Estimation and Prediction of the Probability of Death in a Traffic Accident

Maria Elzaurdi, Guillermo Villanueva, Carme Viñas

August 14, 2022

1 Introduction

1.1 Project Description

The number of accidents that take place on the roads and cause deaths has been a constant matter of concern during the last decades. In the year 2019 there were 104080 traffic accidents with victims in Spain, which led to a total of 1755 deaths. The aim of this study is to estimate the probability of an accident being fatal (leading to one or more deaths) and to find out which variables have an impact on this probability, as well as to know the magnitude of the influence that each of the explanatory variables have on the response. Hence, we could know what to improve in order to reduce the number of fatalities on Spanish roads and we would be able to predict the number of fatal accidents given the total number of traffic accidents.

In the present project, after introducing the variables of interest and the data of use, we are going to create different Bayesian models to answer our questions. The first model built will be a pooled Bayesian model in which we will use several variables, with the goal of differentiating the variables that are relevant to those that are not. We will use this model to predict the proportion of deadly accidents of other years. Secondly, we will build a simpler model with those variables that are more relevant to see if we can still predict the results accurately with it. Finally, we will propose an unpooled and a hierarchical model for the different regions (*Comunidades Autónomas*) to find out the differences and similarities between them.

1.2 Data

The data used for this report has been obtained from the DGT (Dirección General de Tráfico). The following excel file contains data (a total of 72 variables) for traffic accidents registered in 2019. The accidents included in this database (104080) are those where there were victims (either injured or dead people) reported.

https://www.dgt.es/menusecundario/dgt-en-cifras/dgt-en-cifras-resultados/dgt-en-cifras-detalle/?id=00174

The following table contains a few rows of the selected variables of use unfiltered:

TOTAL MU30DF	ZONA AGRUPADA	HORA	CONDICION METEO	MES	DIA SEMANA	COD PROVINCIA
0	1	8	1	2	4	1
0	1	10	1	6	1	1
0	1	16	1	7	5	1
0	1	10	2	11	2	1
0	1	16	1	2	5	1
0	2	22	2	12	7	1
1	1	16	1	3	6	1

The meaning of the categorical variables can be found in https://www.dgt.es/menusecundario/dgt-en-cifras/dgt-en-cifras-resultados/dgt-en-cifras-detalle/?id=00175.

The corresponding code of each province can be found in https://www.ine.es/daco/daco42/codmun/cod_ccaa_provincia.htm. We cluster the provinces into bigger regions (*Comunidades Autónomas*) in the following way:

Provinces Code	**Region**	**Region Code**
4, 11, 14, 18, 21, 23, 29, 41	Andalucía	1
22, 44, 50	Aragón	2
33	Asturias	3
07	Illes Balears	4
35, 38	Canarias	5
39	Cantabria	6
5, 9, 24, 34, 37, 40, 42, 47, 49	Castilla y León	7
2, 13, 16, 19, 45	Castilla La Mancha	8
8, 17, 25, 43	Cataluña	9
3, 12, 46	Comunitat Valenciana	10
6, 10	Extremadura	11
15, 27, 32, 36	Galicia	12
28	Comunidad de Madrid	13
30	Región de Murcia	14
31	Comunidad Foral de Navarra	15
1, 48, 20	País Vasco	16
26	La Rioja	17
51	Ceuta	18
52	Melilla	19

1.2.1 Data Analysis

In this section we briefly analyse the data of interest.

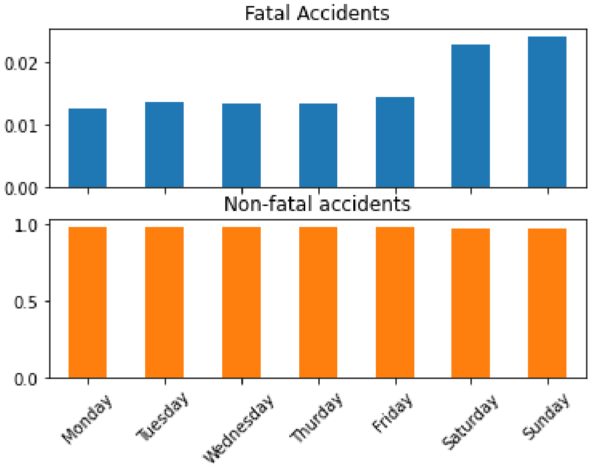

Figure 1: Bar plot showing the fraction of fatal accidents and non-fatal accidents for each day of the week.

Fig.1 shows, for each day the week, the number of fatal/non-fatal accidents divided by the total number of accidents registered in each day of the week. Interestingly, Saturdays and Sundays are the days of the week in which the number of registered fatal accidents increases. Moreover, we notice that the difference with respect to the other days of the week is considerable.

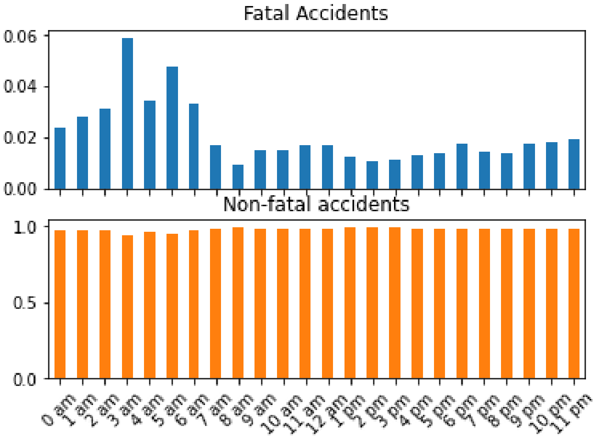

Figure 2: Bar plot showing the fraction of fatal accidents and non-fatal accidents for each day of the week.

Fig.2 shows, for each one time of the day (in 1-hour slots), the number of fatal/non-fatal accidents divided by the total number of accidents registered at that time slot. We notice that it is between 3 am and 6 am where the fraction of fatal accidents is the highest.

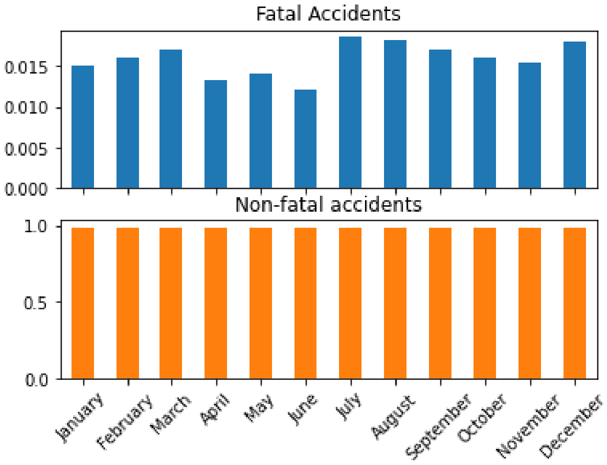

Figure 3: Bar plot showing the fraction of fatal accidents and non-fatal accidents for month of the year.

Fig.3 shows, for each month of the year, the number of fatal/non-fatal accidents divided by the total number of accidents registered in each month. We notice that July and August seem to be the months in which the fraction is higher, although the difference is not significant. Unlike the previous two explanatory variables, the relationship between the type of accident and the month of the year is not very clear.

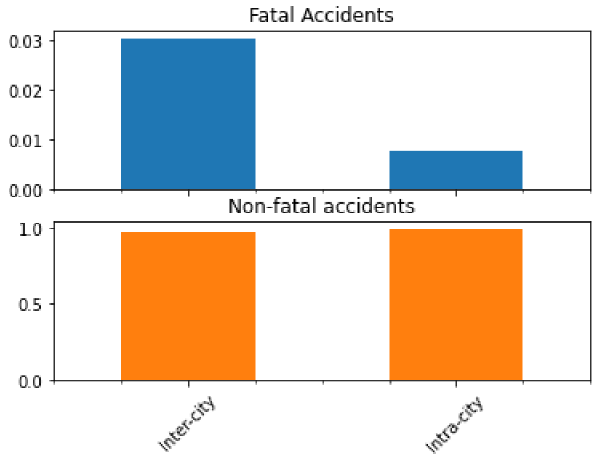

Figure 4: Bar plot showing the fraction of fatal accidents and non-fatal accidents for each type of road.

Fig.4 shows, for each type of road, the number of fatal/non-fatal accidents divided by the total number of registered accidents. We notice that the type of road also affects the fraction of fatal accidents. A higher fraction of fatal accidents are registered in inter-city roads.

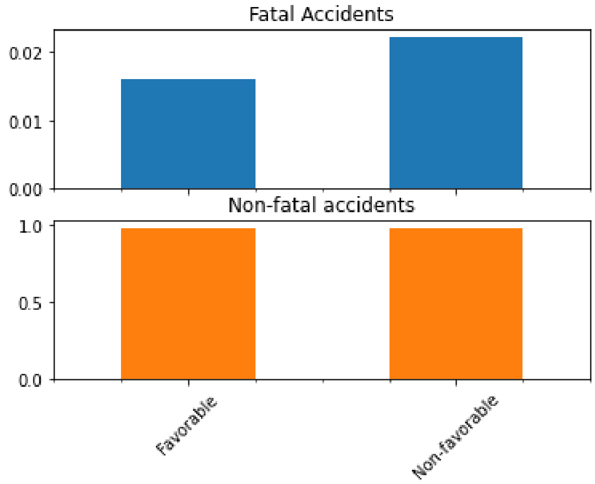

Figure 5: Bar plot showing the fraction of fatal accidents and non-fatal accidents for each type of weather condition.

Finally, Fig.5 shows, for each type of weather condition, the number of fatal/non-fatal accidents divided by the total number of registered accidents. Similarly to the previous explanatory variable, non-favorable weather conditions register a higher fraction of fatal accidents. However, the difference between favorable and non-favorable is smaller in comparison to the difference between intra-city and inter-city accidents.

We will take into account these insights and relationships between the response variable and explanatory variables to build a Bayesian model to predict the type of accident.

1.2.2 Processed data

We filter the data in order to make all the variables binary. We consider the road type to be 1 if it is inter-city (previously 1) and 0 if it is intra-city (previously 2). The time takes 0 value when it is between 8 and 20, and 1 when it is smaller than 8 or larger than 20 (between 21 and 7). Meteo is 0 when it is favourable (previously 1) and 1 when it is not (previously equal or larger than 2). The holidays variable is 1 for holidays peak season (months 7 and 8) and 0 otherwise. The weekend variable takes value 1 for days 5, 6, 7 and 0 otherwise. The response variable is already binary. The only variable that remains non-binary is the region code, which can take integers numbers between 1 and 19 as values, each number corresponding to a specific region. The following table contains a few rows of the selected variables of use after filtering them:

Fatal	**Road**	**Time**	**Meteo**	**Holidays**	**Weekend**	**Region**
0	1	0	0	0	0	16
0	1	0	0	0	0	16
0	1	0	0	1	1	16
0	1	0	1	0	0	16
0	1	0	0	0	1	16
0	0	1	1	0	1	16
1	1	0	0	0	1	16

In order to validate our model, we will use data of the years 2018 (found in `https://www.dgt.es/menusecundario/dgt-en-cifras/dgt-en-cifras-resultados/dgt-en-cifras-detalle/?id=00173`), 2017 (found in `https://www.dgt.es/menusecundario/dgt-en-cifras/dgt-en-cifras-resultados/dgt-en-cifras-detalle/?id=00172`) and 2016 (found in `https://www.dgt.es/menusecundario/dgt-en-cifras/dgt-en-cifras-resultados/dgt-en-cifras-detalle/?id=00171`), which will be filtered and modified in the same way.

1.3 Variables

The following table contains a summary of the variables used in the model. In addition to the response variable, which is a binary variable that indicates whether an accident is fatal (1) or not (0), there are five explanatory variables and one classification variable used to separate the data of the different regions in Spain.

CODE	**NAME**	**EXPLANATION**	**TYPE**
y	Fatal	Binary variable that indicates if an accident is fatal (1) or not (0). Deaths are computed 30 days after the accident.	Response Variable
Z	Road	Binary variable that indicates if a road is inter-city (1) or intra-city (0).	Explanatory Variable
T	Time	Binary variable that indicates if the accident took place during the night (from 9 pm to 7 am) (1) or the day (0).	Explanatory Variable
M	Meteo	Binary variable that indicates if the meteorological conditions are unfavourable (1) or favourable (0).	Explanatory Variable
H	Holidays	Binary variable that indicates if there is the accident took place in July or August (peak holiday season) (1) or not (0).	Explanatory Variable
E	Weekend	Binary variable that indicates if the accident took place during the weekend (from Friday to Sunday) (1) or not (0).	Explanatory Variable
P	Province	Spanish provinces codes.	Classification Variable

2 Bayesian Model

The first thing we want to estimate is the proportion of traffic accidents that are fatal and, in parallel, which variables have an effect on this probability. We will consider that an accident is fatal when it leads to at least one death. Therefore, we propose a regression model in which the event of an accident being fatal y follows a Bernoulli distribution with $y = 0$ when there are not any fatalities and $y = 1$ when there is at least one dead person. Thus, θ will define the probability of an accident leading to at least one fatality.

$$y \sim \text{Bernoulli}(\theta)$$

where $0 < \theta < 1$ and $p(y) = \theta^y(1-\theta)^{1-y}$. The expected value of the model is $E[y] = \theta$ and the variance corresponds to $V[y] = \theta(1-\theta)$.

