Bayesian Modeling of Road Collision Data

Thomas Oman
Student Number: 160646902
Supervisor: Joe Matthews

May 2020

Abstract

Bayesian methods have been efficient at estimating model parameters whilst showing uncertainty in an intuitive way of posterior distributions. In this project we aim to use Bayesian methods to model the rate of road collisions in the state of Florida using spatial and seasonal effects and then more advanced modeling techniques such as kernel density estimate and conditional auto regressive models to reduce the uncertainty in the posterior distributions for the effects in the models so that the models could be used more generally for road collision data world wide.

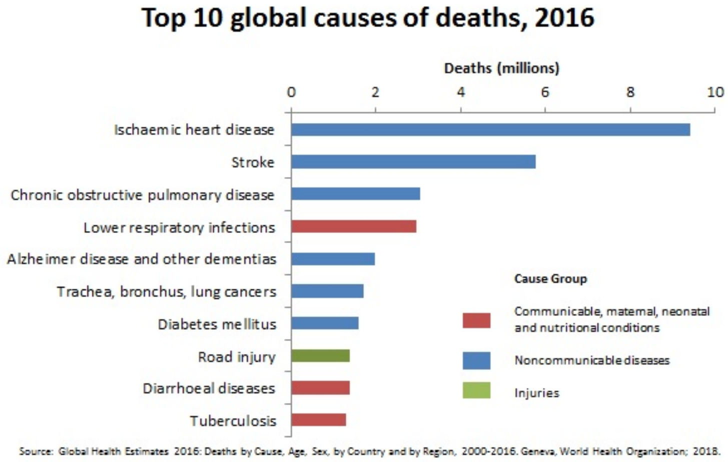
Contents

Chapter 1

Introduction

1.1 Motivation

In 2016, there was a recorded number of 56.9 million deaths worldwide with 1.4 million of these deaths attributed to road injuries, placing road injuries 8th in leading global causes of death. This position has increased since 2000 where road injuries placed 10th and whilst road injuries accounted for approximately 2.5% of deaths recording in 2016, there was a further 50 million injuries due to road collisions. Over 50% of victims of road collisions are vulnerable road users such as pedestrians, cyclist and motorcyclists meaning that simply avoiding driving does not grant safety from road collisions.

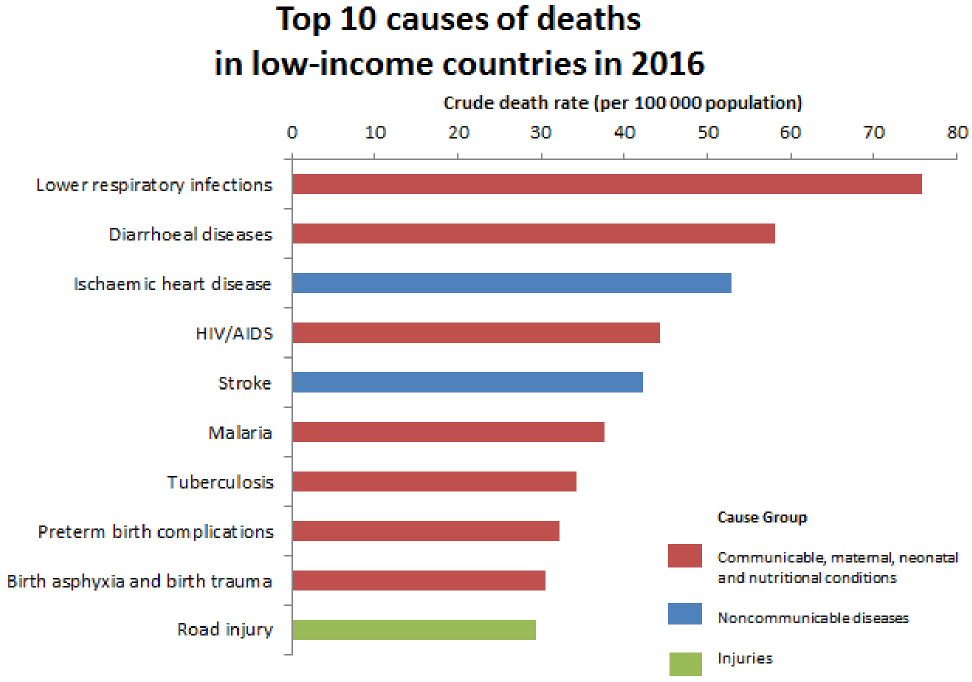


Top 10 global causes of deaths, 2016

Source: Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.

Figure 1.1: World Health Organisation top 10 global causes of death in 2016 [5]

The data show that lower and middle income countries are disproportionately affected by road collision deaths than higher income countries. Despite low and middle income countries only containing 60% of the world's vehicles they account for 93% of the worlds fatalities due to road collisions. Whilst these countries are the most affected by road collisions there is a lack of availability of road collision data in these countries for a number of reasons, one in particular being the obvious lack of funding available to set up in the infrastructure needed to record and store this data. This important point affects how we will approach modeling our data, as we will want it to be as universal as possible so that it can be used to help model data in countries of lower economic standing and also since we have very little data in comparison to the true number of collisions in other countries we will want to extract as much usefulness as possible, so that means reducing the uncertainty in our model as far as possible. Road collision deaths and injuries have wider effects on a country as its be shown to cost approximately 3% of a country's gross domestic product.[5][6]



Source: Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.
World Bank list of economies (June 2017). Washington, DC: The World Bank Group; 2017 (https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups).

Figure 1.2: World Health Organisation top 10 global causes of death In low economic countries in 2016 [6]

1.2 The Data

The data set used in this project originated from 51 traffic zones across the state of Florida taken monthly over a period of 56 years. Each data point is an average collision rate to account for varying zone sizes. In total we have 34,221 data points which represent a collision rate in a specific zone and month however 73% of these data points are missing, so by treating these points as missing at random we can exclude these observations. After removing the missing values and getting the remaining data into the correct format we have 9506 collision rates with the accompanying site number and month. We were also provided with an additional data set which contained the longitude and latitude co-ordinates to the centroid of each zone which we over laid onto a map of Florida to geographically see the center of each of the traffic zones.

In Figure 1.3 we see the map of Florida overlaid with the centroids of the 51 traffic zones. The spread of of the traffic zones covers both the east and west coast of Florida as well as many that are in the center. Given the large spread of sites across the state we might expect to see a spatial effect that will improve our model.

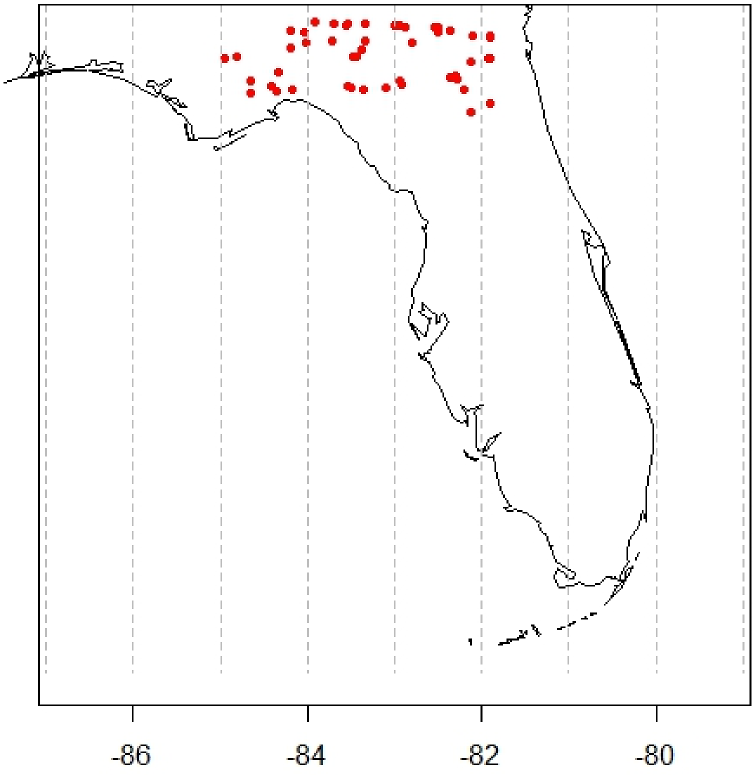


Figure 1.3: Centroid location of each traffic zone in Florida

Florida also has very specific hurricane seasons which affect the state between the months of June of October. Looking at the map again we see that some of the centroids of the zones are in land where as others are much more coastal, this adds to the evidence for expecting a spatial effect as we expect coastal zones to be affected more by hurricanes and cause more road collisions than in other in land zones. We might expect that this will cause a difference in road collisions between each month so we expect there to be a seasonal effect also. If we do some exploratory data analysis on the data we find some evidence to support our expectations of seasonal and spatial effects.

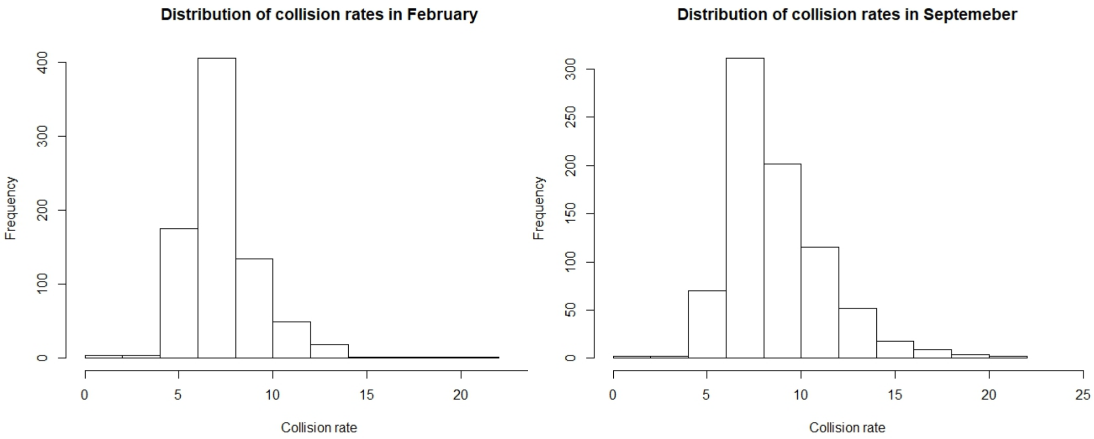


Figure 1.4: Comparison of collision rates in February and September

In Figure 1.4, we see the distribution of the recorded collision rates across all zones in the months of February and September, the importance of these two months is there relative positions to the hurricane season. Florida encounters very few hurricanes in February whereas September falls during peak hurricane season so we expect the collision rates in September to be higher than in February. We can see from the plots that the peak of the distributions is equal around 6-7 collisions per unit of distance, however the distribution for September is more right skewed and shows there are more collision rates of a higher value recorded in September than in February. Furthermore, the mean collision rate in September is 8.68 which is higher than the mean in February of 7.31.

