Statistical Methodology for the Analysis of Road Safety Data

Joseph T. Matthews

A thesis submitted to Newcastle University
for the degree of Doctor of Philosophy

Submitted: October 2019

Acknowledgements

I would first like to offer huge thanks to my fantastic supervisors Dr Neil Thorpe and Dr Lee Fawcett for providing me with the opportunity to pursue this PhD, their immense guidance on how to be an academic, and their companionship on the many travels I've been so fortunate to go on as part of my research.

Massive thanks must go to the industrial partners for their support. In particular Paulo Humanes and Karsten Kremer from PTV Group for their generous funding and giving of their time to provide valuable insights into the applications and future development of the project. Also thanks to Peter Slater and the Northumbria Safer Roads Initiative, without whose initial interest, this project as a whole would likely have never come to be.

Huge thanks must go to the fantastic Newcastle University staff I have been so lucky work alongside, Andy Golightly, Kevin Wilson, Daniel Henderson, Jackie Martin, and Georgina Kay-Black. A special mention must be made for the late Professor Richard Boys, without whose intervention my academic career would literally not be possible, and without whose constant demands for trips to the pub, this work would likely have been finished much sooner. In a similar vein, thanks to Stuart Douglas and the staff of the Hotspur pub for ensuring I was never in danger of dying of thirst of a Friday evening.

A special mention for the wonderful PhD cohort I was so lucky to be a part of: Aamir Khan, Alex Svalova, Ryan Doran, Robbie Bickerton, Jack Aiston, Naomi Hannaford, Liam Dobson, Jack Walton, Clarissa Barratt, Keith Newman, George Stagg, Hayley Moore, Jack Aiston and Nicola Hewett. Particular thanks must go to my incredible PHD3 colleagues: David Pescod, Tom Lowe, Matthew Robinson and Stephen Johnson. Particular mention goes to Tom Bland, whose often nocturnal work schedule provided invaluable company, and whose excellent taste in food outlets enriched my PhD experience beyond measure. Special mention also for David Robertson, whose encyclopaedic knowledge of everything related to maths and computing saved my life on countless occasions, whose commitment to ridiculous impressions and dancing to 80s music ensured life never got too serious, and whose barrage of compliments has inflated my confidence in my taste in socks to frankly untenable levels. A particular mention must also go to my fellow Brains Trust members, David Cushing, Em Rickinson, David Walshaw, Stuart Hall, David Stewart, Jen Wood and Sarah Jowett. A huge thank you to Sierra Mon, who despite being more than 3,000 miles away for the majority of my study, could always be relied on for camaraderie, humour, and a unique ability to make any situation feel better. Finally of course I have to thank my family, without whose continuing love, support, kindness and patience, this person let alone this work would not be possible. You can now finally stop wondering when I'm going to finish.

Thanks guys.

Contents

Chapter 1

Introduction

Notation

Below is a summary of the notation used to describe the statistical framework used in this chapter. This notation is retained throughout all future chapters.

Notation	Meaning
i	Site indicator, $i = 1, \ldots, N$
j	Covariate indicator, $j = 1, \ldots, P$
P	Number of covariates in the SPF
$x_{i,j}$	The value of covariate j at site i
$y_{i,\mathrm{BEF}}/y_{i,\mathrm{AFT}}$	Observed collision counts in the before/after period at site i
λ_i	Underlying collision rate at site i
ρ_i	RTM effect at site i
κ_i	Trend effect at site i
τ_i	Treatment effect at site i
μ_i	Fitted estimate from the SPF for site i
β_j	SPF regression coefficient for covariate j
ξ_i	Estimated multiplicative change in collision counts due to trend at site i
γ	Overdispersion parameter of the Negative Binomial error structure of the SPF
m/M	The iteration count/total number of iterations in an MCMC algorithm

1.1 Motivation

Ensuring and maintaining road safety is an important task facing government bodies the world over. It is estimated that approximately 1.35 million people die every year (one every 24 seconds), making road traffic injuries the 8th largest cause of death globally [World Health Organisation, 2018]. The dangers of road collisions are particularly noticeable when considering young people, which a 2018 World Health Organisation (WHO) report claimed was responsible for more deaths of children aged 5-14 and young adults aged 15-29 than any other single cause. *Road Safety by Sweden* take this claim further, claiming that without intervention, road traffic injuries will rise to be the 5th largest cause of death globally by 2030, in line with increased road vehicle ownership. In order to counteract this, in 1994 the organisation created the "Vision Zero Initiative" which has now become part of Swedish law, with the ultimate target being no deaths or serious injuries as a result of road accidents whatsoever [Traffic Safety By Sweden,]. In addition, the United Nations has declared the decade spanning 2011-2020 to be a "Decade of Action on Road Safety" [United Nations,], with the goal being to halve the annual number of global road accident deaths by the end of 2020, where it is hoped around 5 million lives may be saved as a result of this endeavour [FIA Foundation,]. It is particularly striking that while high-income countries are home to 40% of the world's vehicles, they have only 7% of the world's traffic related deaths. This is in contrast to low-income countries which house just 1% of the world's vehicles yet have 13% of the world's traffic related deaths [World Health Organisation, 2018]. This not only demonstrates the role of well funded infrastructure in maintaining road safety, it also highlights the importance of improving road safety in poorer areas. To that extent therefore it is important that for techniques to have the greatest potential to improve road safety globally, they should not depend on nor assume the large amounts of data/expertise which are only available in higher income countries.