As we are in a regression model, the probability of an accident being fatal, θ, will have the following expression:

$$\log\left(\frac{\theta}{1-\theta}\right) = \beta_0 + \beta_1 Z + \beta_2 T + \beta_3 M + \beta_4 H + \beta_5 E$$

where

$$Z = \begin{cases} 1 & \text{if national road} \\ 0 & \text{otherwise} \end{cases}$$

$$T = \begin{cases} 1 & \text{if night time} \\ 0 & \text{otherwise} \end{cases}$$

$$M = \begin{cases} 1 & \text{if unfavourable meteorological conditions} \\ 0 & \text{otherwise} \end{cases}$$

$$H = \begin{cases} 1 & \text{if holiday season} \\ 0 & \text{otherwise} \end{cases}$$

$$E = \begin{cases} 1 & \text{if weekend} \\ 0 & \text{otherwise} \end{cases}$$

The prior distributions of parameters β_i (with $i = 0, 1, 2, 3, 4, 5$) will be non-informative distributions as we do not have any previous knowledge about these parameters. We define all of them as uniform distributions taking values between -100 and 100 because we think that, since these parameters are modifying the slope of a straight line, this range is wide enough to include all the possible values that they could take.

2.1 Results

We obtain the following results for the parameters β_i (with $i = 0, 1, 2, 3, 4, 5$):

PARAMETER	**95% CI**	**EXPECTED VALUE**
β_0	$[-5.206, -4.990]$	-5.102
β_1	$[1.266, 1.473]$	1.372
β_2	$[0.408, 0.619]$	0.511
β_3	$[-0.327, -0.046]$	-0.185
β_4	$[-0.027, 0.216]$	0.097
β_5	$[0.190, 0.381]$	0.285

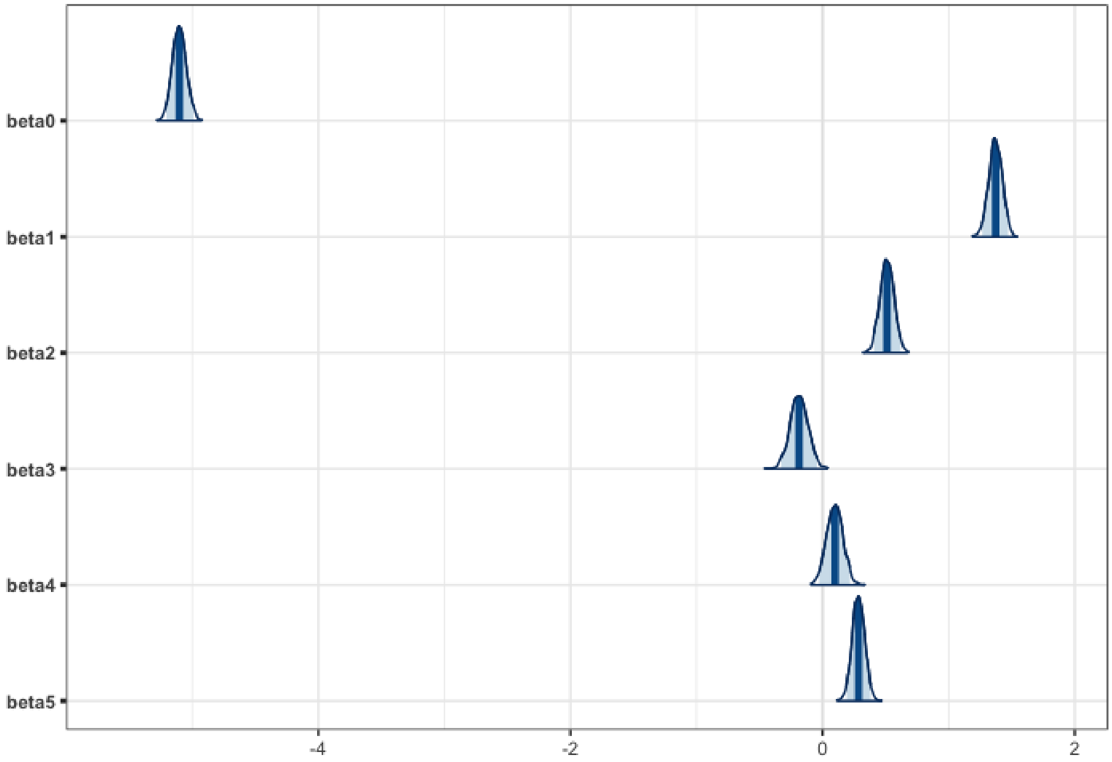

Figure 6: Posterior distributions of the parameters β_i $(i = 0, 1, 2, 3, 4, 5)$ with medians and 95% credible intervals.

From the results above, we can say that the variables corresponding to the type of road and the time have a big impact on the fatality probability in a car accident. That is, if an accident takes place in a national road (inter-city road) and/or at night (between 9 pm and 7 am) the chances of it being deadly increase. Among the two, the effect of the type of road is is higher rather than the time. Regarding the other three variables, they have a smaller impact on the response. The meteorological variable has a negative effect, which means that, according to the model and data of use, in a situation where the meteorological conditions are unfavourable the proportion of fatal accidents is reduced. The other two variables, holidays and weekends, have a very small influence on the response variable, meaning that the percentage of accidents that end up being fatal is barely affected by the month or day of the week when the accident take place. Furthermore, the intercept is large and negative.

2.2 Prediction and Validation

In order to validate the first model, we try to simulate and predict results from other years (2016, 2017 and 2018).

Firstly, we generate the distribution of theta for each individual accident of the years 2016, 2017 and 2018 using the explanatory variables (Z, T, M, H and E) of each accident and the distributions of beta previously obtained using 2019 data. From each distribution of theta, we can simulate the number of expected fatal accidents in a year using the Bernoulli distribution. We plot in a histogram the expected number of fatal accidents out of the total predicted by the explanatory variables of each accident. Furthermore, we calculate the mean of all the values of the total number of fatal accidents obtained and compare it with the real value.

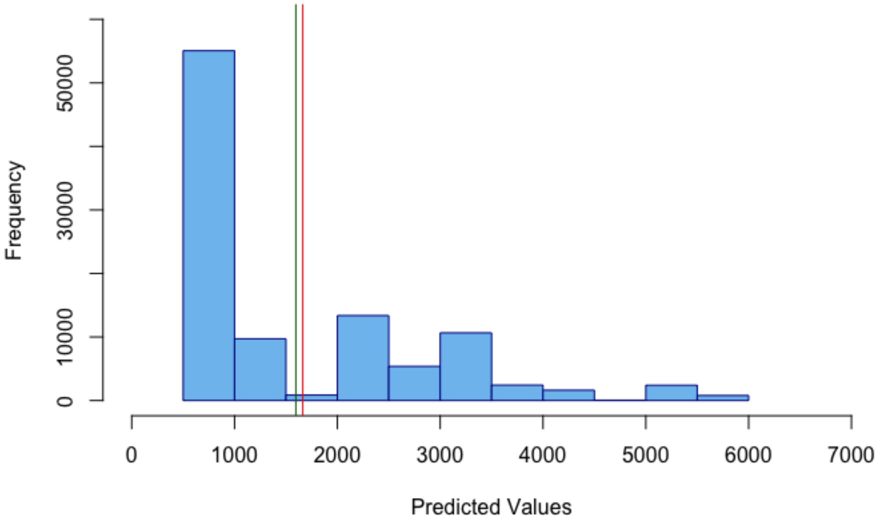

Figure 7: Predicted numbers of fatal accidents in 2016. The red line corresponds to the real number of fatal accidents and the green line, to the mean predicted value.

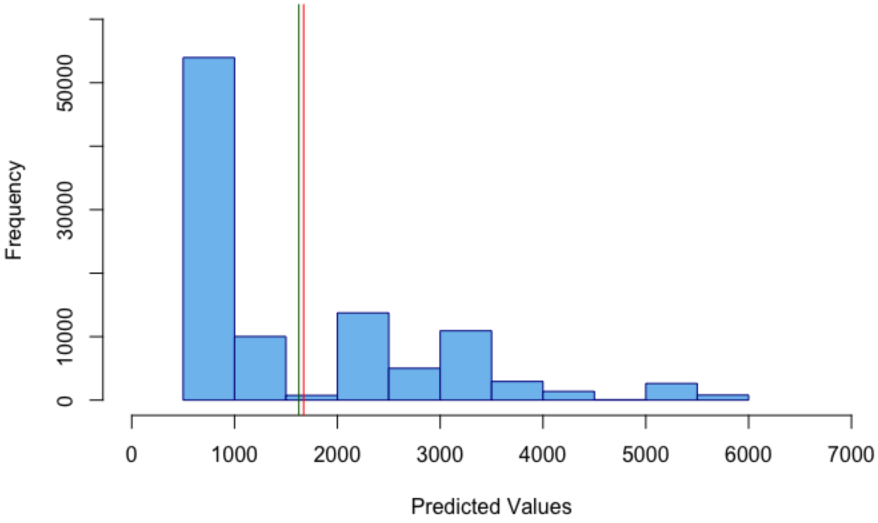

Figure 8: Predicted numbers of fatal accidents in 2017. The red line corresponds to the real number of fatal accidents and the green line, to the mean predicted value.

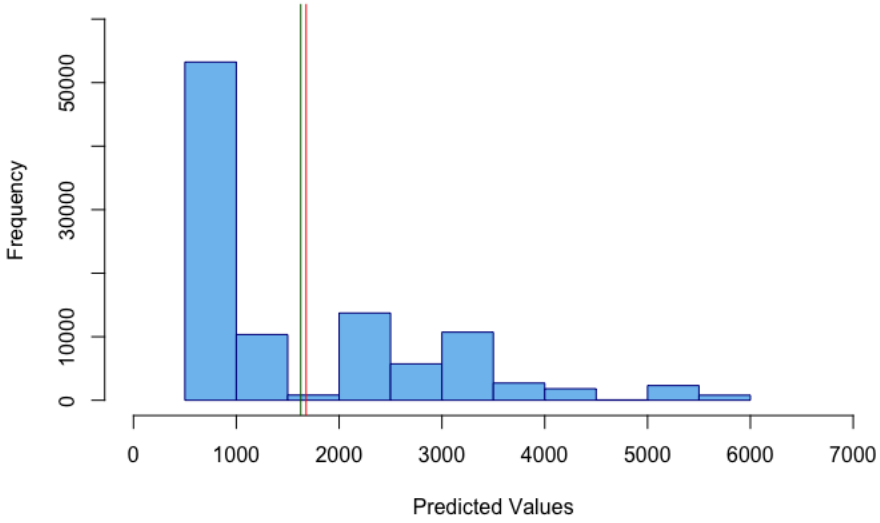

Figure 9: Predicted numbers of fatal accidents in 2018. The red line corresponds to the real number of fatal accidents and the green line, to the mean predicted value.

YEAR	**ACCIDENTS**	**REAL NUMBER OF FATAL ACCIDENTS**	**PREDICTED NUMBER OF FATAL ACCIDENTS**
2016	102363	1663	1596
2017	102234	1672	1625
2018	102300	1679	1626

From the obtained results it seems that the first Bayesian model predicts the proportion of deadly accidents with high accuracy, even though the predicted values are slightly lower than the real ones. However, due to the total number of accidents, the number of fatal accidents and the ratio between the two is similar in every year, we would expect that the predicted results are close to the ones used when calculating the posterior distributions of the parameters.

Another way to validate the model is to predict, given the explanatory variables of each individual accident (making use of data from 2016, 2017 and 2018), the response variable of each of those accidents using the parameter distributions obtained and check if the event of an accident being fatal or not is correctly predicted in each case. By this method, we get that the first Bayesian model proposed predicts the response variable with about 97%of accuracy. However, while the accuracy when predicting the non-fatal accidents is as high as 98.5%, less than 1% of fatal accidents are correctly predicted. Hence, we conclude that, even though the model seems to give a precise number of fatal accidents out of the total, it is not good for predicting if a certain accident is going to be fatal given the values of the explanatory variables of such accidents.

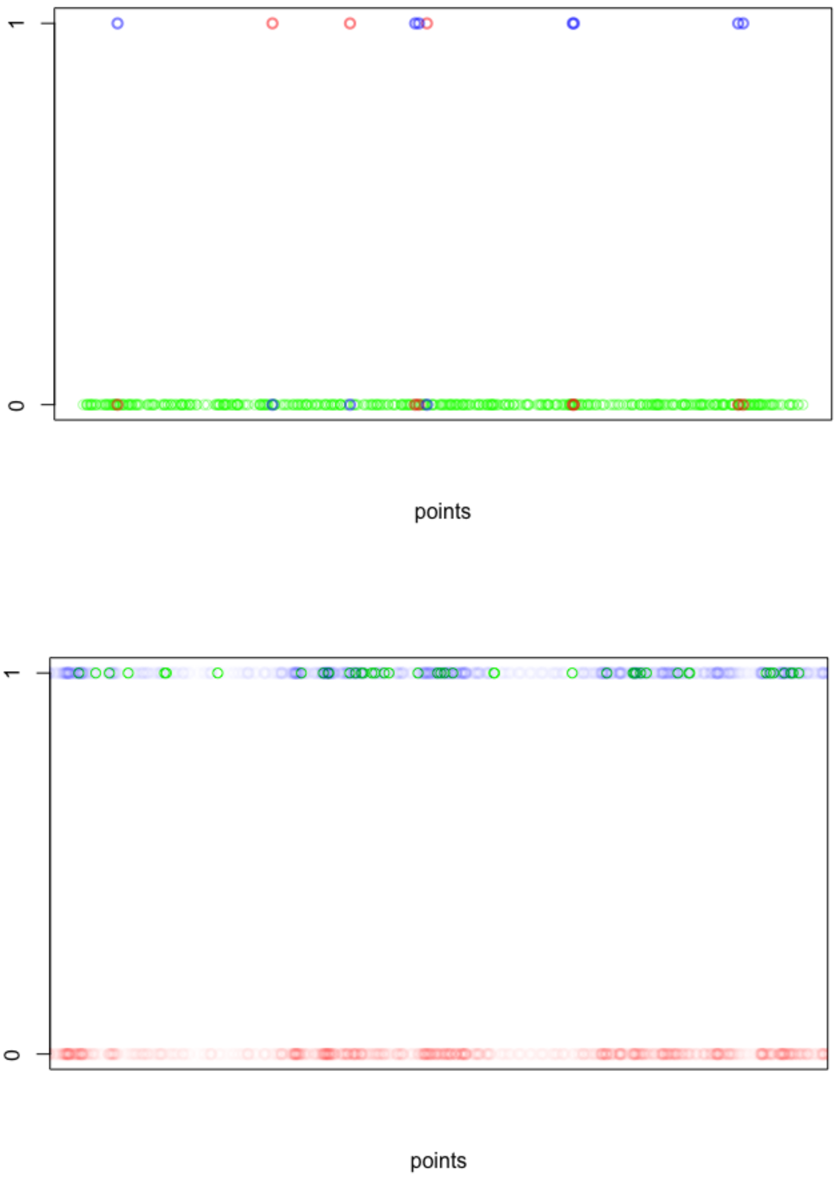

Figure 10: Accuracy of prediction of 500 randomly selected accidents (top) and all fatal accidents (bottom) in 2016. The green points indicate the accidents in which the response variable has been correctly predicted. Alternatively, the red and blue points indicate the accidents in which the response variable has not been predicted well. In those cases, the red color is the predicted value and the blue color is the true value. We can observe that the prediction is good for non-fatal accidents but bad for the fatal ones.