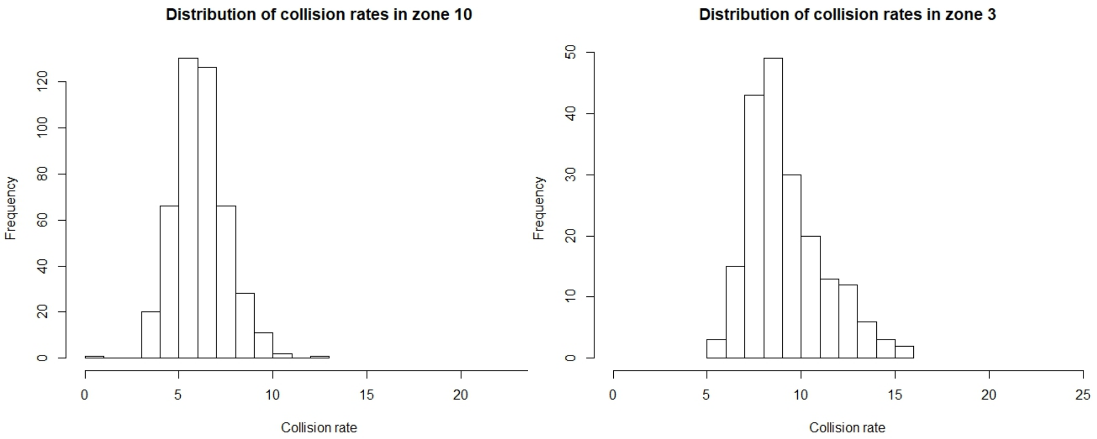


Figure 1.5: Comparison of collision rates in zone 3 and zone 10

In Figure 1.5, we see the distribution of recorded collision rates in zones 3 and 10. The significance of these two zones is their location and its relation to the hurricane season. Zone 3 is a coastal zone and because of this we expect it to be affected more by the hurricane season and have higher collision rates whereas zone 10 is an inland zone and therefore affected less by the hurricane season. The plots show clearly that the values for the collision rates in zone 3 is significantly higher than of zone 10 and whilst zone 10 distribution looks symmetric around its peak value, the distribution for zone 3 is right skewed meaning there are more collision rates recorded with higher values than the peak than there is with lower values of the peak. Finally, the mean collision of zone 3 is 9.21 which is significantly larger than the mean collision rate of zone 10 which is 6.13 which supports our expectations.

1.3 Bayesian Inference

In Bayesian inference we assume a parametric model to describe some phenomena, and represent our beliefs about those parameters through distributions. We chose a prior distribution $\boldsymbol{\pi(\theta)}$ to represent prior knowledge as a way of including external information or relative ignorance. This method allows for the inclusion of exterior information and an intuitive way of representing uncertainty that traditional statistical inference does not allow. Using Bayes Theorem we are able to include this external information in the form of a prior distribution with the observed data as follows

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto \pi(\boldsymbol{\theta}) \times f(\mathbf{x}|\boldsymbol{\theta})$$

This equation shows that the posterior distribution for a set of parameters $\boldsymbol{\theta}$ is proportional to the prior distributions of $\boldsymbol{\theta}$ multiplied by the likelihood of the data. If the prior distribution chosen is conjugate to the likelihood function then the posterior distribution will be of the same family of probability distribution as the prior and therefore we are able to define the posterior distribution. However, if the prior chosen is not conjugate to the likelihood function then the posterior distribution will take the form of a non-standard distribution and therefore we cannot find the posterior moments or plot the posterior density.

We can simulate realisations from the posterior distributions using Markov Chain Monte Carlo (MCMC) methods. In this project we used a mixture of Gibbs sampling and Metropolis Hastings sampling algorithms to generate realisations from the posterior distribution.[1]

Algorithm 1 Metropolis-Hastings Sampling

```
1. Initialise the iteration counter to j = 1
2. Initialise the state of the chain to $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_p^{(0)})^T$
3. Generate a proposed value $\boldsymbol{\theta}^*$ using the proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(j-1)})$
4. Evaluate the acceptance probability $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*)$ of the proposed move, where
       $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = min\left(1, \frac{\pi(\boldsymbol{\theta}^*|\mathbf{x})q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}|\mathbf{x})q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}\right)$
5. set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^*$ with probability $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*)$, set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$ otherwise
6. Change counter j to j+1, and return to step 3
```

Algorithm 2 Gibbs Sampling

```
1. Initialise the iteration counter to j = 1
2. Initialise the state of the chain to $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})^T$
3. Obtain a new value $\boldsymbol{\theta}^{(j)}$ from $\boldsymbol{\theta}^{(j-1)}$ by successive generation of values
       $\theta_1^{(j)} \sim \pi(\theta_1 | \theta_2^{(j-1)}, \theta_3^{(j-1)}, \ldots, \theta_p^{(j-1)}, \boldsymbol{x})$
       $\theta_2^{(j)} \sim \pi(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_p^{(j-1)}, \boldsymbol{x})$
       .
       .
       .
       $\theta_p^{(j)} \sim \pi(\theta_p | \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_{p-1}^{(j)}, \boldsymbol{x})$
4. Change counter j to j+1, and return to step 3
```

The MCMC scheme we ran to sample from the posterior distribution was a component-wise Metropolis-Hastings algorithm with normal random walk proposals. This algorithm runs very similar to the Gibbs sampler in structure but instead of drawing from a full conditional distribution for each parameter, we define a proposal distribution to draw a proposal from and calculate and acceptance probability using the posterior distribution at the proposed value and the posterior at the current value. We chose to draw the proposal from a normal distribution and since the normal distribution is symmetric it cancels out in the acceptable probability.

Chapter 2

Independent Fixed Models

2.1 The Null Model

The first model we decided to fit to the data follows a Normal distribution with mean μ and variance $\frac{1}{\tau}$. This null model is used as a baseline model to compare against the model built using seasonal and zonal effects where each observation in the full model, $y_{i,s}$, corresponds to a collision rate in month s and zone i. We assume vague prior knowledge for both mu and tau as linking back to the lower economic standing countries, we want the model to be as universal as possible to we choose the priors to be a wide as possible. so the priors are set as follows.

$$\begin{aligned} y &\sim N\left(\mu, \frac{1}{\tau}\right) \\ \mu &\sim N(0, 1000) \\ log(\tau) &\sim N(0, 1000) \end{aligned}$$

Algorithm 3 Component-wise Metropolis-Hastings Sampling

```
1. Initialise the iteration counter to j = 1
2. Initialise the state of the chain to $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})^T$
Let $\boldsymbol{\theta}_{-i}^{(j)} = (\theta_1^{(j)}, \ldots, \theta_{i-1}^{(j)}, \theta_{i+1}^{(j)}, \ldots, \theta_p^{(j)})^T, i = 1, 2, \ldots, p$
3. Obtain a new value $\boldsymbol{\theta}^{(j)}$ from $\boldsymbol{\theta}^{(j-1)}$ by successive generation of values
$\theta_1^{(j)} \sim \pi(\theta_1|\boldsymbol{\theta}_{-1}^{(j)}, \boldsymbol{x})$ using M-H step with proposal distribution $q_1(\theta_1|\theta_1^{(j-1)}, \boldsymbol{\theta}_{-1}^{(j)})$
$\theta_2^{(j)} \sim \pi(\theta_2|\boldsymbol{\theta}_{-2}^{(j)}, \boldsymbol{x})$ using M-H step with proposal distribution $q_2(\theta_2|\theta_2^{(j-1)}, \boldsymbol{\theta}_{-2}^{(j)})$
.
.
.
$\theta_p^{(j)} \sim \pi(\theta_p|\boldsymbol{\theta}_{-p}^{(j)}, \boldsymbol{x})$ using M-H step with proposal distribution $q_p(\theta_p|\theta_p^{(j-1)}, \boldsymbol{\theta}_{-p}^{(j)})$
4. Change counter j to j+1, and return to step 3
```

So as shown in Algorithm 3, for each parameter in our model, in this instance just μ and τ, we would update it using a Metropolis-Hastings step with a proposed value generated from a Normal distribution with a mean value as the current value of the parameter and the variance set as an innovation parameter which can be tuned such that the acceptance rates, the percentage of proposed values accepted, are in the needed range of 20% - 40%. We define a burn-in period as the time it takes for the Markov chain to reach its stationary distribution and so after removing the burn in period, we were looking to generate approximately 10,000 realisations of the posterior distributions of the parameters and from looking at the trace plot below of both μ and τ we can the that the Markov chains have converged to their stationary distributions as we are accepting values within a consistent range and therefore we have generated realisations from the posterior distribution.

