations from Bayesian Inference*. <https://CRAN.R-project.org/package=ggmcnc>.
- Fernández-i-Marín, Xavier. 2016. “ggmcnc: Analysis of MCMC Samples and Bayesian Inference.” *Journal of Statistical Software* 70 (9): 1–20. <https://doi.org/10.18637/jss.v070.i09>.
- Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2023. *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. <https://CRAN.R-project.org/package=e1071>.
- Ripley, Brian. 2023a. *Class: Functions for Classification*. <http://www.stats.ox.ac.uk/pub/MASS4/>.
- . 2023b. *MASS: Support Functions and Datasets for Venables and Ripley’s MASS*. <http://www.stats.ox.ac.uk/pub/MASS4/>.
- . 2023c. *Tree: Classification and Regression Trees*. <https://CRAN.R-project.org/package=tree>.
- Sing, Tobias, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. 2020. *ROCR: Visualizing the Performance of Scoring Classifiers*. <http://ipa-tys.github.io/ROCR/>.
- Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer. 2005. “ROCR: Visualizing Classifier Performance in r.” *Bioinformatics* 21 (20): 7881. <http://rocr.bioinf.mpi-sb.mpg.de>.
- Su, Yu-Sung, and Masanao Yajima. 2024. *R2jags: Using r to Run JAGS*. <https://CRAN.R-project.org/package=R2jags>.
- Venables, W. N., and B. D. Ripley. 2002b. *Modern Applied Statistics with s*. Fourth. New York: Springer. <https://www.stats.ox.ac.uk/pub/MASS4/>.
- . 2002a. *Modern Applied Statistics with s*. Fourth. New York: Springer. <https://www.stats.ox.ac.uk/pub/MASS4/>.
- Wickham, Hadley. 2007. “Reshaping Data with the Reshape Package.” *Journal of Statistical Software* 21 (12). <https://www.jstatsoft.org/v21/i12/>.
- . 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- . 2022. *Reshape: Flexibly Reshape Data*. <http://had.co.nz/reshape>.
- Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2023. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. <https://CRAN.R-project.org/package=ggplot2>.
- Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>.
- Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. <https://CRAN.R-project.org/package=readr>.
- Zhu, Hao. 2021. *kableExtra: Construct Complex Table with Kable and Pipe Syntax*. <https://CRAN.R-project.org/package=kableExtra>.