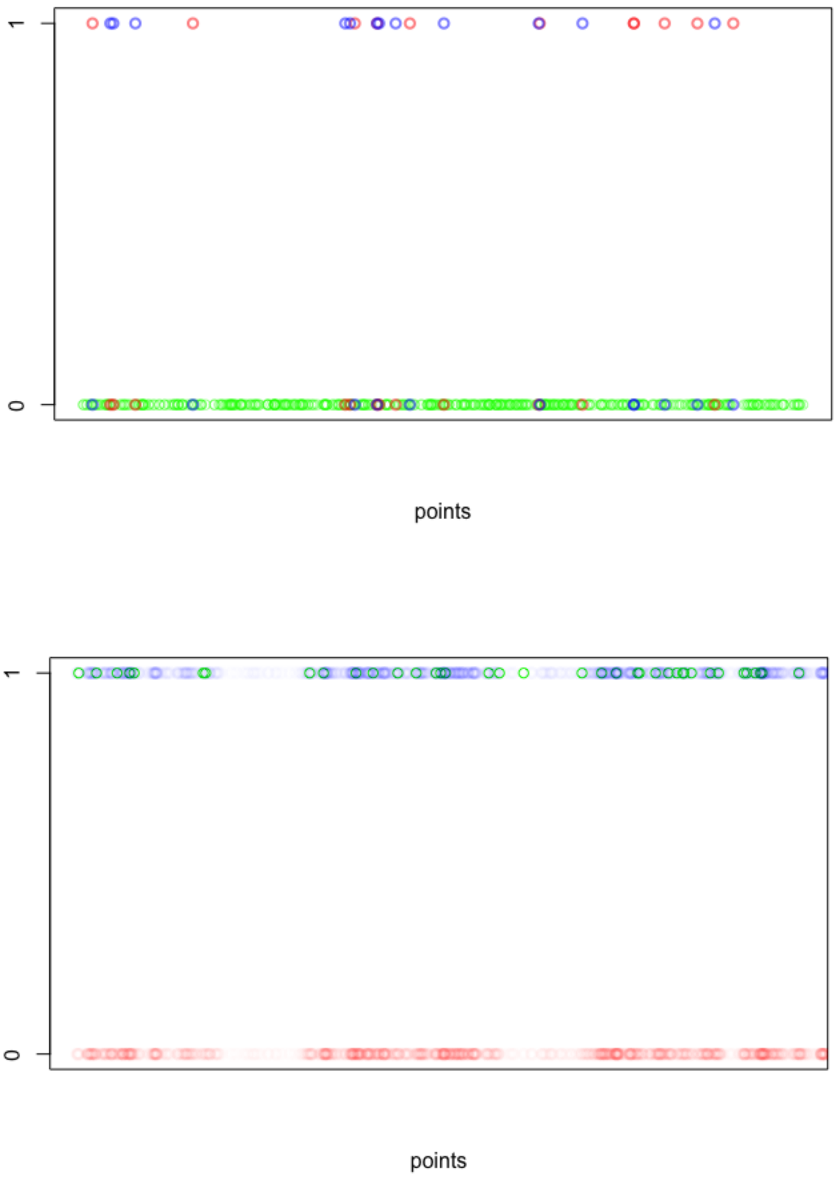

Figure 11: Accuracy of prediction of 500 randomly selected accidents (top) and all fatal accidents (bottom) in 2017. The green points indicate the accidents in which the response variable has been correctly predicted. Alternatively, the red and blue points indicate the accidents in which the response variable has not been predicted well. In those cases, the red color is the predicted value and the blue color is the true value. We can observe that the prediction is good for non-fatal accidents but bad for the fatal ones.

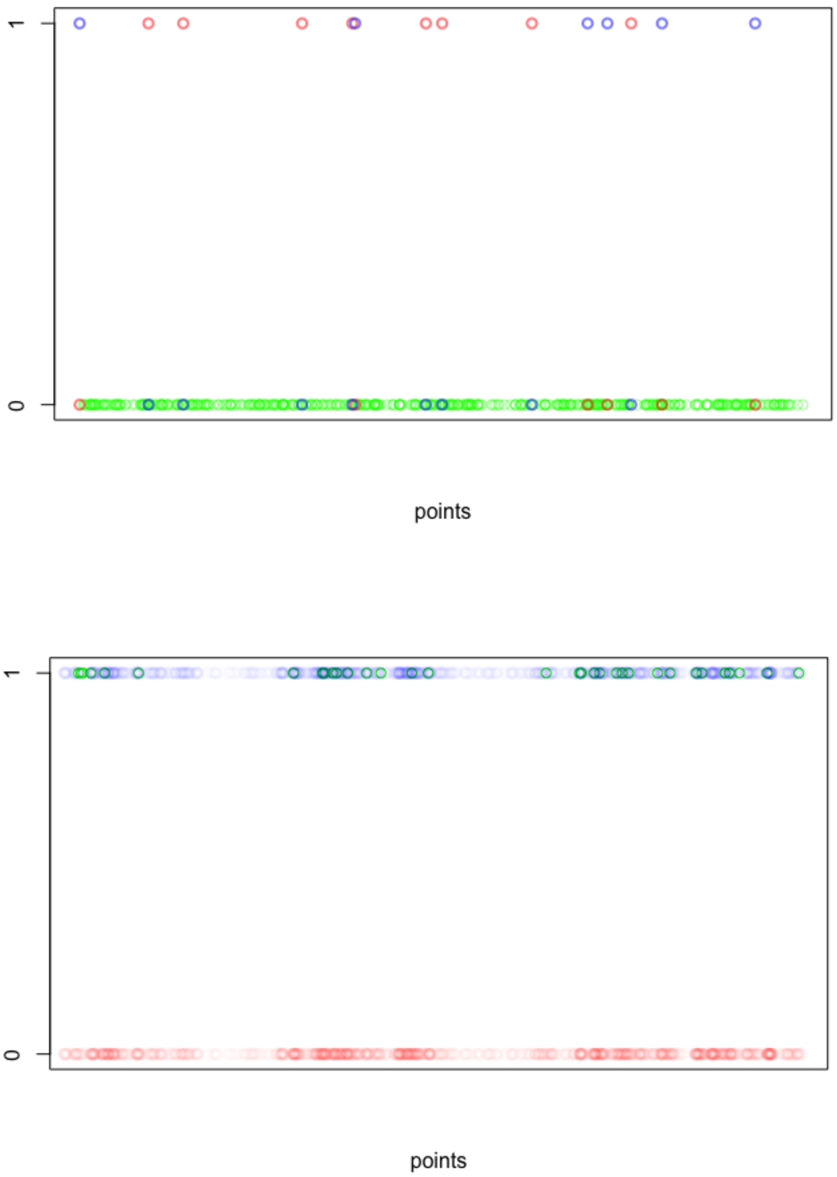

Figure 12: Accuracy of prediction of 500 randomly selected accidents (top) and all fatal accidents (bottom) in 2018. The green points indicate the accidents in which the response variable has been correctly predicted. Alternatively, the red and blue points indicate the accidents in which the response variable has not been predicted well. In those cases, the red color is the predicted value and the blue color is the true value. We can observe that the prediction is good for non-fatal accidents but bad for the fatal ones.

3 Simplified Bayesian Model

The second Bayesian model proposed, similarly as before, is a regression model in which the event of an accident being fatal y follows a Bernoulli distribution with $y = 0$ when there are not any fatalities and $y = 1$ when there is at least one dead person. Thus, θ will define the probability of an accident leading to at least one fatality.

$$y \sim \text{Bernoulli}(\theta)$$

In this case, we will use only the two explanatory variables that had a highest influence in the first model (the type of road and the time at which the accident occurs). Hence, the second Bayesian model is a simplified version of the first model proposed. The probability of an accident being fatal θ will follow the expression:

$$\log\left(\frac{\theta}{1-\theta}\right) = \beta_0 + \beta_1 Z + \beta_2 T$$

where

$$Z = \begin{cases} 1 & \text{if national road} \\ 0 & \text{otherwise} \end{cases}$$

$$T = \begin{cases} 1 & \text{if night time} \\ 0 & \text{otherwise} \end{cases}$$

The prior distributions of parameters β_i (with $i = 0, 1, 2$) will be non-informative distributions as we do not have any previous knowledge about these parameters. We define all of them as uniform distributions taking values between -100 and 100 because we think that, since these parameters are modifying the slope of a straight line, this range is wide enough to include all the possible values that they could take.

3.1 Results

We obtain the following results for the parameters β_i (with $i = 0, 1, 2$):

PARAMETER	**95% CI**	**EXPECTED VALUE**
β_0	[-5.087,-4.904]	-4.993
β_1	[1.281,1.493]	1.386
β_2	[0.422,0.628]	0.529

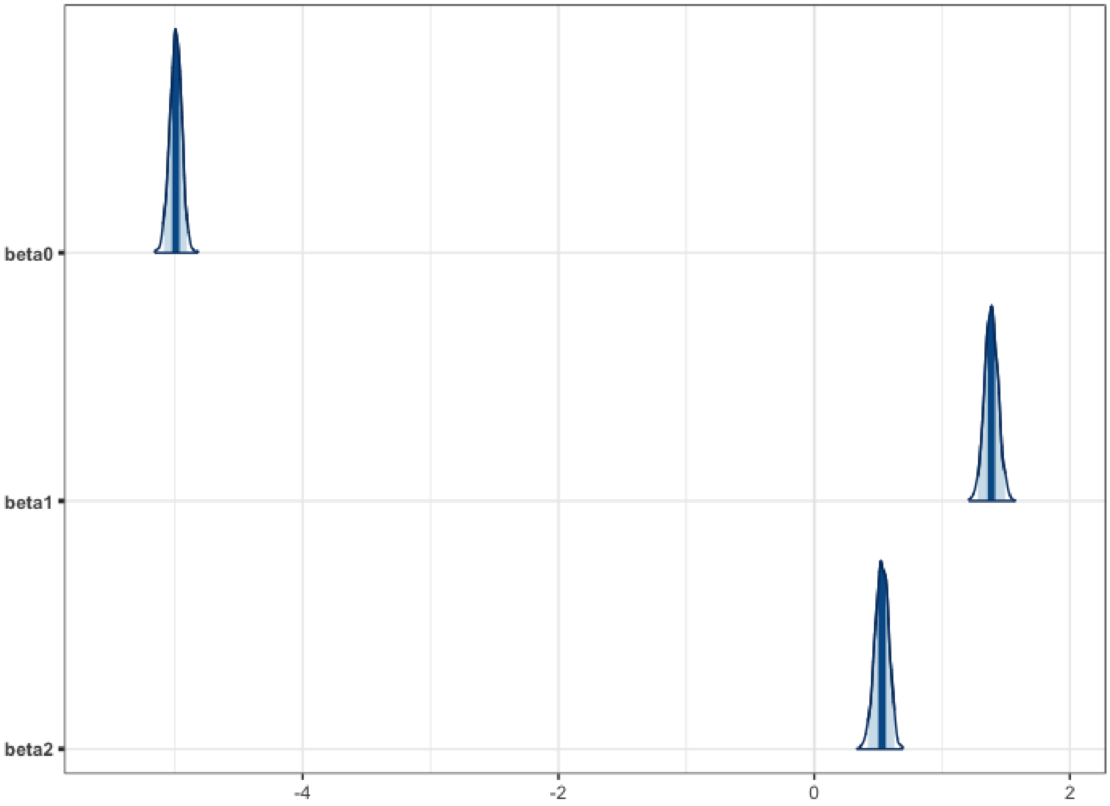

Figure 13: Posterior distributions of the parameters β_i $(i = 0, 1, 2)$ with medians and 95% credible intervals.

Using the simplified model, the distribution of the parameters β_i $(i = 0, 1, 2)$ are almost identical to the ones obtained before.

3.2 Prediction and Validation

In order to validate the first model, we try to simulate and predict results from other years (2016, 2017 and 2018).

First, we generate the distribution of theta for each individual accident of the years 2016, 2017 and 2018 using the explanatory variables (Z and T) of this accident and the distributions of beta previously obtained using 2019 data. From each distribution of theta, we can simulate the number of expected fatal accidents in a year using the Bernoulli distribution. We plot in a histogram the expected number of of fatal accidents out of the total predicted by the explanatory variables of each accident. Furthermore, we calculate the mean of all the values of the total number of fatal accidents obtained and compare it with the real value.