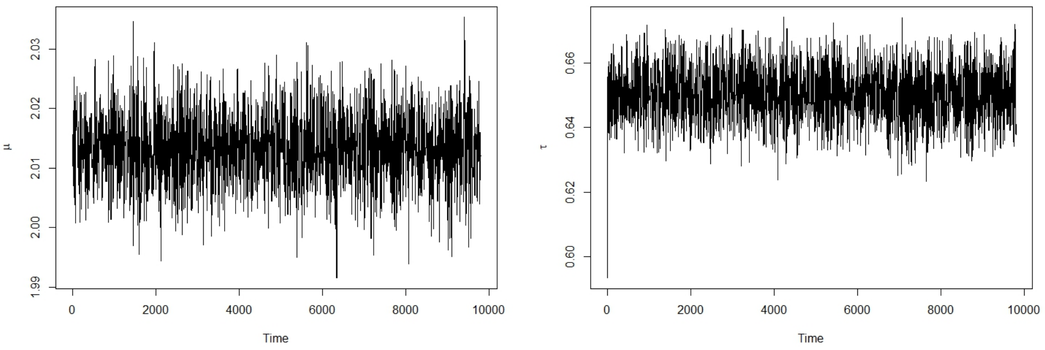
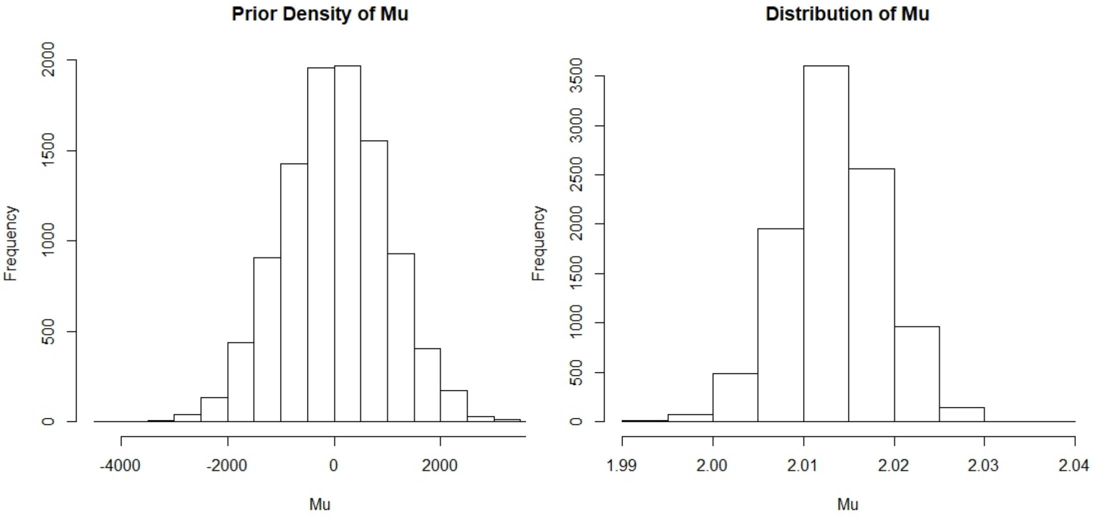


Figure 2.1: Trace plot of μ and τ

If we take a look at the posterior distribution of μ in comparison to the prior distribution we can see that the mean value for μ has shifted to the right by a value of 2 and still seems to follow a normal looking distributions with significantly smaller variance.



(a) Prior

(b) Posterior

Figure 2.2: Comparison of prior and posterior densities of μ

2.2 The Seasonal Model

For the seasonal model we are assuming there is difference in collision rates between each month so we model the data using a seasonal effect $\boldsymbol{\phi_s}$, for s = 1,2,...,12. As before with the null model we assume vague prior knowledge for the seasonal effects and so we use the following distributions to model the data.

$$y_{i,s} \sim N\left(\phi_s, \frac{1}{\tau}\right)$$
$$\phi_s \sim N(0, 1000)$$
$$log(\tau) \sim N(0, 1000)$$

We use the same component-wise metropolis-Hastings sampling algorithm as we used to sample from the posterior distributions in the null model except now $\boldsymbol{\theta}$ consists of 12 separate seasonal effects and $\boldsymbol{\tau}$. So we update each seasonal effect in turn by generating a proposal value from the proposal distribution with mean set to the current value of the parameter and calculating an individual acceptance probability for each parameter in $\boldsymbol{\theta}$. In turn we end up with generating realisations from 12 separate posterior distributions that show the distributions of collision rates during each month. To check whether there seems to be a clear seasonal effect we expect to see a difference in the posterior means for each of the seasonal effects.

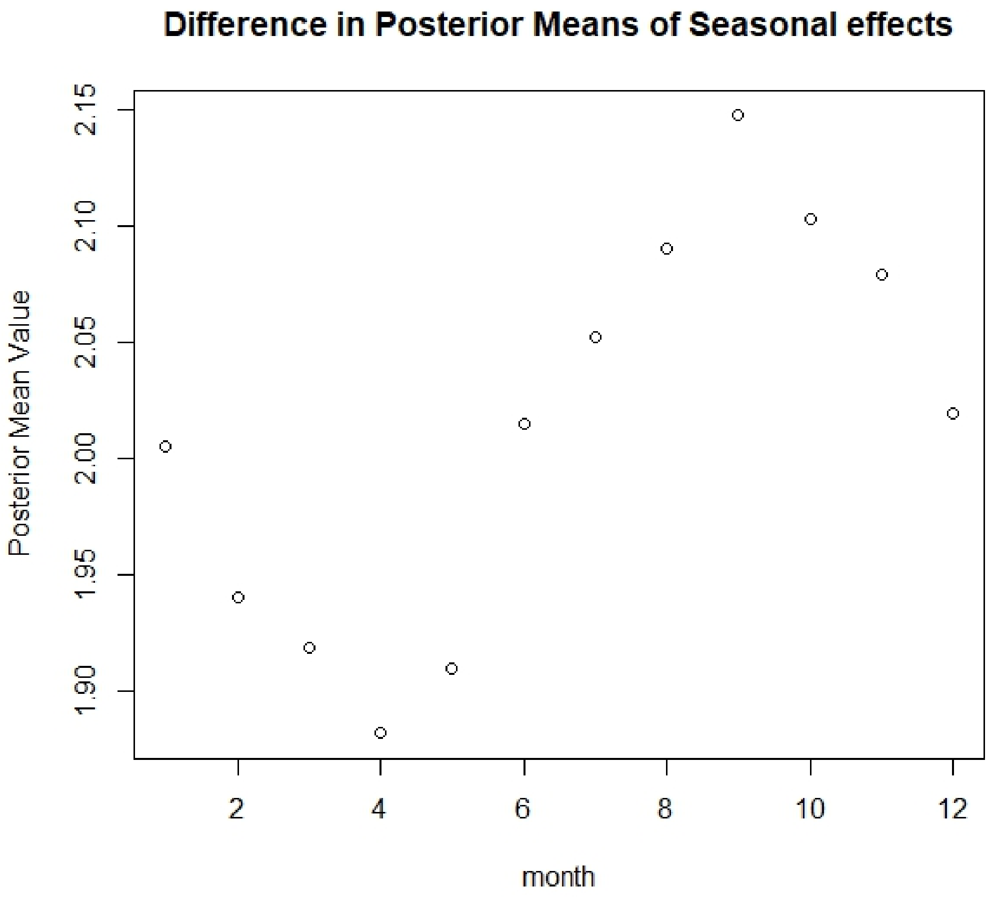
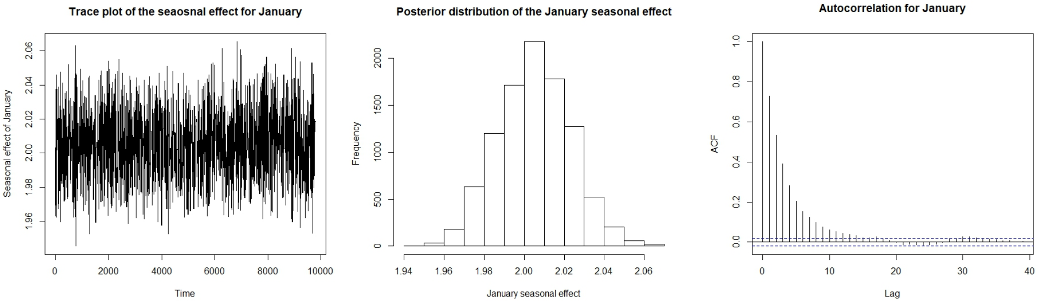


Figure 2.3: Difference between the seasonal effects posterior means

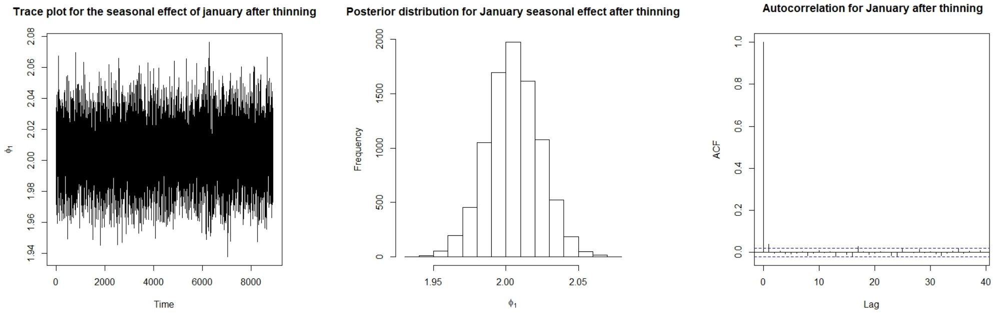
Looking at figure 2.3 we can see a clear trend in the posterior mean values for each of the seasonal effects. April has the lowest posterior mean values suggesting there are less collisions in April than in any other month. The posterior means then rise from April to reach a peak in September where we expect a higher collisions rate. From September we get a decreasing trend all the way back to April and so we can clearly see there is a difference between the seasonal effects and that they should not be left out from the model.

In Figure 2.4 we look in more detail at the seasonal effect for January. The first plot the effect shows a trace plot of the realisations generated using the MCMC scheme and we can see that the chain has converged and is exploring a consistent range of values. The plot shows that the mixing is good and the chain does not go for long periods without accepting a proposal and so we can say that our realisations are from the stationary distribution of this Markov chain. The second plot shown in the figure shows the posterior density of the seasonal effect and we can see that posterior density looks to have a normal like distribution with a mean value shown in the posterior means plot in figure 2.3. The final plot shows the autocorrelation for each of the seasonal effect and guides us to how much thinning is required to achieve an effective sample size of approximately 10000 after the burn in period was removed. The autocorrelation plot for the seasonal effect shown suggest we need to thin by a factor of 10- 12.



(a) Trace plot (b) Posterior density (c) Autocorrelation

Figure 2.4: January seasonal effect before thinning



(a) Trace plot (b) Posterior density (c) Autocorrelation

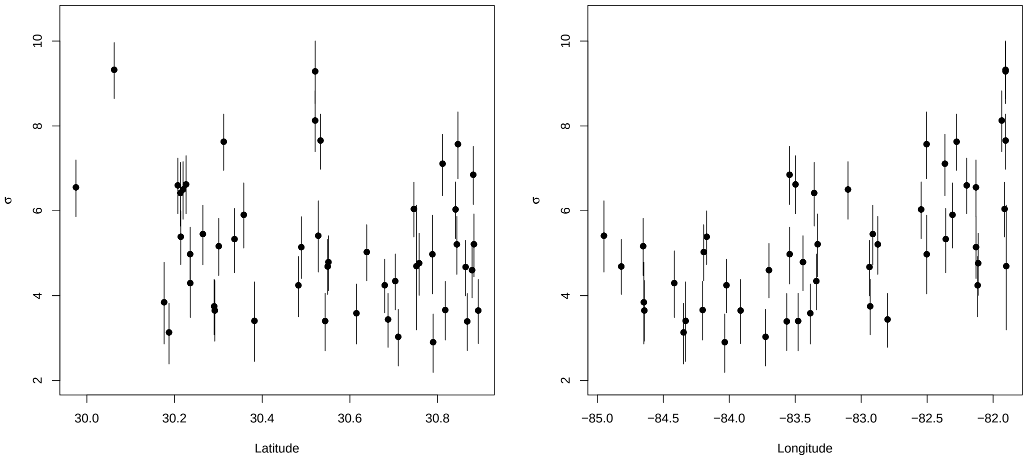
Figure 2.5: January seasonal effect after thinning

2.3 The Zonal Model

In the zonal model we are assuming there is a difference in collision rates between each of the zones and we therefore to chose to model the data with the inclusion of a spatial effect σ_i, for i=1,2,...,51. Like with the previous two models we are assuming vague prior knowledge and so we use the following distributions to define our model.

$$y_{i,s} \sim N(\sigma_i, \frac{1}{\tau})$$
$$\sigma_i \sim N(0, 1000)$$
$$log(\tau) \sim N(0, 1000)$$

We continue to use the component-wise metropolis hastings algorithm to generate realisations of the posterior distributions and for this model $\boldsymbol{\theta}$ consists of 51 separate spatial effects and $\boldsymbol{\tau}$ and we generate these realisations in the same way as the previous two models. Now we have 51 posterior distributions for each of the spatial effects we want to check if there seems to be a difference in the posterior means against longitude and latitude co-ordinates of the 51 zone centroids.



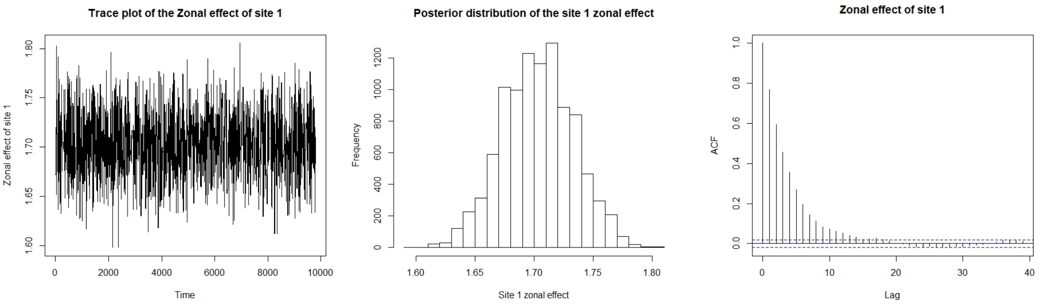
(a) Latitude

(b) Longitude

Figure 2.6: Spatial effect posterior means against the geographical co-ordinates of the zones

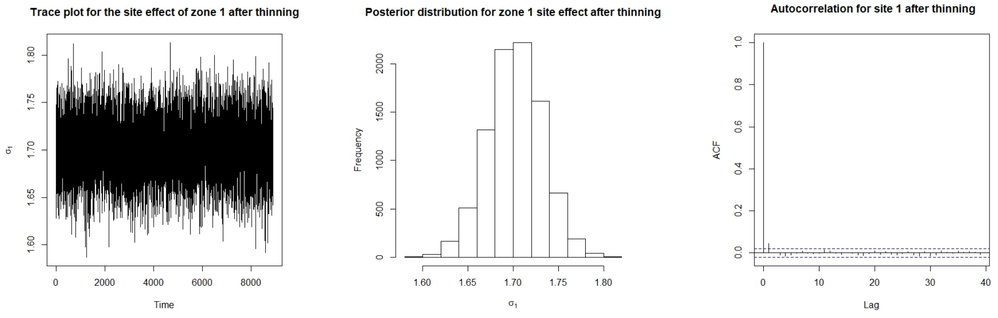
When looking at Figure 2.6 there does not seem to be a clear trend in posterior means against the latitude co-ordinates but there does seem to be a slight increasing trend in the plot against the longitude co-ordinates and so we do see a difference in the posterior means between spatial effects and thus they should not be left out of the model.

In Figure 2.7 we look specifically at the spatial effect for site 1. The first of the three plots is a trace plot of the realisations generated by the MCMC scheme. As with the seasonal model we can see from the plot that the chains have converged, the mixing is good and that the realisations are from the stationary distribution. In the second plot we see the posterior densities for the spatial effect and appear to have a normal like distribution with a mean value near 1.7. Finally in the last plot we see the autocorrelation is high and like in the seasonal model we would need to thin by a factor of 10-12 to lower the autocorrelation.



(a) Trace plot (b) Posterior density (c) Autocorrelation

Figure 2.7: Site 1 zonal effect before thinning



(a) Trace plot (b) Posterior density (c) Autocorrelation

Figure 2.8: Site 1 zonal effect after thinning

2.4 The Seasonal and Zonal Model

The fourth model built is a combination of the previous two models, we include the seasonal and zonal effects to create the full model which includes all the effects. For this model each observation $y_{i,s}$ is a collision rate in a specific zone and a specific month.We again assume vague prior knowledge and have set the following priors.

$$\begin{aligned} y_{i,s} &\sim N\left(\sigma_i + \phi_s, \frac{1}{\tau}\right) \\ \phi_s &\sim N(0, 1000) \\ \sigma_i &\sim N(0, 1000) \\ log(\tau) &\sim N(0, 1000) \end{aligned}$$

When running this model initially it became clear that there was an identifiability issue. To define identifiability, let $\boldsymbol{Y}$ be a vector of observed random variables and define $\boldsymbol{f}$ as the probability distribution function our model specified by the parameters $\boldsymbol{\theta}$. If there exists some $\theta_1 \neq \theta_2$ that satisfies

$$f(\boldsymbol{Y}|\theta_1) = f(\boldsymbol{Y}|\theta_2)$$

for all $\boldsymbol{Y}$, then the parameters are of the model are not identifiable. So this means that all possible sets of observations have identical probabilities for two different sets of parameters. To fix this issue, we set the first seasonal effect, the effect of January, to be equal to 0 and then not updated through the MCMC scheme. For all structured models based on this model, we must also set January's seasonal effect equal to 0.

In Figure 2.9 and 2.10 the output of our MCMC scheme is shown in trace plots for the seasonal effect in October and the zonal effect at site 1. In both trace plots for the effects we can see that the chain has converged, we are generating samples from the stationary distribution and that the mixing of the chain is good. The posterior distribution for both effects have a Normal like distribution but differ at their mean values. For the October seasonal effect, we see that it is comparatively small to the zone 1 effect as the posterior means are 0.086 and 1.70 respectively.

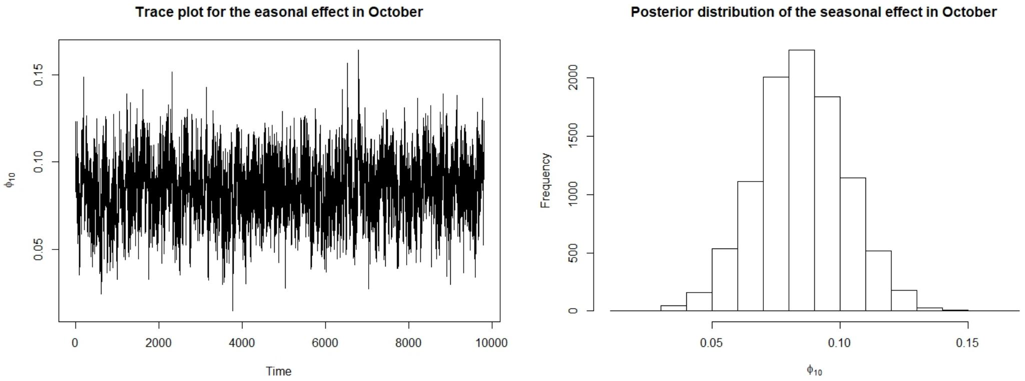


Figure 2.9: Trace plot and posterior distribution for the seasonal effect in October

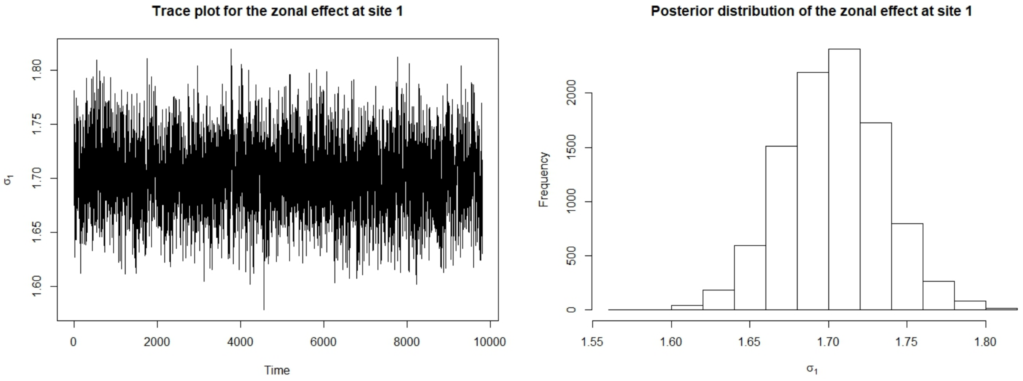
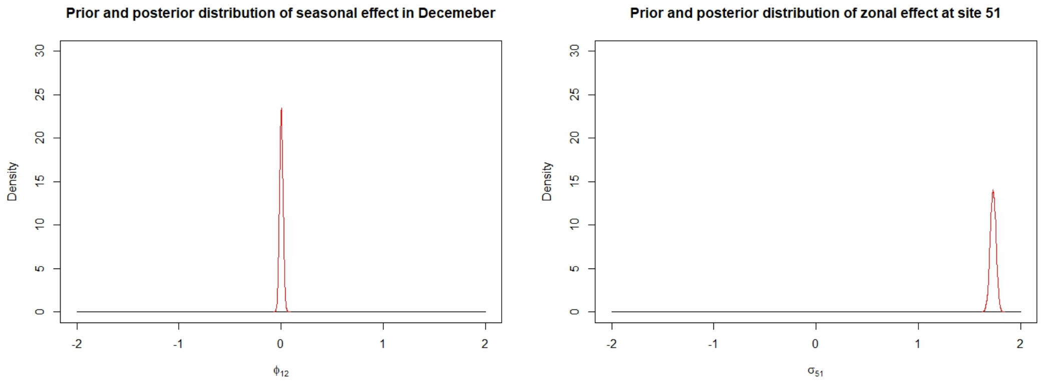