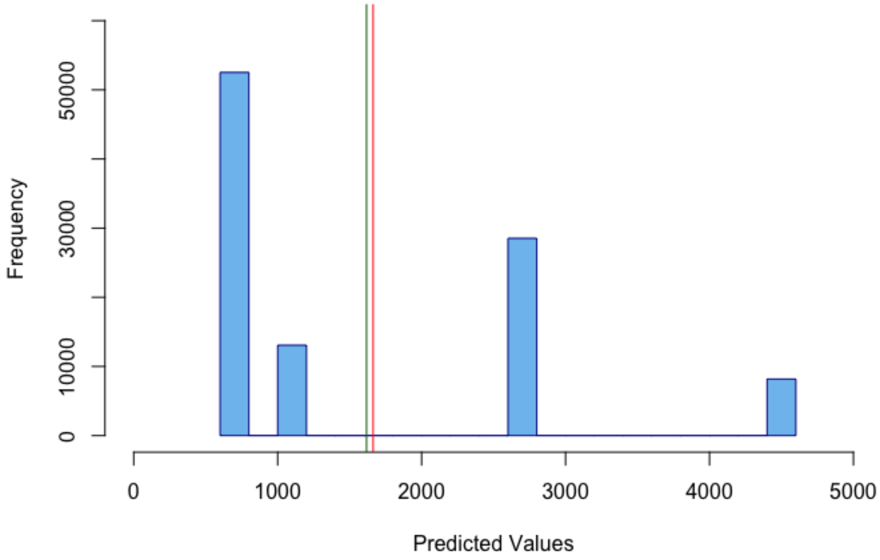

Figure 14: Predicted numbers of fatal accidents in 2016. The red line corresponds to the real number of fatal accidents and the green line, to the mean predicted value.

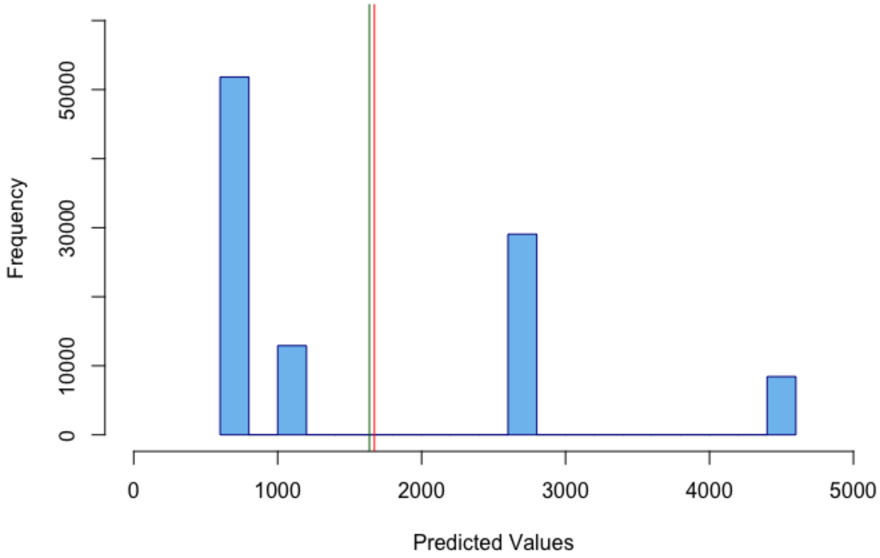

Figure 15: Predicted numbers of fatal accidents in 2017. The red line corresponds to the real number of fatal accidents and the green line, to the mean predicted value.

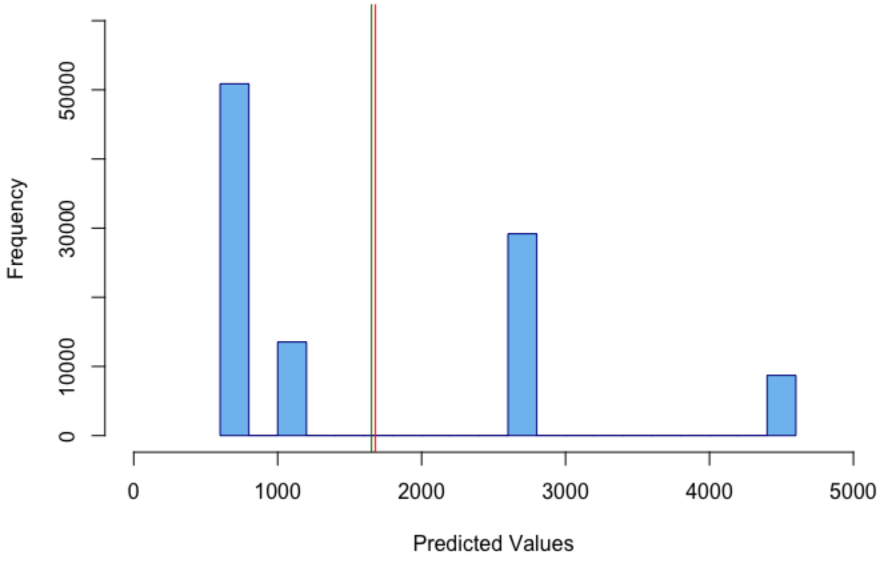

Figure 16: Predicted numbers of fatal accidents in 2018. The red line corresponds to the real number of fatal accidents and the green line, to the mean predicted value.

YEAR	ACCIDENTS	REAL NUMBER OF FATAL ACCIDENTS	PREDICTED NUMBER OF FATAL ACCIDENTS
2016	102363	1663	1617
2017	102234	1672	1637
2018	102300	1679	1652

From the obtained results it seems that the modified Bayesian model predicts the proportion of deadly accidents with high accuracy, somewhat better than the previous one. Again, the predicted values are slightly below the real values.

As before, we try to validate the model by predicting, given the explanatory variables of each individual accident (making use of data from 2016, 2017 and 2018), the response variable of each of those accidents using the parameter distributions obtained and check if the event of an accident being fatal or not is correctly predicted in each case. The result we get is that the simplified Bayesian model proposed predicts the response variable with about a 97% accuracy. However, while the accuracy when predicting the non-fatal accidents is as high as 98.5%, less than 1% of fatal accidents are correctly predicted. Hence, similarly as before, we conclude that, even though the model seems to give a precise number of fatal accidents out of the total, it is not good for predicting if a certain accident is going to be fatal given the values of the explanatory variables of such accidents.

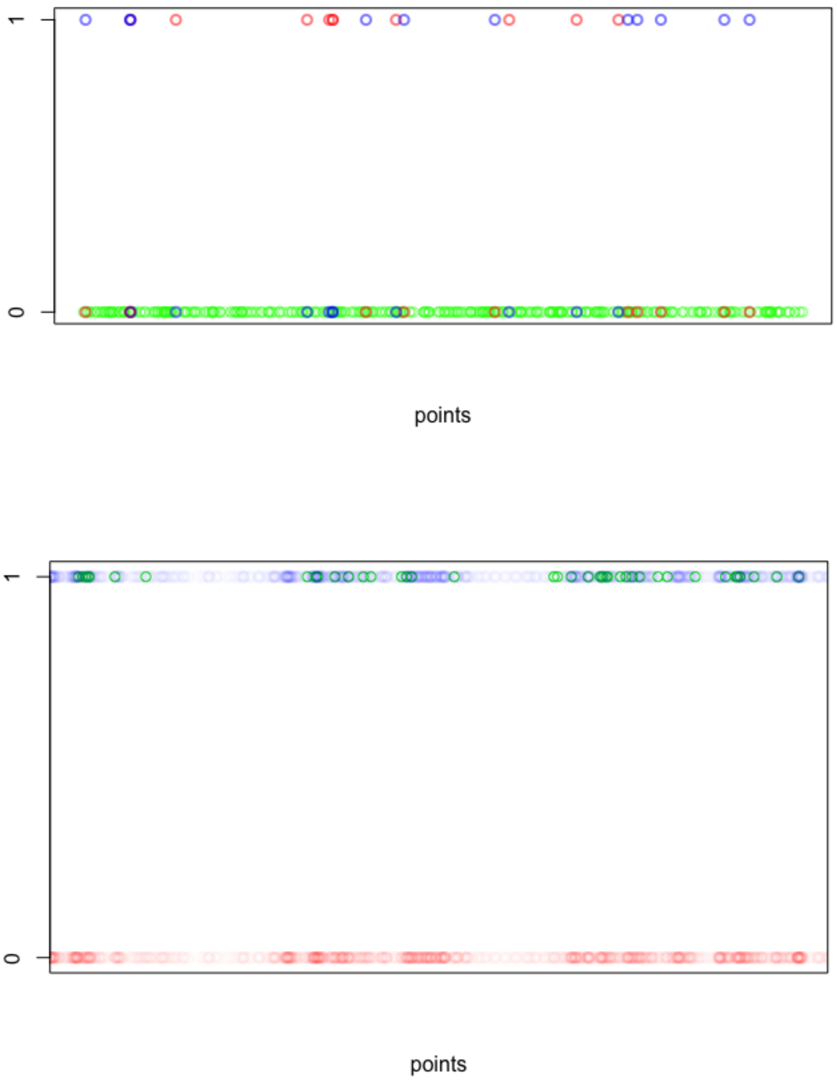

Figure 17: Accuracy of prediction of 500 randomly selected accidents (top) and all fatal accidents (bottom) in 2016. The green points indicate the accidents in which the response variable has been correctly predicted. Alternatively, the red and blue points indicate the accidents in which the response variable has not been predicted well. In those cases, the red color is the predicted value and the blue color is the true value. We can observe that the prediction is good for non-fatal accidents but bad for the fatal ones.

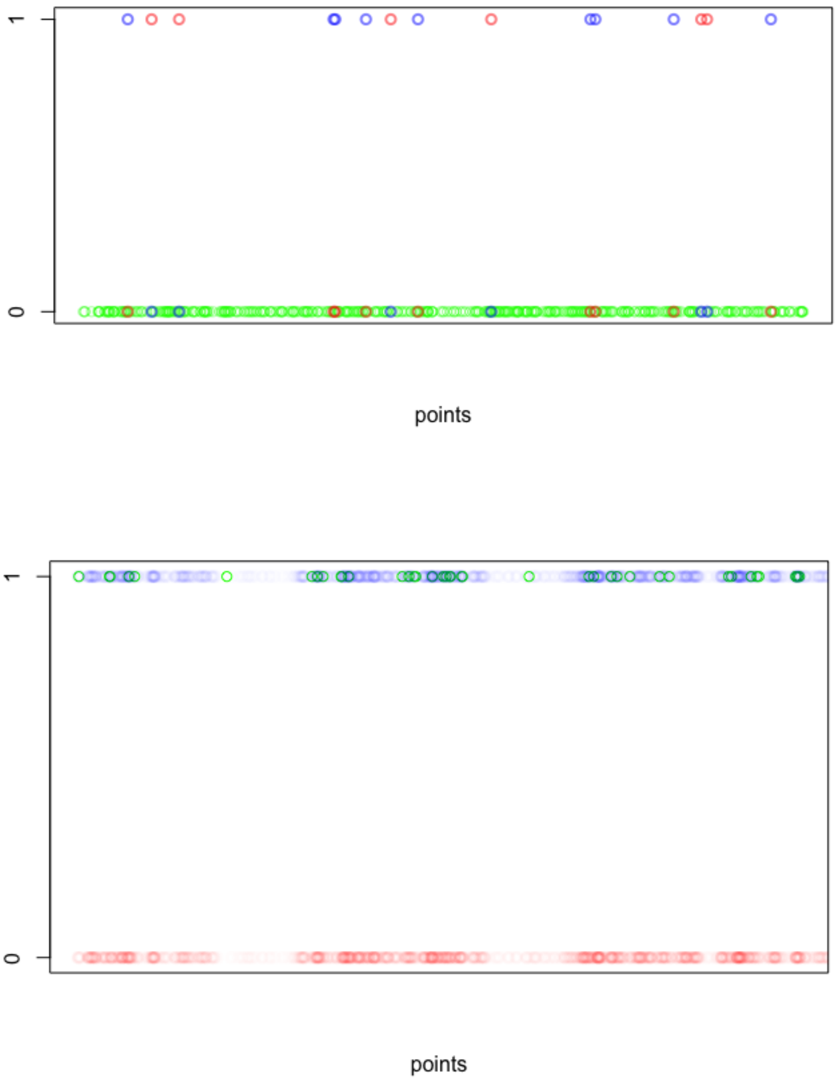

Figure 18: Accuracy of prediction of 500 randomly selected accidents (top) and all fatal accidents (bottom) in 2017. The green points indicate the accidents in which the response variable has been correctly predicted. Alternatively, the red and blue points indicate the accidents in which the response variable has not been predicted well. In those cases, the red color is the predicted value and the blue color is the true value. We can observe that the prediction is good for non-fatal accidents but bad for the fatal ones.