Figure 2.10: Trace plot and posterior distribution of zonal effect at site 1

In Figure 2.11, we show the a comparison of the prior and posterior distributions for the seasonal effect in December and the zonal effect at site 51. In both plots we see that the prior distribution is as wide as possible as we assumed vague prior knowledge so looks like a flat line just above zero density and both posterior distributions have significantly narrowed and shifted to their posterior mean values of 0.003 and 1.72 for the seasonal and zonal effect respectively.


Prior and posterior distribution of seasonal effect in Decemeber

Prior and posterior distribution of zonal effect at site 51

Figure 2.11: Prior in black and posterior in red distributions of seasonal effect in December and the zonal effect at site 51

In order to compare the uncertainty of this full model with the structured models we need to compute confidence intervals for a few of the observations in a specific months and zones. For an observations $y_{i,s}$ the following are 95% credible intervals for an a collision rate at zone i in month s.[?]

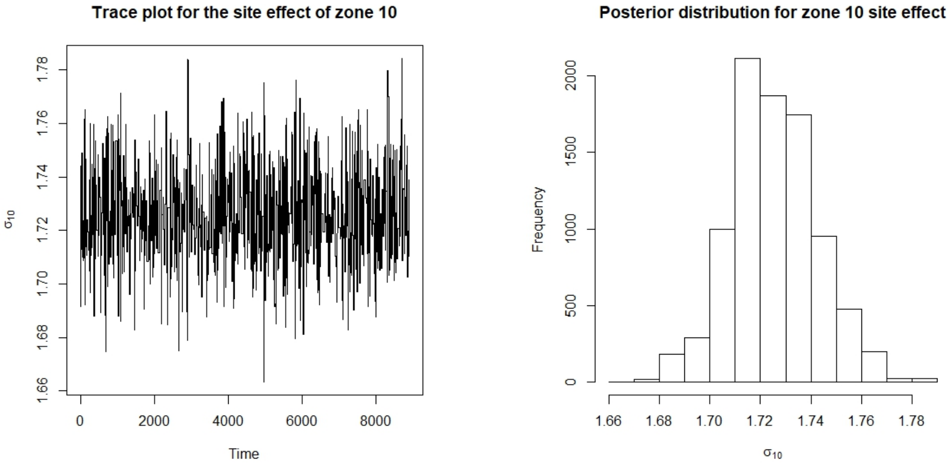
$$\begin{aligned} y_{1,2} &= (1.578, 1.702) \\ y_{10,4} &= (1.602, 1.680) \\ y_{25,7} &= (1.775, 1.883) \\ y_{48,10} &= (2.116, 2.191) \end{aligned}$$

2.5 The Summer Season Model

To test if all of the seasonal effects are required in the model with the zonal effect we built a summer effect model. This model now changes the seasonal effect in to a binary variable, in which the effect is equal to 1 if observations were recorded during the months of June to October, in line with the hurricane season. For this model we used the Metropolis-Hastings component-wise algorithm to sample from the posterior distributions of the effects and like with all other models previously we assume vague prior knowledge on both the summer effect, ϕ_{sum}, and all the zonal effects σ_i for i = 1,...51.

$$\begin{aligned} y_i &\sim N\left(\sigma_i + \phi_{sum}, \frac{1}{\tau}\right) \\ \phi_{sum} &\sim N(0, 1000) \\ \sigma_i &\sim N(0, 1000) \\ log(\tau) &\sim N(0, 1000) \end{aligned}$$

In Figure 2.12, it shows the trace plot and posterior distribution of zonal effect at site 10. The trace plot is exploring a consistent range of values showing that the Markov chain has converged to its stationary distribution and the output are samples from the summer effects posterior distribution. The histogram of the posterior distribution shows a Normal like distribution of values with a mean value of 1.72.

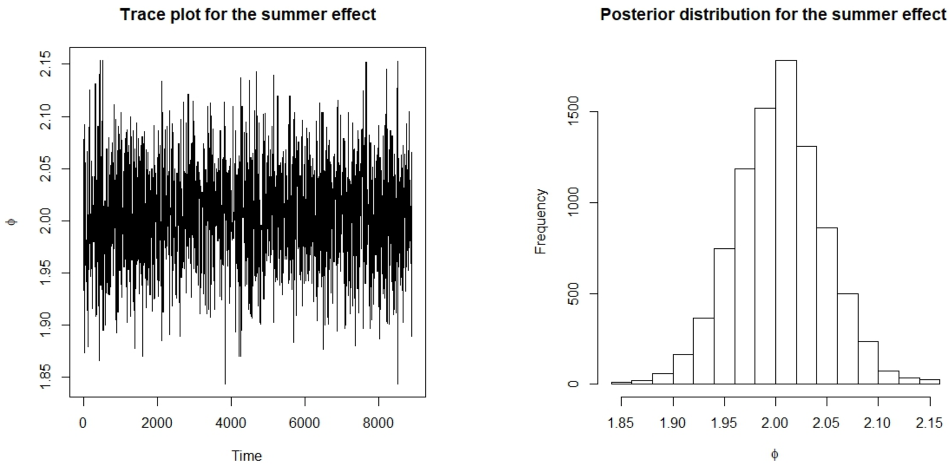


(a) Trace plot

(b) Posterior density

Figure 2.12: Site 10 zonal effect

In Figure 2.13, we show the trace plot and posterior density for the summer effect. As before the the chain has converged and the mixing is good so we know that the samples generated are from the posterior distribution. The posterior distribution has a mean values of 2.00 and has Normal like distribution of values which is slightly left skewed.



(a) Trace plot

(b) Posterior density

Figure 2.13: Summer effect

2.6 Deviance Information Criterion

In order to chose the best model we to introduce a way of comparing these models using a singular value. The deviance information criterion (DIC) is a Bayesian equivalent to the known Akaike Information Criterion (AIC) and like the AIC the DIC is a single value that balances a measure of model adequacy and a measure of model complexity. To calculate the DIC we must first define the value of the deviance. The deviance is defined by

$$D(\boldsymbol{\theta}) = -2log(f(\boldsymbol{y}|\boldsymbol{\theta}))$$

So the deviance is just the log-likelihood multiplied by -2. Whilst we don't use the deviance as a measure of fit for a model because as more parameters are added the deviance will continue to decrease which can cause issues such as overfitting and misleading conclusions, we do use the deviance to calculate the DIC for each of the models. The DIC penalises models for having too many parameters much like the AIC, so models that balance number of parameters and fit to the data get the lowest DIC values. The DIC is defined by

$$DIC = 2\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$$

To calculate the DIC we must first calculate deviance values for each iteration of the MCMC output and then find the mean of these deviance values. Next we calculate the posterior means of the parameters we generated realisations from and we use this values to calculate a separate deviance value. We now can calculate a singular DIC value for each of the models, lower values of DIC suggest a better model for this data. The DIC values for each model of the independent fixed models are defined below

$$\begin{aligned} Null &= 14615 \\ Seasonal &= 14402 \\ Zonal &= 7087 \\ Summer &= 7326 \\ Full &= 6585 \end{aligned}$$

As we can see from the DIC values, as the seasonal and zonal effects are added to the model, the DIC values decrease and the model becomes better suited for the data. We added the summer model after noticing that adding seasonal effects reduced the DIC value by a significantly smaller amount that adding zonal effects. So, we tested whether all of the seasonal effects were needed by fitting a model with zonal effects and an additional summer effect. However, the DIC value increased from the full model and therefore we cannot leave out the seasonal effects. [2][4]

Chapter 3

Adding structure to the full model

3.1 Motivation

We have created 5 individual models for the data and used the DIC to compare them in their effectiveness of modeling the data, the full seasonal and zonal model had the lowest DIC value and was therefore decided to be the best model. Using this model we were then able to model a collision rate for a specific zone in a specific month with 95% credible interval and linking back to the lack of accessibility for road collision data in lower economic standing countries, we need to try and reduce the width of these confidence intervals as much as possible to get the maximum amount of impact from the data we have. We aim to do this by adding structure to our full mode seasonal and zonal model to create 3 new models in which we allow for sharing of information between the seasonal and zonal effects. The model is the same as the full seasonal and zonal model where observations $y_{i,s}$ are collision rates in a specific zone and a specific month.

$$\begin{aligned} y_{i,s} &\sim N\left(\sigma_i + \phi_s, \frac{1}{\tau}\right) \\ \phi_s &\sim N(0, 1000) \\ \sigma_i &\sim N(0, 1000) \\ log(\tau) &\sim N(0, 1000) \end{aligned}$$

3.2 The CAR Model

The conditional autoregressive (CAR) models were introduced in 1974 but have been used more extensively in the past 20 years due to the convenience of employment in MCMC methods for fitting complex hierarchical spatial models. This model allows for sharing of information of the seasonal effects by generating a new proposal from the proposal distribution where the mean of the proposal distribution is a weighted sum of all the other current seasonal effects. We chose the weights according to how many of the seasonal effects we wish to effect the proposed value for the current seasonal effect.

$$q(\boldsymbol{\phi}^{*}|\boldsymbol{\phi}^{(j-1)}) \sim N(w_1\phi_1 + w_2\phi_2 + \ldots + w_{12}\phi_{12}, \epsilon)$$

Mathematically, for seasonal effect $\boldsymbol{\phi_s}$, we calculate a vector of weights $\boldsymbol{w} = (w_1, \ldots, w_{s-1}, w_{s+1}, \ldots, w_{12})$ where $\sum(\boldsymbol{w}) = 1$. In our specific models we decided to share information between the seasonal effects that landed previously and immediately after the current month and for the second model we chose to include the 2 seasonal effects that landed previously and after the current month.