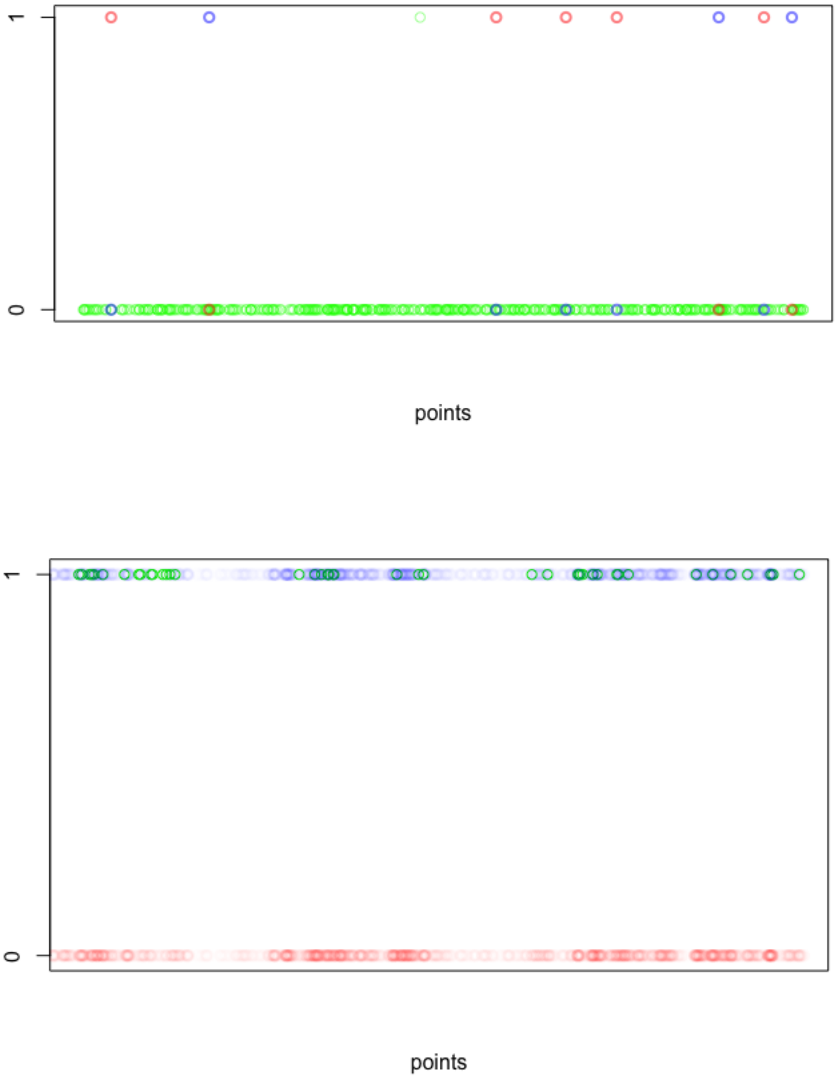

Figure 19: Accuracy of prediction of 500 randomly selected accidents (top) and all fatal accidents (bottom) in 2018. The green points indicate the accidents in which the response variable has been correctly predicted. Alternatively, the red and blue points indicate the accidents in which the response variable has not been predicted well. In those cases, the red color is the predicted value and the blue color is the true value. We can observe that the prediction is good for non-fatal accidents but bad for the fatal ones.

4 More about Prediction and Validation

In this section we give more details about the prediction and validation of our Bayesian models. We have previously seen that the models proposed seem to be inaccurate when predicting if an accident is fatal or not. However, we think this might be because we are not defining the correct classification rule for our classification problem. The aim of this section is to propose a more accurate rule for classifying accidents into fatal and non-fatal.

Firstly, the Bayesian model provides us with the posterior distribution of the parameters β_i. In this case, we consider as a point estimator of these parameters the expected value. Then, for accident (prediction) we will obtain a value of θ, which represents the probability that the traffic accident is fatal. If the value of this probability is greater than a threshold (cut point c), we would classify the accident as fatal. The cut point $c = 0.025$ seems to be accurate enough, leading to good performance in the prediction. We use data from 2018 in order to predict the accuracy of both models with the new proposed value of the cut point.

4.1 Full Bayesian Model

We first use the full Bayesian model in which all explanatory variables are considered.

The following table shows the confusion matrix when prediction traffic accidents from 2018:

		Predicted	
		Non-fatal	Fatal
Actual	Non-fatal	69509	31111
	Fatal	636	1043

Table 1: Full Bayesian model. Confusion matrix for traffic accidents with victims registered in 2018. Accuracy: 0.69, Sensitivity: 0.69, Specificity: 0.62. Classification rule: c = 0.025.

4.2 Simplified Bayesian Model

We consider the simplified Bayesian model in which only the type of road and the time slot of the day are used as explanatory variables.

The following table shows the confusion matrix when prediction traffic accidents from 2018:

		Predicted	
		Non-fatal	Fatal
Actual	Non-fatal	63923	36697
	Fatal	484	1195

Table 2: Confusion matrix for traffic accidents with victims registered in 2018. Accuracy: 0.64, Sensitivity: 0.64, Specificity: 0.71. Classification rule: c = 0.025.

4.3 Comparison and Summary

The probability of correct classification for both models is similar, although for the simplified Bayesian model this probability is slightly higher.

$$\text{Probability of correct classification (full model)} = 0.66$$

$$\text{Probability of correct classification (simplified model)} = 0.68$$

In summary, given an accident with victims, if we know the type of road and the time when it took place, the proposed simplified Bayesian model is able to predict, with probability 2/3, whether or not there was at least one death.

If the cut point is increased, the accuracy increases but the specificity decreases and almost all accidents are classified as non-fatal, which is what happened in the earlier prediction before changing the cut point.

5 Differentiating Regions

In the data set we find data for the different regions (*Comunidades Autónomas*) in Spain, for which the result and impact of the different variables of use could have important differences. For instance, the amount of accidents varies between regions, as well as the percentage of fatal ones. We believe that different results could be obtained for the distributions of the parameters if we differentiate by regions and, hence, it is interesting to do some research on it. To do so, we develop an unpooled and a hierarchical model, and study the differences between the two, as well as the dissimilarities with the original model.

5.1 Unpooled Bayesian Model

In this part, we define an unpooled model for the probability of an accident being fatal for each of the different regions.
Considering that we have 19 different groups in our data set (on for each region where the accident happened), our aim is to study the probability of an accident being fatal, θ_j, in the j^{th} group so that this parameter can be changed in each group.

We denote as y_{ij} the observation of the accident i in group j, which can take values 0 (if it is not fatal) or 1 (if it is fatal). Thus, the unpooled model is the following:

$$y_{ij} \sim \text{Bernoulli}(\theta_j)$$

As it is a regression model, the probability of an accident being fatal in each region, θ_j, will have the following expression:

$$\log\left(\frac{\theta_j}{1-\theta_j}\right) = \beta_0 + \beta_1 Z + \beta_2 T + \beta_3 M + \beta_4 H + \beta_5 E$$

where Z, T, M, H and E have the same meaning as before and the prior distributions of parameters β_l (with $l = 0, 1, 2, 3, 4, 5$) will be non-informative distributions, considered as uniforms between -100 and 100.

5.1.1 Results

The unpooled model provides the following results in each community. The posterior distributions for each community:

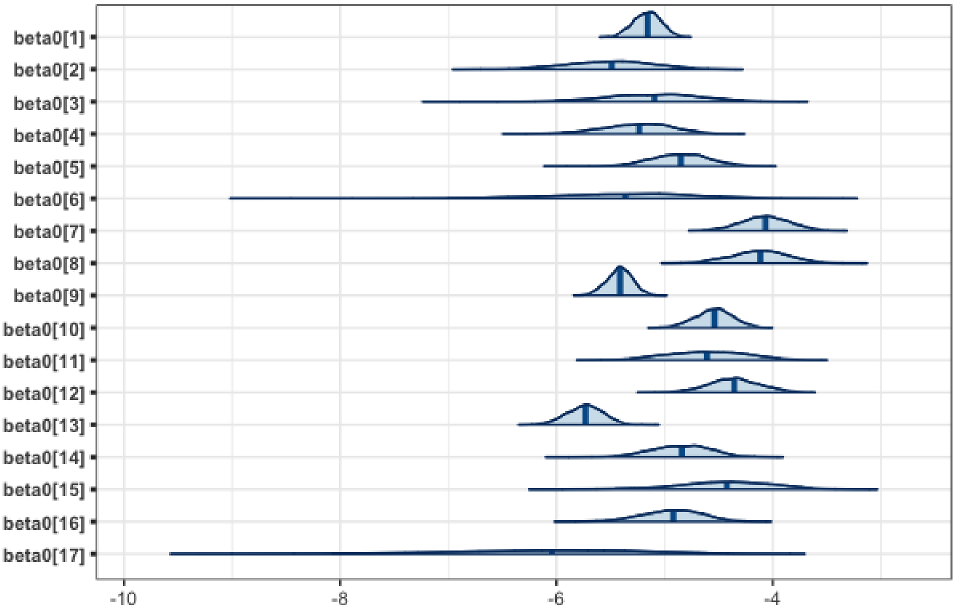

Figure 20: Posterior distributions of the parameter β_0 with medians and 95% credible intervals with respect to each community.

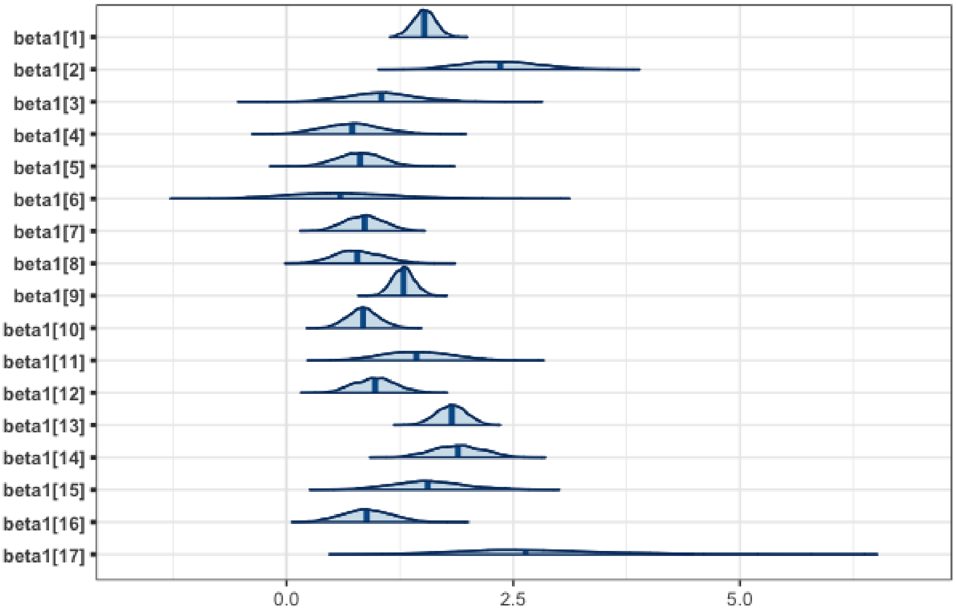

Figure 21: Posterior distributions of the parameter β_1 with medians and 95% credible intervals with respect to each community.

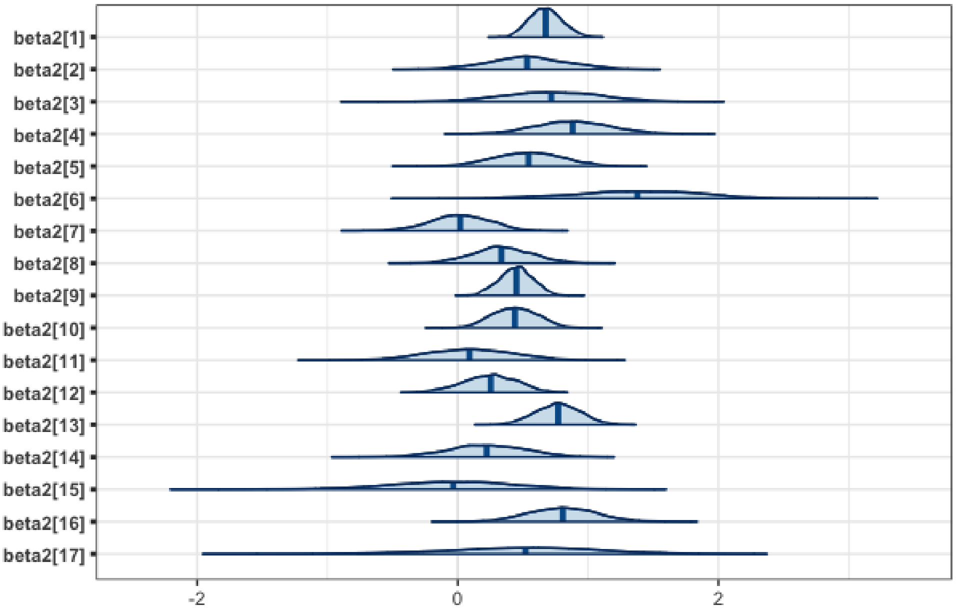

Figure 22: Posterior distributions of the parameter β_2 with medians and 95% credible intervals with respect to each community.

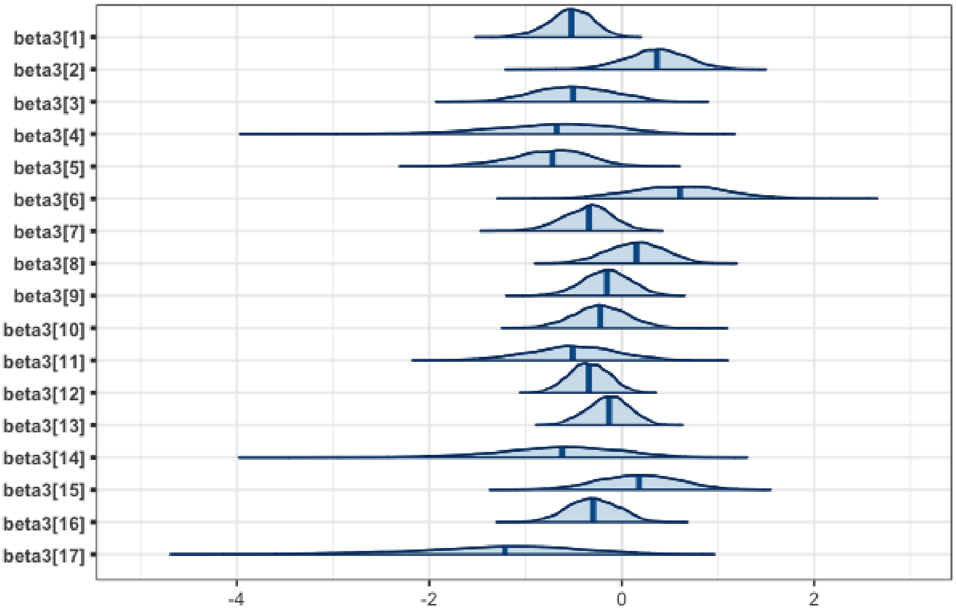

Figure 23: Posterior distributions of the parameter β_3 with medians and 95% credible intervals with respect to each community.