The first model we built has a vector of weights $\boldsymbol{w} = (0, \ldots, w_{s-1} = 0.5, w_{s+1} = 0.5, \ldots, 0)$ so the effects either side of the current month have weights 0.5 and the rest have a weight of 0. For a model of monthly data it is unlikely that there will be long term dependence and so we don't expect the effect in March to tell us anything about effect in September. The second model has weights $\boldsymbol{w} = (0, \ldots, w_{s-2} = \frac{1}{6}, w_{s-1} = \frac{1}{3}, w_{s+1} = \frac{1}{3}, w_{s+2} = \frac{1}{6}, \ldots, 0)$ so the effects directly before and after the current month have a weight of $\frac{1}{3}$ and the effects that are two months before and after the current effect has weights $\frac{1}{6}$ so the closer the effect to the current effect, the more impact it will have on the proposed value of the effect.

In Figures 3.1 and 3.2, we again look at the same seasonal effect , $\boldsymbol{\phi_{10}}$, and zonal effect, $\boldsymbol{\sigma_1}$, as we did for the full seasonal and zonal model. The trace plots show that the chains have converged and that the posterior distributions looks almost identical to the unstructured model.

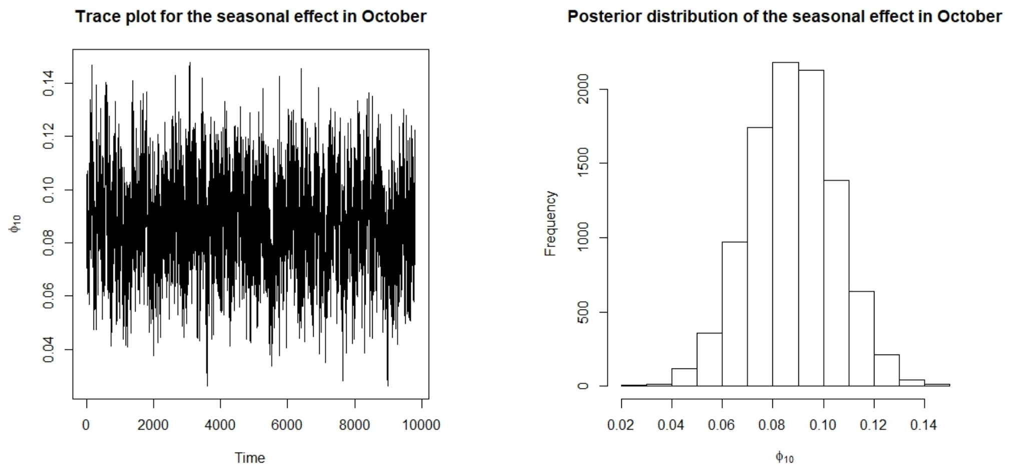


Figure 3.1: Trace plot and posterior distribution for the seasonal effect in October

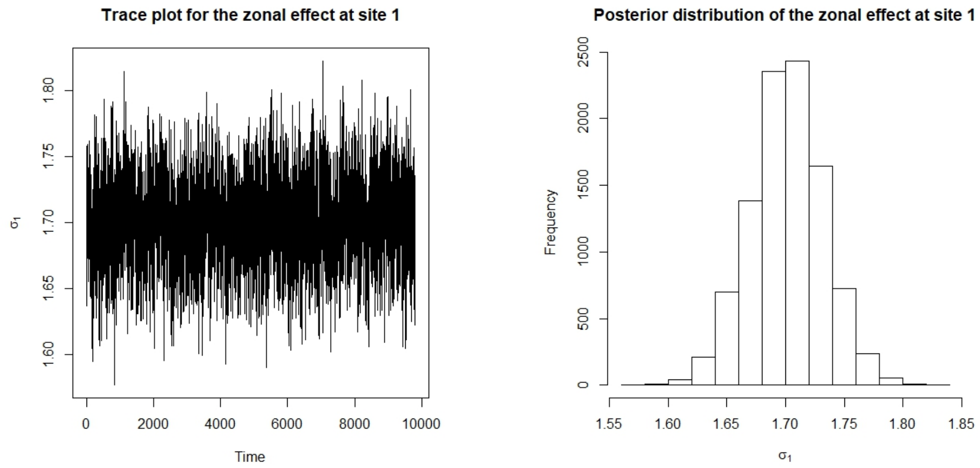


Figure 3.2: Trace plot and posterior distribution of zonal effect at site 1

In Figure 3.3 we show a prior and posterior distribution comparison for the same seasonal effect, ϕ_{12}, and zonal effect, σ_{51}, as we looked at in the full seasonal and zonal model. Again we see that the posterior distributions looks identical except for a slight increase in the density of their peaks showing the distributions have narrowed.

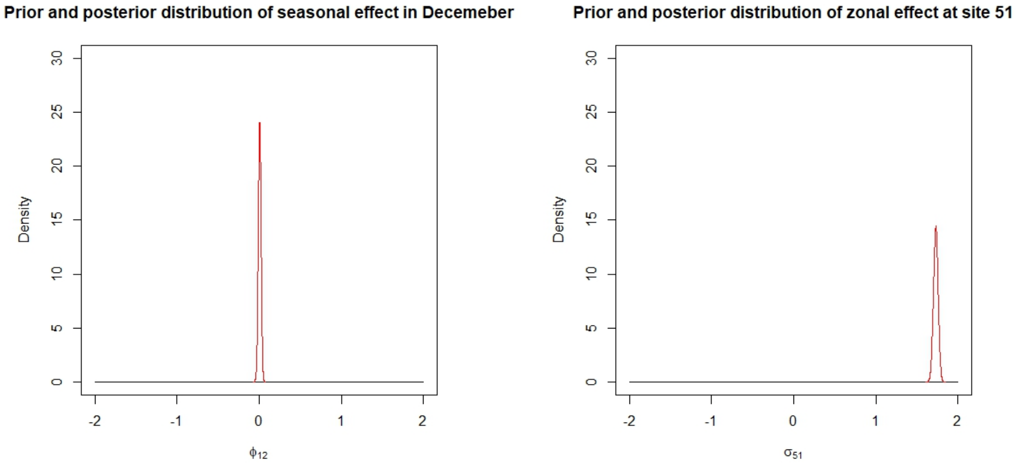


Figure 3.3: Prior and posterior distributions of seasonal effect in December and the zonal effect at site 51

So we check the same 95% credible intervals as before to compare. For an observations $y_{i,s}$ the following are 95% credible intervals for an a collision rate at zone i in month s.

$$\begin{aligned} y_{1,2} &= (1.579, 1.702) \\ y_{10,4} &= (1.604, 1.681) \\ y_{25,7} &= (1.776, 1882) \\ y_{48,10} &= (2.117, 2.190) \end{aligned}$$

So this model has narrowed these confidence intervals by an average of 0.002 and has therefore reduced the uncertainty from our previous model. We can also calculate a DIC value for the CAR model and we find that it has a value of 6566, which is lower than the previous model by 19, therefore suggesting this is a better model.

We had also produced plots from the second CAR model in which we expanded the influence of seasonal effects to two months before and after the current effect with weights $\frac{1}{3}, \frac{1}{6}$ respectively. The posterior distributions look similar to the previous CAR model except for the seasonal effect, ϕ_{10}, has slightly lower values.

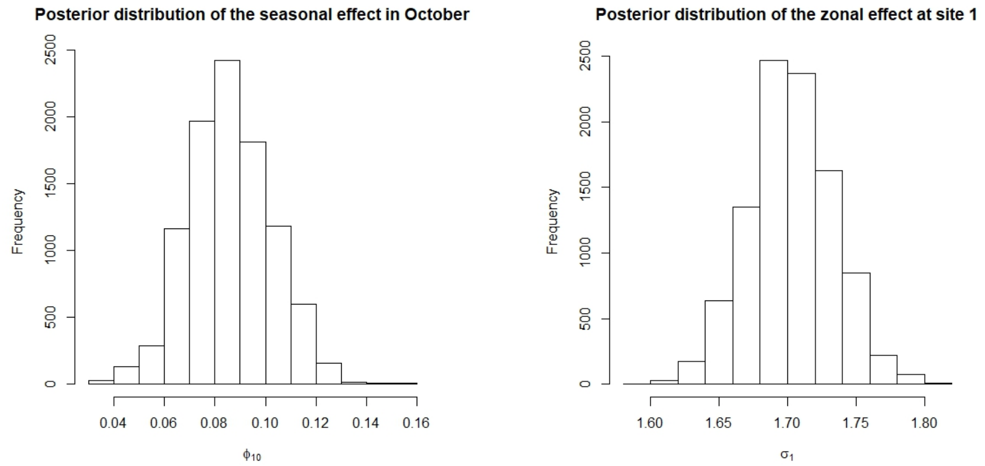


Figure 3.4: Posterior distribution for the seasonal effect in October and zonal effect at site 1

If we look at the 95% confidence intervals for the same observations we did in the first CAR model, these intervals are actually 0.001 wider and thus has larger uncertainty. The DIC value for the second CAR model is also larger than the first at a value of 6571, therefore suggesting the first CAR model is preferred to this model. This may be because we are oversmoothing when we allow effects to share information that are 2 months apart. [7][3]