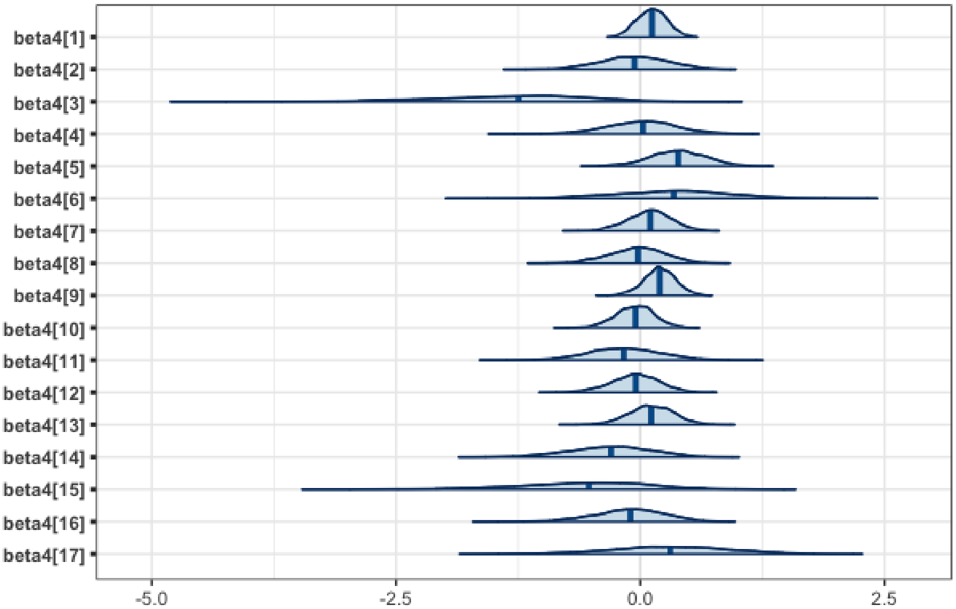

Figure 24: Posterior distributions of the parameter β_4 with medians and 95% credible intervals with respect to each community.

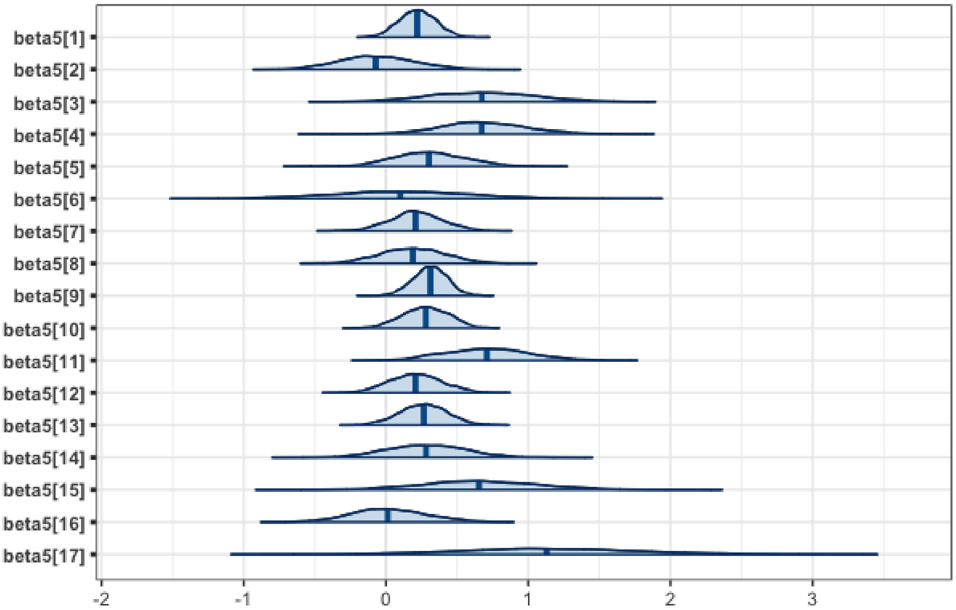

Figure 25: Posterior distributions of the parameter β_5 with medians and 95% credible intervals with respect to each community.

Because the last two regions (*Ceuta* and *Melilla*) have less data, the β parameters for them have a much larger variance and, hence, we plot them separately in order to obtain a better visualisation.

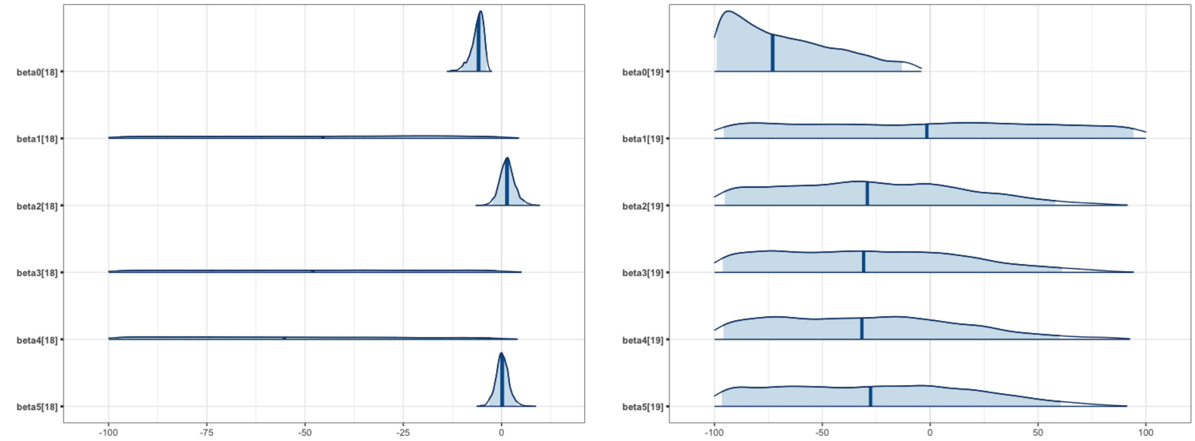

Figure 26: Posterior distributions of the parameters β_i $(i = 1, 2, 3, 4, 5)$ with medians and 95% credible intervals for the regions *Ceuta* (left) and *Melilla* (right).

The values of the following table represent the mean value of each parameter β_l for $l = 0, ..., 5$ with respect to each community.

COMMUNITIES	β_0	β_1	β_2	β_3	β_4	β_5
1	-5.160034	1.5178648	0.67603618	-0.5328037	0.11853696	0.22130107
2	-5.506181	2.3695506	0.53723422	0.3525167	-0.06567012	-0.06466991
3	-5.110131	1.0577466	0.71621753	-0.5024294	-1.31775638	0.67621985
4	-5.256699	0.7240936	0.88476510	-0.7351505	0.01387449	0.68505181
5	-4.857214	0.8098382	0.53867025	-0.7481468	0.38641845	0.30628333
6	-5.418619	0.6275879	1.36568434	0.5918122	0.31474454	0.10424501
7	-4.068374	0.8582823	0.01693186	-0.3598939	0.09447853	0.21239444
8	-4.123377	0.7923549	0.33795391	0.1491422	-0.03460774	0.18880388
9	-5.415729	1.2873621	0.44681633	-0.1602462	0.19860899	0.31036381
10	-4.545601	0.8448466	0.43794245	-0.2244411	-0.05738245	0.27796250
11	-4.611718	1.4384494	0.08909064	-0.5196155	-0.16637359	0.71139574
12	-4.357535	0.9751822	0.24465463	-0.3394462	-0.05656519	0.21028524
13	-5.737554	1.8167059	0.77031235	-0.1392352	0.11116835	0.26625731
14	-4.848261	1.8877146	0.22233867	-0.6597129	-0.31489111	0.28311787
15	-4.445956	1.5689599	-0.04467325	0.1761981	-0.58004143	0.66425066
16	-4.934723	0.8892216	0.80548697	-0.2971835	-0.11588502	0.02209310
17	-6.118155	2.7084407	0.49159215	-1.2914888	0.30020558	1.15161508
18	-6.144853	-47.3915358	1.45839343	-48.7188722	-53.46038728	0.21117302
19	-67.915973	-2.5618343	-26.95874470	-28.4674779	-29.63435954	-26.43377408

We can appreciate that β_1 has a high impact on the 2^{nd} and 17^{th} regions. With respect to β_2, regions 6 and 18 have a higher values of it. β_3 is significant in the region 6 and β_4 in regions 5,6 and 17. Finally, we can observe that the highest value of $beta_5$ has the regions 17 and 11.

5.2 Hierarchical Bayesian Model

Hierarchical models involve several parameters in such a way that the distributions of some depend significantly on the values of others. That is, when separating the data in regions, the posterior distributions of the model parameters are different, but they are still influenced by one another, and those with more data will have a larger effect.

We define a hierarchical model for the probability of an accident being fatal, θ_j, in each region j. We consider a Bernoulli distribution:

$$y_{ij} \sim \text{Bernoulli}(\theta_j)$$

As we are using a regression model, the probability of an accident being fatal in each region, θ_j, will have the following expression:

$$\log\left(\frac{\theta_j}{1-\theta_j}\right) = \beta_0 + \beta_1 Z + \beta_2 T + \beta_3 M + \beta_4 H + \beta_5 E$$

where Z, T, M, H and E have the same meaning as before and the prior distributions of parameters β_l (with $l = 0, 1, 2, 3, 4, 5$) will be non-informative distributions:

$$\beta_l \sim \text{Normal}(a, b)$$

Finally, we assign a uniform distribution to the hyperparameters a and b:

$$a \sim \text{Uniform}(-10, 10)$$

$$b \sim \text{Uniform}(0, 100)$$

5.2.1 Results

The hierarchical model provides the following results in each community. The posterior distributions for each community:

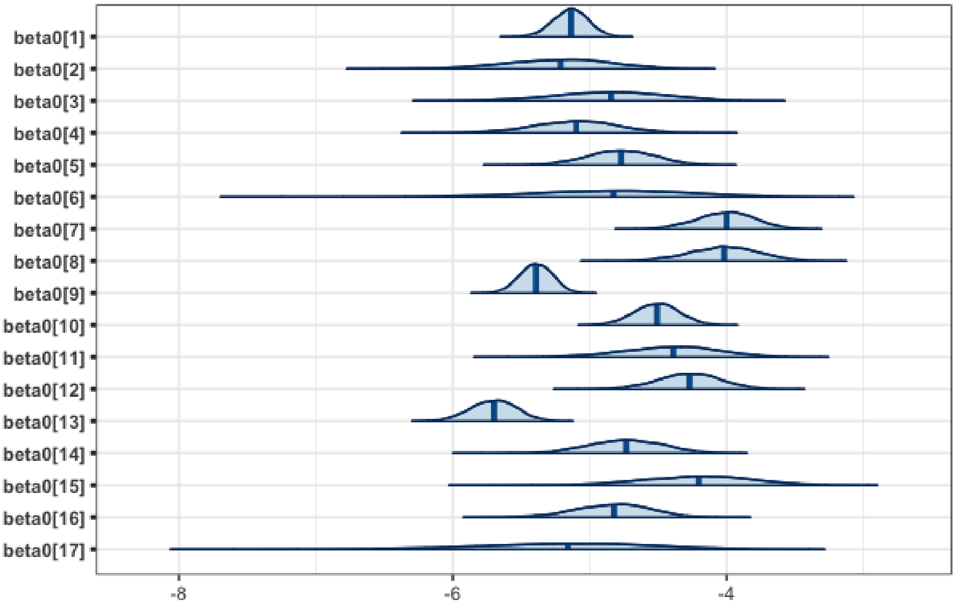

Figure 27: Posterior distributions of the parameter β_0 with medians and 95% credible intervals with respect to each community.

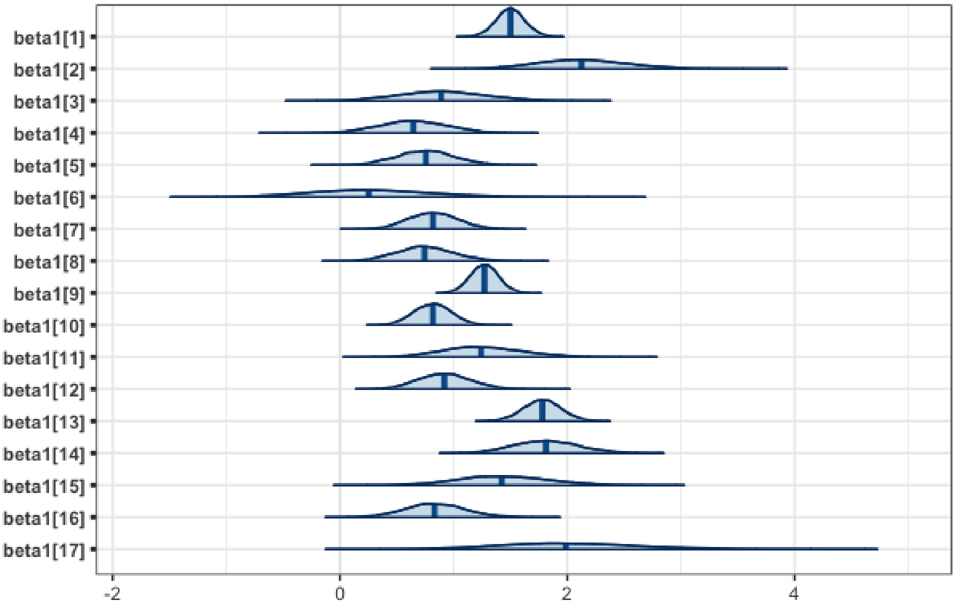

Figure 28: Posterior distributions of the parameter β_1 with medians and 95% credible intervals with respect to each community.