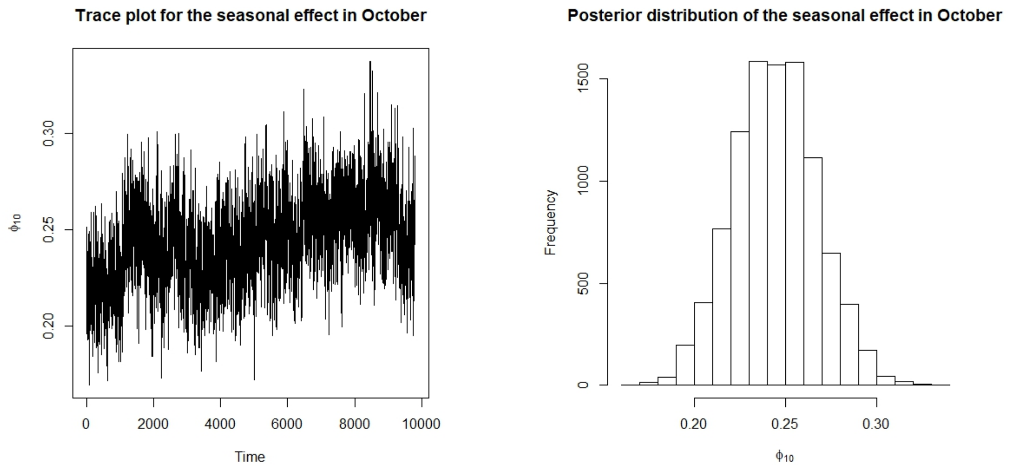
3.3 The KDE Model

The kernel density estimator (KDE) model works much like the CAR model, by sharing information between the effects to generate a new proposal value, but for the KDE model we are sharing information between the zonal effects instead of the seasonal effects. In the same way as the CAR model, we set the the mean of the proposal distribution for the current effect as the weighted sum of all of the effects. The way in which we define the weights is no longer done by choice, but by a Gaussian density kernel, so $w_{ij} = exp(\frac{-d_{ij}}{b})$ where d_{ij} is the distance between the the centroids of the zones i and j and b is the bandwidth parameter which defines how the weights are distributed to zones closer to the current zone in comparisons to zones that are further away.

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(j-1)}) = N(w_1\sigma_1 + w_2\sigma_2 + ... + w_{51}\sigma_{51}, \epsilon)$$

The aim of this structure is to have the zonal effects that are closer to the current zone have a greater impact on the proposal value for the current effect rather than having all the zones have an equal affect. So we chose the value of our bandwidth parameter to be $b = 0.1$ so that zonal effects that are closer to the the current zone will have a larger weight. For each zonal effect we propose we will first need to calculate a new set of weights, then generated a proposal from the proposal distribution and finally evaluate the acceptance probability.

In Figures 3.5 and 3.6 we see the trace plot and posterior distribution of the seasonal and zonal effects ϕ_{10}, σ_1. After a large burn-in period of 200,000 realisations the chain seem to have converged so the samples we have generated are from the posterior distribution. The posterior distributions, whilst having the same Normal distribution shape, have changed to all the other others. The distribution of ϕ_{10} has shifted to the right to a mean value of 0.25 and the distribution of σ_1 has also shift to the right by approximately 0.1 to a new mean value of 1.80.


Trace plot for the seasonal effect in October

Posterior distribution of the seasonal effect in October

Figure 3.5: Trace plot and posterior distribution for the seasonal effect in October

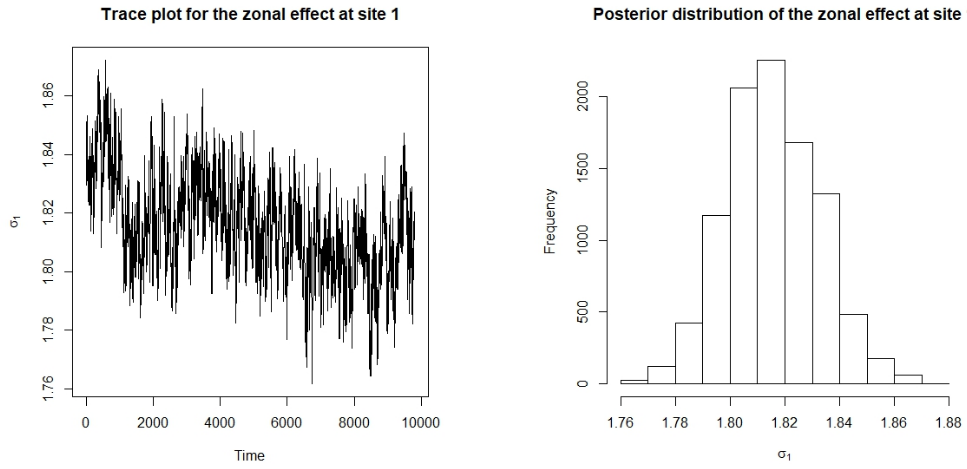


Figure 3.6: Trace plot and posterior distribution of zonal effect at site 1

In Figure 3.7 we compare the prior and posterior distributions of the seasonal and zonal effect ϕ_{12}, σ_{51}. As with the posterior distributions in Figures 3.5 and 3.6, these distributions have shifted to the right from previous models but we can still see the huge decrease in variance in the posterior distributions in comparison to the priors.

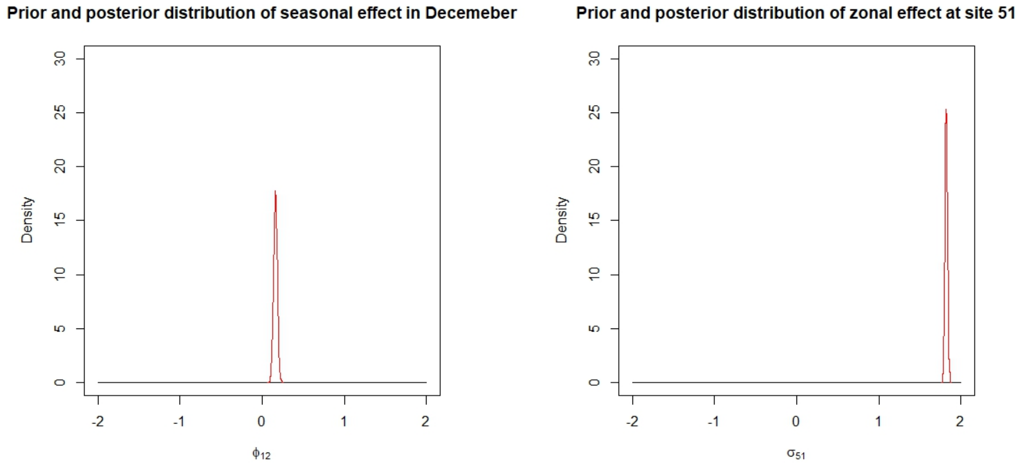


Figure 3.7: Prior and posterior distributions of seasonal effect in December and the zonal effect at site 51

In order to compare the uncertainty in this model to other models we calculated the 95% credible intervals for the same observations as we did previously where an observations $y_{i,s}$ the following are 95% credible intervals for an a collision rate at zone i in month s.

$$\begin{aligned} y_{1,2} &= (1.859, 1.938) \\ y_{10,4} &= (1.830, 1.910) \\ y_{25,7} &= (1.977, 2.054) \\ y_{48,10} &= (2.035, 2.118) \end{aligned}$$

The 95% confidence intervals all shifted to higher values than on previous models, however the width of the new credible intervals are narrower than the unstructured models but still wider than the first CAR model. So whilst this has less uncertainty than the previous models, it also predicts a different range of collision rates.[8]

3.4 The Hierarchical Model

The hierarchical model allows for sharing of information by allowing other effects inform the prior of the current effect. The model is a 3 stage hierarchical model where the priors and hyper priors are set as follows.

$$\begin{aligned} y_{i,s} &\sim N\left(\sigma_i + \phi_s, \frac{1}{\tau}\right) \\ log(\tau) &\sim N(0, 1000) \\ \phi_s &\sim N\left(0, \frac{1}{\tau_\phi}\right) \\ \sigma_i &\sim N\left(\mu_\sigma, \frac{1}{\tau_\sigma}\right) \\ \mu_\sigma &\sim N(a, b) \\ \tau_\sigma &\sim Ga(c, d) \\ \tau_\phi &\sim Ga(e, f) \end{aligned}$$

To generate samples from the posterior distributions from this model, we combine both the Metropolis-Hastings component-wise sampling with the Gibbs sampling algorithms. We updated the hyper parameters $\mu_\sigma, \tau_\sigma, \tau_\phi$ using the Gibbs sampling update of drawing a value from their full conditional distributions and we then used these new values in the Metropolis-Hastings component-wise algorithm to generate realisations for our seasonal and zonal effects ϕ_s, σ_i.

Firstly we had to calculate the full conditional distributions of the hyper parameters. The full conditional posterior distribution for μ_σ

$$\pi(\mu_\sigma|.) \propto \pi(\mu_\sigma)\pi(\sigma_i)$$

$$\propto \Pi_{i=1}^{51} exp\left(-\frac{\tau_\sigma}{2}(\sigma_i - \mu_\sigma)^2\right) exp\left(-\frac{1}{2}\frac{(\mu_\sigma - a)^2}{b}\right)$$

$$\propto exp\left(-\frac{1}{2}\left(\mu_\sigma^2\left(51\tau_\sigma + \frac{1}{b}\right) - 2\mu_\sigma\left(\frac{a}{b} + 51\tau_\sigma \sum \sigma_i\right)\right)\right)$$

$$\propto exp\left(-\frac{51\tau_\sigma + \frac{1}{b}}{2}\left(\mu_\sigma - \frac{\frac{a}{b} + 51\tau_\sigma \sum \sigma_i}{51\tau_\sigma + \frac{1}{b}}\right)^2\right)$$

$$\pi(\mu_\sigma|.) \sim N\left(\frac{\frac{a}{b} + 51\tau_\sigma \sum \sigma_i}{51\tau_\sigma + \frac{1}{b}}, \frac{1}{51\tau_\sigma + \frac{1}{b}}\right)$$

Likewise for both τ_σ, τ_ϕ

$$\pi(\tau_\sigma|.) \propto \pi(\tau_\sigma)\pi(\sigma_i)$$

$$\propto \tau_\sigma^{\frac{49}{2}+c} exp\left(-\tau_\sigma\left(d + \sum_{i=1}^{51}(\sigma_i + \mu_\sigma)^2\right)\right)$$

$$\pi(\tau_\sigma|.) \sim Ga\left(\frac{51}{2} + c, d + \sum_{i=1}^{51}(\sigma_i + \mu_\sigma)^2\right)$$

$$\pi(\tau_\phi|.) \sim Ga\left(e + 6, f + \sum_{s=1}^{12}\phi_s^2\right)$$

After computing the full conditional distributions for the hyper parameters we are able to generate realisations from the posterior distributions of the seasonal and zonal effects. In Figures 3.8 and 3.9 we show the trace plots and posterior distributions of the seasonal and zonal effects ϕ_{10}, σ_1. The trace plots show that the chain has converged and so we know that the samples are from the posterior distributions. The posterior distributions are similar to those from the previous models with the width of the distribution narrower than the unstructured full seasonal and zonal model.