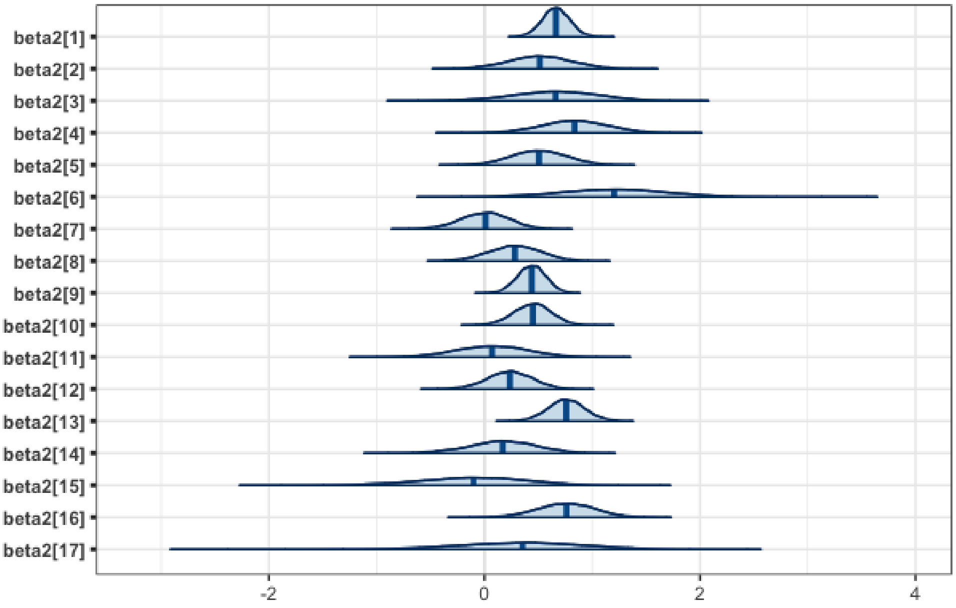

Figure 29: Posterior distributions of the parameter β_2 with medians and 95% credible intervals with respect to each community.

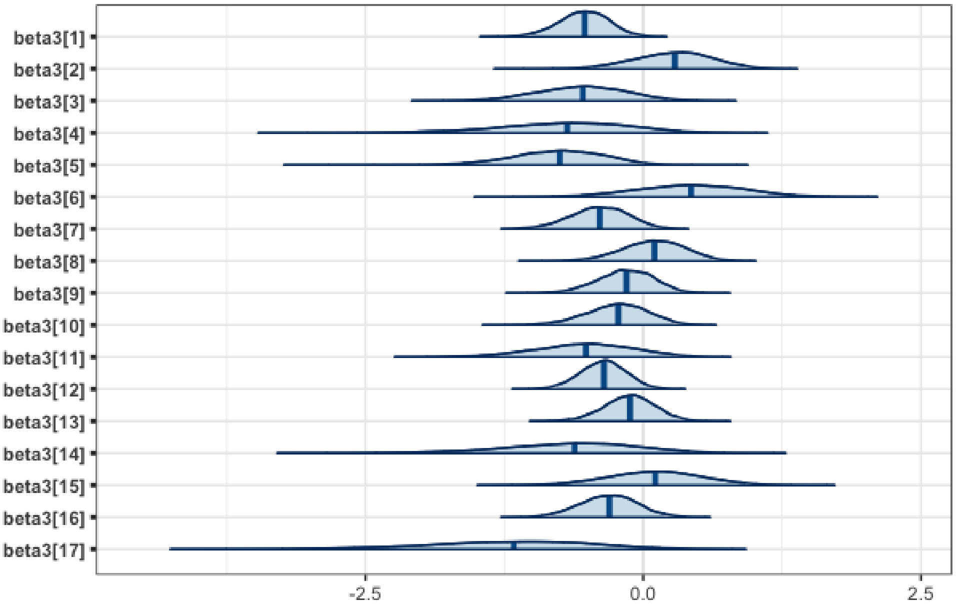

Figure 30: Posterior distributions of the parameter β_3 with medians and 95% credible intervals with respect to each community.

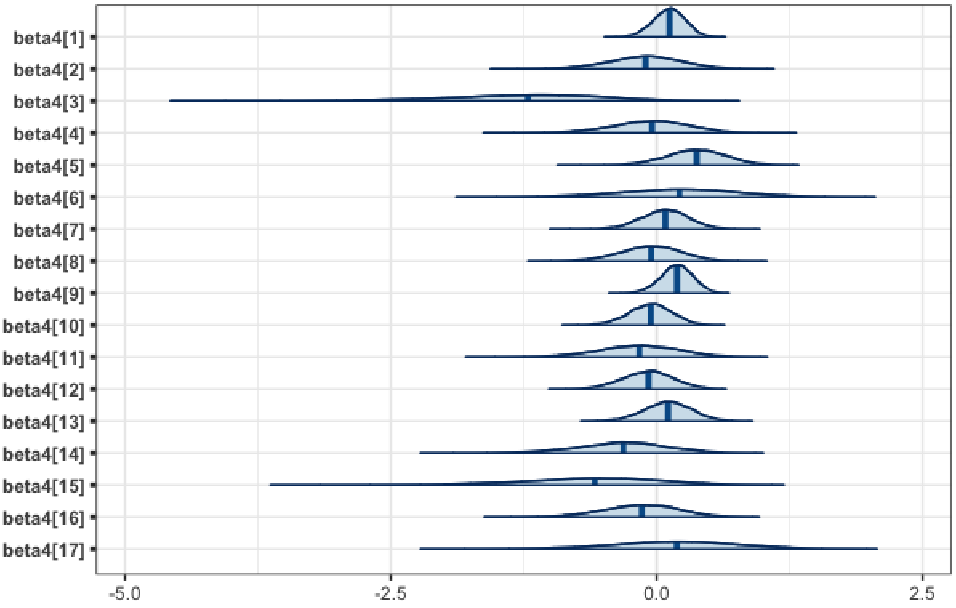

Figure 31: Posterior distributions of the parameter β_4 with medians and 95% credible intervals with respect to each community.

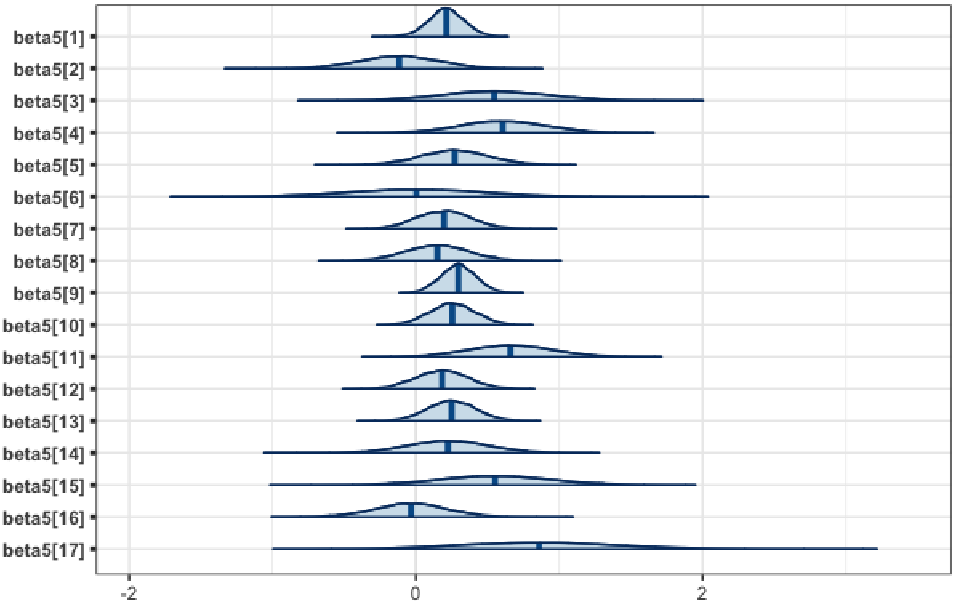

Figure 32: Posterior distributions of the parameter β_5 with medians and 95% credible intervals with respect to each community.

Because the last two regions (*Ceuta* and *Melilla*) have less data, the β parameters for them have a much larger variance and, hence, we plot them separately in order to obtain a better visualisation.

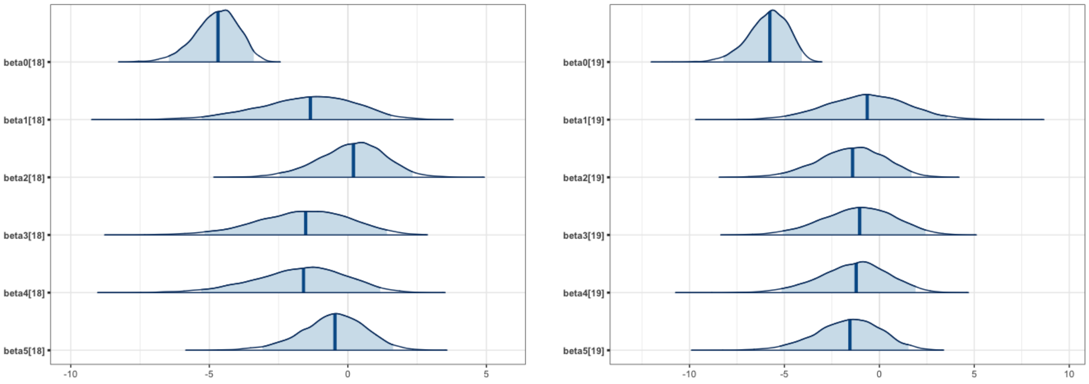

Figure 33: Posterior distributions of the parameters β_i $(i = 1, 2, 3, 4, 5)$ with medians and 95% credible intervals for the regions *Ceuta* (left) and *Melilla* (right).

The values of the following table represent the mean value of each parameter β_l for $l = 0, ..., 5$ with respect to each community.

COMMUNITIES	β_0	β_1	β_2	β_3	β_4	β_5
1	-5.139544	1.501025	0.6638797	-0.5360669	0.1171517	0.2138458
2	-5.235962	2.140049	0.5129536	0.2693747	-0.1147699	-0.119448
3	-4.856027	0.8991892	0.6532744	-0.5561966	-1.267177	0.5493901
4	-5.107487	0.6465512	0.8352997	-0.7259409	-0.05886253	0.6078631
5	-4.773373	0.7540492	0.5047915	-0.7692286	0.3733058	0.2719174
6	-4.845782	0.2771541	1.197119	0.4243067	0.2016703	0.003571044
7	-4.008199	0.8203493	0.004782583	-0.3966394	0.07564224	0.1946972
8	-4.021797	0.7504035	0.2810549	0.09450357	-0.06076106	0.1488676
9	-5.391336	1.270828	0.4367513	-0.1585975	0.1876585	0.2966312
10	-4.510724	0.8170777	0.445011	-0.2380239	-0.06075391	0.2558053
11	-4.408061	1.257542	0.06446758	-0.5376462	-0.172362	0.6604253
12	-4.275921	0.9215622	0.2328372	-0.3600652	-0.08443132	0.1819389
13	-5.702878	1.780477	0.7604651	-0.1255022	0.1045843	0.2504111
14	-4.737864	1.815785	0.1669037	-0.6665527	-0.330634	0.2246537
15	-4.22085	1.437961	-0.1149111	0.1023725	-0.6203392	0.5537564
16	-4.833206	0.8386512	0.7556018	-0.3117614	-0.1528129	-0.03589133
17	-5.195846	2.023925	0.329363	-1.217478	0.1765413	0.8642478
18	-4.745743	-1.499268	0.1335809	-1.611855	-1.722964	-0.5274259
19	-5.865938	-0.6604871	-1.496242	-1.119081	-1.348944	-1.642314

We can appreciate that β_1 has a high impact on the 2^{nd} and 17^{th} regions. With respect to β_2, regions 6, 13 and 16 have a higher values of it. β_3 is significant in the region 6 and β_4 in regions 5 and 6. Finally, we can observe that the highest value of β_5 has the regions 4 and 10.

5.3 Comparison of the models

In the following plots one can observe the comparison between the pooled, the unpooled and the hierarchical Bayesian models.

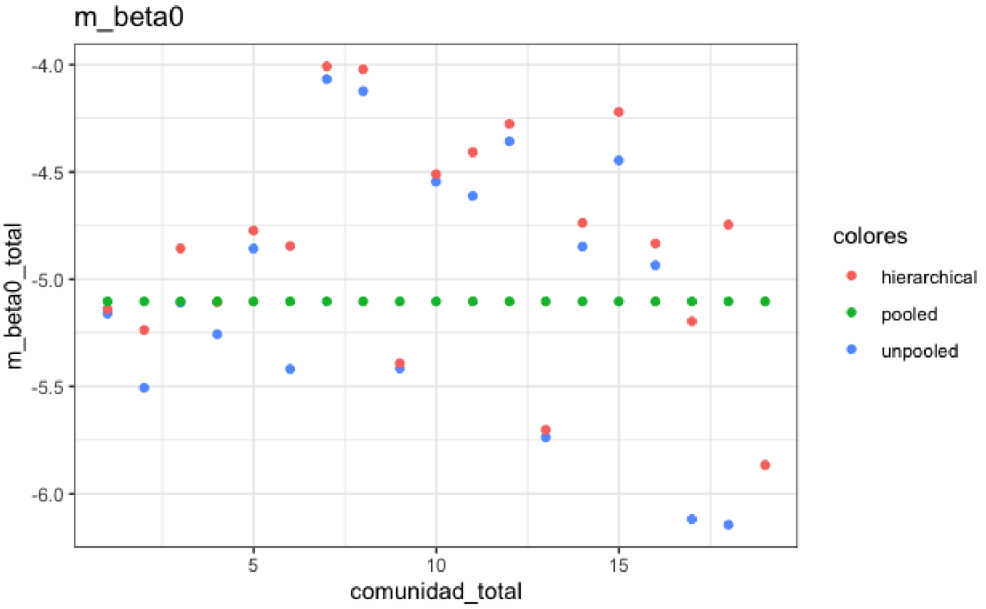

Figure 34: Comparison of the obtained values of β_0 using the hierarchical, pooled and unpooled models. The x-axis represent the community and the y-axis the mean values of the parameter β_0 for each model.

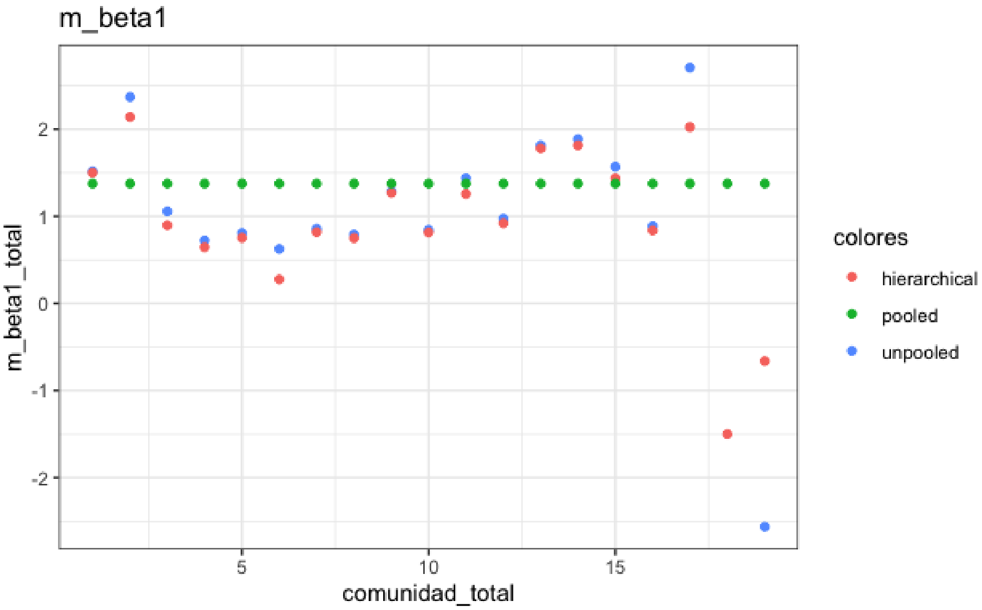

Figure 35: Comparison of the obtained values of β_1 using the hierarchical, pooled and unpooled models. The x-axis represent the community and the y-axis the mean values of the parameter β_1 for each model.

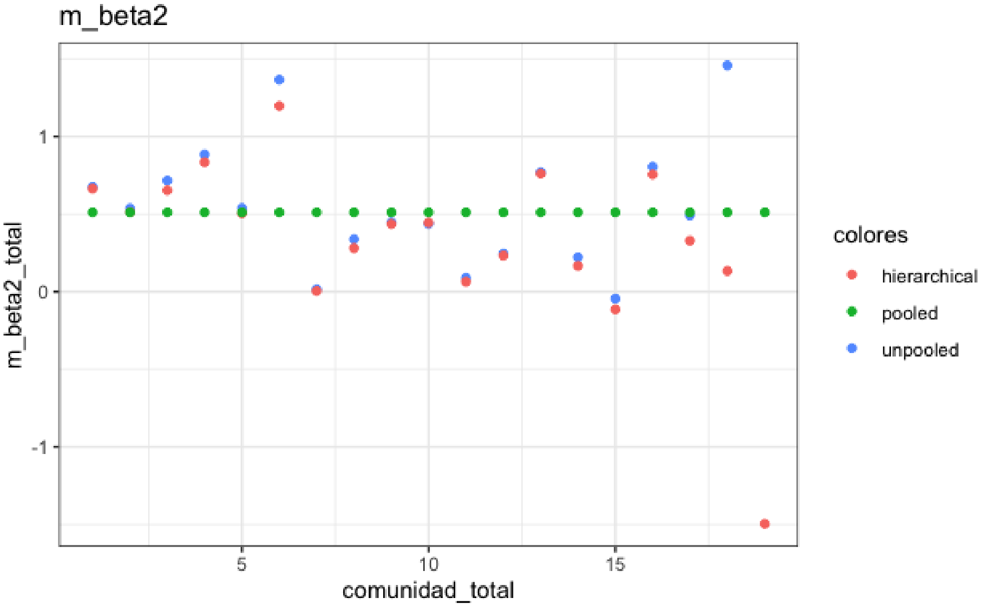

Figure 36: Comparison of the obtained values of β_2 using the hierarchical, pooled and unpooled models. The x-axis represent the community and the y-axis the mean values of the parameter β_2 for each model.

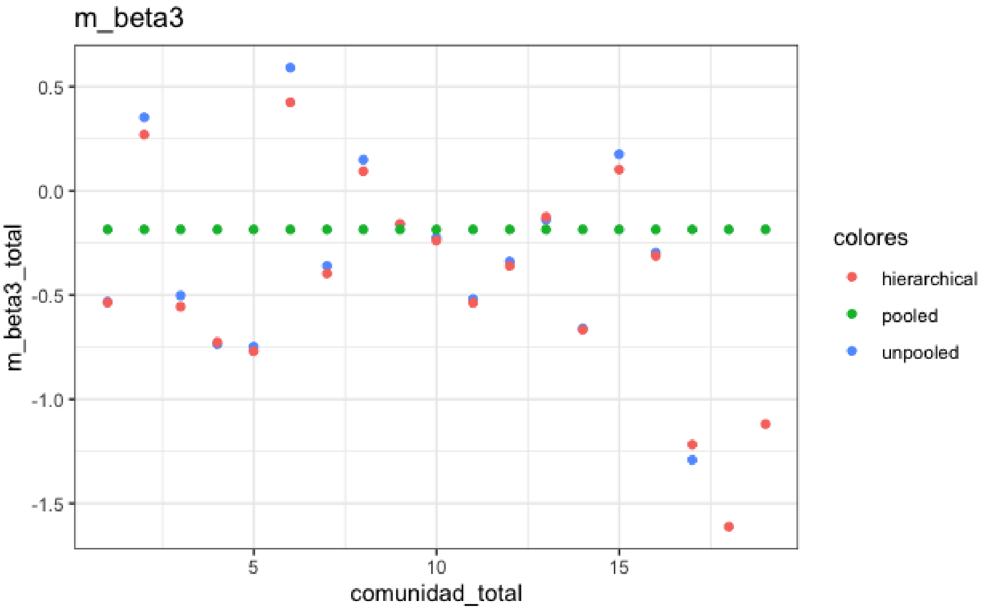

Figure 37: Comparison of the obtained values of β_3 using the hierarchical, pooled and unpooled models. The x-axis represent the community and the y-axis the mean values of the parameter β_3 for each model.

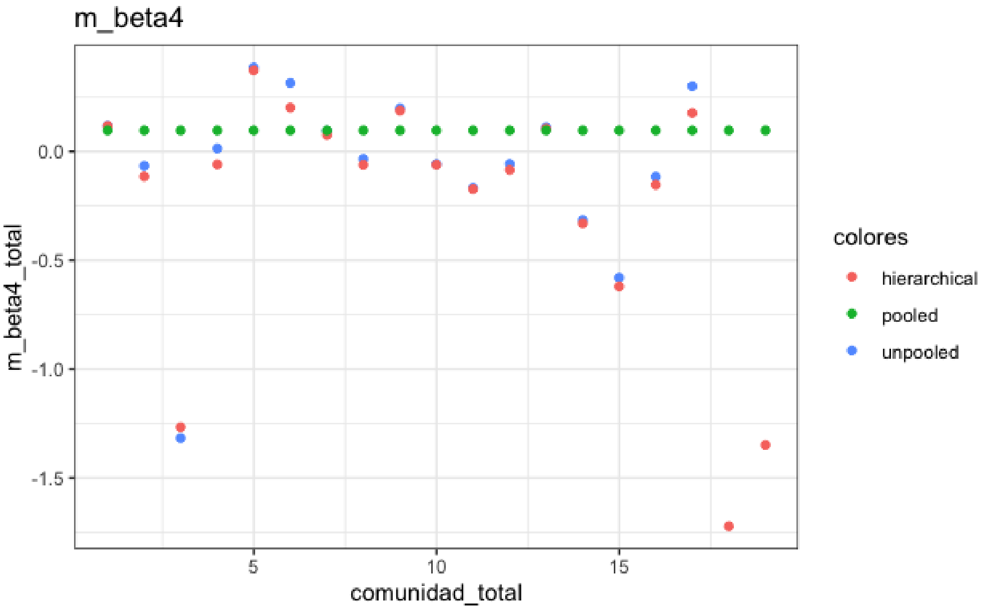

Figure 38: Comparison of the obtained values of β_4 using the hierarchical, pooled and unpooled models. The x-axis represent the community and the y-axis the mean values of the parameter β_4 for each model.

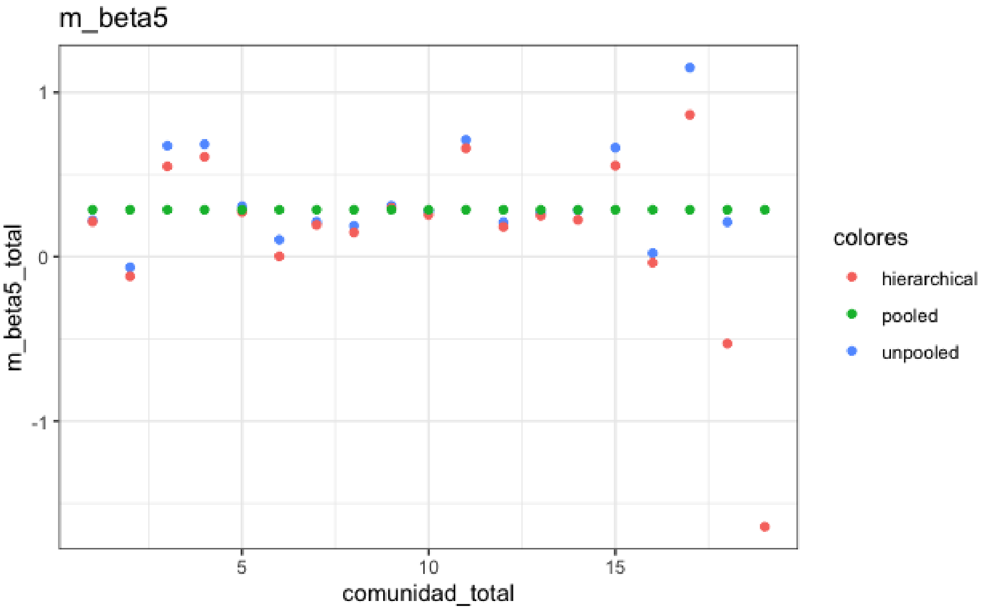

Figure 39: Comparison of the obtained values of β_5 using the hierarchical, pooled and unpooled models. The x-axis represent the community and the y-axis the mean values of the parameter β_5 for each model.

From the previous plots the first thing we can see is that in case of the pooled model there are not any differences between the different regions, which is obvious.

When we use the unpooled model we are defining different regions as completely independent from one another and, thus, for those regions with a small amount of data the parameters' values obtained are extreme and with large variance. Furthermore, these values of β could be unreliable, as in some cases we could have a sample that could be too small.

On the other hand, when we build the hierarchical model we are implying that the different regions are somehow independent, but still have some influence on one another. Moreover, those regions with a large number of data will have a higher weight, and vice versa. Hence, we obtain a better approximation of the real values of the parameters if we want to make predictions, as we are assuming that the variables might be different between the different regions, but still allowing some relationship between them.

6 Conclusions

Initially, we constructed two different Bayesian models in order to estimate and predict the probability of an accident leading to at least one death or not. The first model built considered a total of five different explanatory variables, two of which had a higher impact than the others on the response. Hence, we built a simplified Bayesian model with only these two explanatory variables.

When we tried to predict the number of fatal accidents out of the total number of accidents for other years, we observed that both models were providing an accurate result. However, this could have been possible because the proportion in all years were comparable, we tried to predict specifically the event of fatality for each specific accident and found out that the model was not precise at all since less than 1% of the fatal accidents were correctly predicted. We thought this could be because we were using a cut point that was too high and, consequently, was leading to a very low specificity, causing almost all accidents to be classified as non-fatal. Therefore, we tried using a smaller value of the cut point and managed to obtained a probability of correct classification approximately of 2/3 in both models.

After building and validating the main models, we studied the differences between the different regions (*Comunidades Autónomas*) in Spain. We created an unpooled model where the data is completely separated in regions and we do not allow any influence between them. The result obtained was that those regions with a larger sample have parameters' values that are closer to the pooled model and, on the contrary, the ones with a small amount of data provide extreme values of the parameters and large variances, which seems to be less reliable. Finally, we used a hierarchical model, where the data is also divided in regions, but there was influence between them. Thus, those regions with more data will have a greater influence on the others. This lead to observe results closer to the average values obtained when the regions are not differentiated, but we can still see differences between them.

References

1. https://doi.org/10.1016/j.aap.2007.04.002
2. https://www.sciencedirect.com/science/article/pii/S0001457504001186