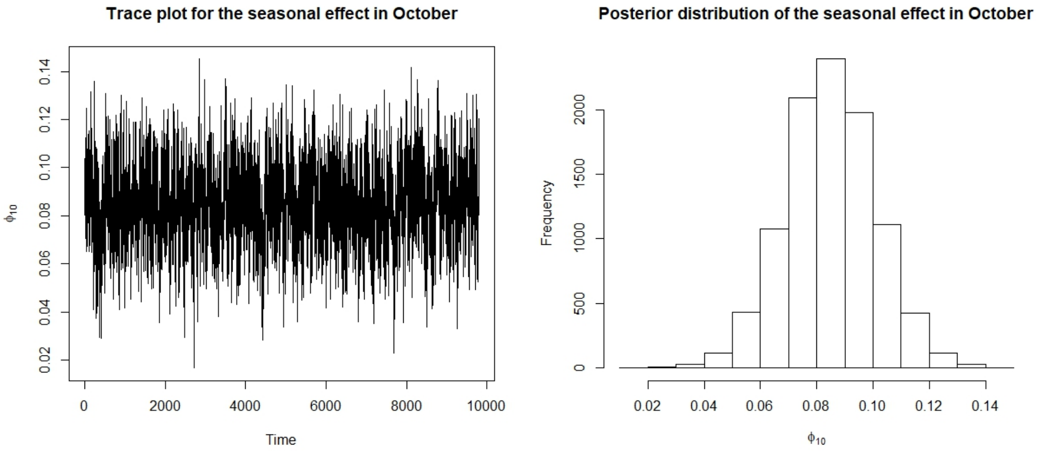


Figure 3.8: Trace plot and posterior distribution for the seasonal effect in October

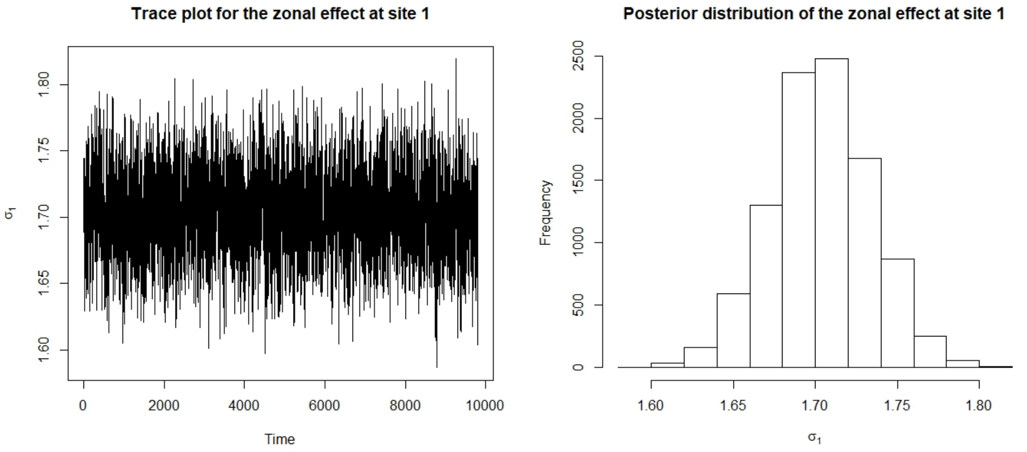


Figure 3.9: Trace plot and posterior distribution of zonal effect at site 1

As with the other models, we need to compute 95% credible intervals to compare the levels of uncertainty in this model. For an observations $y_{i,s}$ the following are 95% confidence intervals for an a collision rate at zone i in month s.

$$
\begin{aligned}
y_{1,2} &= (1.580, 1.702) \\
y_{10,4} &= (1.605, 1.679) \\
y_{25,7} &= (1.775, 1.882) \\
y_{48,10} &= (2.117, 2.190)
\end{aligned}
$$

This model has 95% credible intervals that are on average 0.00175 narrower than the unstructured full seasonal and zonal model, so this model has lower uncertainty. We can also calculate the DIC for this model which has a value of 6574 which is lower than the unstructured model but larger than the CAR model and therefore is not preferred.[9]

Chapter 4

Conclusions

4.1 The Best Model

We started this project with the aims creating a model that accurately represented road collisions in Florida and have use of modeling road collision in a more general setting. Our initial research on this topic led to the the studies from the World Health Organisation that show that countries with lower economic standing are affected most by road collisions, combined with the fact the road collisions cost countries up to 3% of their gross domestic product, it shows an extremely damaging cycle in developing countries.

So, with our aims set we built 5 models for the data and compared them using the deviance information criterion (DIC), with lower values of DIC suggesting a good balance between number of parameters and fit to the data. We decided on the data we have, that we should incorporate a seasonal effect and a spatial effect as Florida has a unique climate and endures at times strong hurricane seasons between June and October, with the fact that our data was collected form zones from across the state, it was our expectation that the season and location in Florida would effect the collision rate.

We found from our models that as the seasonal effects and the zonal effects were added, our DIC value decreased, suggesting that the added decrease in deviance in our output out weighted the penalising effect of adding extra parameters. Our full seasonal and zonal model that contain all 12 seasonal effects and 51 spatial effects had the lowest DIC value and was therefore decided to be our best model. We are then able to find values for what we expect a collision rate to be at a specific zone in a specific month, for example our results showed that we expected a collision rate of 1.670 at site 1 in February.

We wanted to go further in our research of modeling our data and ways of extracting as much usefulness as possible. Developing countries hit the hardest by road collisions also have little data available to them, so we needed to make the most out of the data we had and to also ensure that we reduced our uncertainty as much as possible. The advantage of using Bayesian modeling is that the uncertainty of a model is shown in a simple and intuitive way as a posterior distribution of samples, so using these we are able to create 95% confidence intervals as a way to visualize our uncertainty. To narrow these confidence intervals and lower our uncertainty we would have to add structure to our models.

4.2 Structured Models

The structured models we built all allowed sharing of information between parameters, we had 3 separate models in which we shared information between the seasonal effects in the CAR model, the spatial effects in the KDE model and in the hierarchical model we set the prior parameters to have a hyperprior distribution and thus allowed sharing of information between the parameters.

In the CAR model, we set the mean of the proposal distribution equal to the weighted sum of seasonal effects apart from the current effect. We built two of these models in which we let the weight for the first model be $\frac{1}{2}$ for the seasonal effects either side of the current effect and 0 otherwise and another where the weights are $\frac{1}{3}$ and $\frac{1}{6}$ for the seasonal effects either side of the current effect and the seasonal effects that are two months apart from the current effect respectively. The output from the MCMC scheme gave us posterior distributions that are narrower than the unstructured model and we calculated that the 95% confidence intervals for the first CAR model was on average 0.002 narrower and therefore has reduced the uncertainty. We also calculated that the first CAR model had a DIC value of 6566 which is lower than all previous models and is therefore the best model.

In the KDE model, we set the mean of the proposal distribution to be the sum of the the weighted site effects where the weights for each current effect are calculated using a Gaussian kernel that gives larger weights to sites closer to the current zonal effect,$w_{ij} = exp(\frac{-d_{ij}}{b})$ where d_{ij} is the distance between the the centroids of the zonal effects i and j and b is the bandwidth parameter which we set to 0.1. The MCMC output showed us that the 95% confidence intervals calculated for this model were on average 0.00165 narrower than the unstructured model but still wider than the first CAR model and therefore would not be preferred.

In the Hierarchical model, we set the prior distributions values to be hyper parameters with hyper priors to allow the sharing of information. We used a Gibbs update to calculate new values of the hyper parameters so we first had to calculate the full conditional distributions for the hyper parameters. Once the new values of the hyper parameters were calculated, we were able to use the same Metropolis-Hastings component-wise sampling algorithm as the previous models. The output of the MCMC scheme showed us that the 95% confidence intervals that were calculated had been narrowed by an average of 0.00175 showing that the uncertainty had be reduced in this model but less so than the first CAR model. The DIC value for this model was equal to 6574 which is lower than the unstructured model but higher than the first CAR model and is therefore not the best model.

Bibliography

[1] G.E.P. Box and G.C. Tiao. *Bayesian Inference in Statistical Analysis.* Wiley Classics Library. Wiley, 2011.

[2] G. Celeux, F. Forbes, C. P. Robert, and D. M. Titterington. Deviance information criteria for missing data models. *Bayesian Anal.*, 1(4):651–673, 12 2006.

[3] Victor De Oliveira. Bayesian analysis of conditional autoregressive models. *Annals of the Institute of Statistical Mathematics*, 64(1):107–133, 2012.

[4] Yong Li, Tao Zeng, and Jun Yu. Robust deviance information criterion for latent variable models. *CAFE research paper*, (13.19), 2013.

[5] World Health Organisation. The top 10 causes of death, 2018. Last accessed 24 April 2020.

[6] World Health Organisation. Road traffic injuries, 2020. Last accessed 24 April 2020.

[7] Cuirong Ren and Dongchu Sun. Objective bayesian analysis for car models. *Annals of the Institute of Statistical Mathematics*, 65(3):457–472, 2013.

[8] Mike West. *Bayesian kernel density estimation.* Institute of Statistics and Decision Sciences, Duke University, 1990.

[9] Mike West and Michael D Escobar. *Hierarchical priors and mixture models, with application in regression and density estimation.* Institute of Statistics and Decision Sciences, Duke University, 1993.