Clearly there are many arms to this strategy implemented in a variety of ways, however a key contributor to the overall aims of this project are the safety practitioners working for road safety organisations, who are tasked with, among other things, deploying and ensuring the effectiveness of road safety countermeasures. These countermeasures, or simply "treatments" are usually deployed at set locations, and come in a variety of forms from schemes designed to alter road users behaviour, such as speed cameras; traffic calming measures; or behaviour awareness campaigns, to measures designed to improve the condition of the road segment itself; such as improving drainage in cases of bad weather; removing any obstructions to driver visibility (e.g. from overgrown plants); to adding traffic signals at busy intersections. The methods discussed henceforth do not discriminate between such treatments and can be applied to all such schemes equally (however in Chapter 2 we shall assume the same scheme has been implemented at multiple sites across a network), and so we shall focus more on the statistical analysis involved in the analysis of the road safety treatments, rather than the treatments themselves (however the area of treatment selection shall be looked at briefly in Sections 6.5 and A.1.4).

Whilst the amount of public money budgeted for the management of road safety appears substantial, for example £175 million was included in the 2017 U.K. Autumn Statement for the sole purpose of improving safety on the most dangerous roads in England [Department for Transport, 2016], the costs of most road safety schemes is also significant, for example it is estimated permanent average speed cameras cost £100,000 per mile [RAC Foundation, 2016].

This represents a substantial investment of public funding, and as such it is important to ensure these funds are invested in an appropriate way. As will be discussed in this thesis, the majority of decisions made in relation to funding allocation through things like countermeasure selection and allocation are primarily data and analysis driven. As such it is of paramount importance to ensure that techniques used to analyse road safety data are as effective and efficient as possible.

Naturally, the optimal allocation of countermeasure treatments would be that which causes the greatest reduction in danger, where here the level of danger shall be assumed to be represented by the collision/casualty rate directly due to road collisions at a location, for a given amount of investment. In this regard the collision/casualty rate should not be viewed in isolation as a determinant of candidacy for treatment allocation. Clearly locations which would be expected to have a greater number of collisions by their very nature (e.g. locations where there is a high amount of traffic flow) cannot necessarily be judged to be a better candidate for treatment than a location which would be expected to have a low number of collisions (e.g. because it has a particularly low amount of traffic flow). Therefore from a diagnostic perspective these rates must be viewed contextually, relative to rates observed at other similar locations, or (if data are available) historic rates at the same location. Inherent in this attempt to classify locations as dangerous is a variety of different criteria by which danger can be ascribed e.g. a relatively high collision rate, a rate which is not abnormally high but appears to be increasing etc. Further discussion of this is given in Section 5.8.

The challenge of determining the optimal allocation, henceforth described as the issue of hotspot identification and prediction, requires extensive analysis of road safety data, so as to best inform the subsequent decision-making. It is not sufficient however to simply determine where a treatment should be applied, we must also ensure that the treatment, once applied, is successfully improving road safety, and reducing collision/casualty numbers as intended. Again, this challenge of determining the effect of a given road safety treatment, henceforth referred to as the problem of scheme evaluation, must also make use of road safety data, so as to clear numerical evidence as to how effective a scheme has been, and therefore determine whether it should be maintained/implemented at further locations. Unfortunately, this usage of road safety data can be problematic and lead to biased conclusions if the data are not handled correctly, an effect felt most keenly where the datasets are relatively small. The main source of such bias is known most commonly as the regression to the mean effect.

1.2 Regression To the Mean

Understandably, road safety practitioners wish to have an empirical basis on which to base decisions regarding the implementation and retention of road safety treatments. Hence they rely largely on road safety data in order to make their decisions (while the role of intuition of an experienced practitioner familiar with the network cannot be understated, such experience is not always available). Unfortunately this can lead to problems (exacerbated when data are scarce) due to the potential confounding effects of regression to the mean (RTM) and temporal trends which may be present in the data.

The RTM effect was first documented by Sir Francis Galton in his paper "Regression towards mediocrity in hereditary stature", [Galton, 1886], in which he documented the effect whereby particularly tall parents would have children who were themselves tall, but not quite as tall as their parents, and hence their height "regressed to the mean" population height. This effect was mirrored in particularly shorter parents whose children's height would also regress to the population mean height and hence be slightly taller than their parents.

The first significant investigation into the RTM effect from a road safety perspective was by Ezra Hauer, see for example [Hauer, 1980], [Hauer, 1986], and can be defined in this context as the selection bias induced when a treatment is applied non-randomly, based on the responses of the individuals that are treated.

In the context of road safety, this can be observed when a site is monitored over a number of years, and the number of collisions observed in a year recorded. Due to the extensive number of possible reasons for a collision to occur, this value will fluctuate stochastically from year to year, despite there being no significant systemic change in the underlying collision rate determined by the inherent danger of the site itself. The reasons for this apparent random fluctuation are numerous, with key factors being the element of human error involved in many collisions, the blame for which cannot be ascribed to the location at which the error occurred, as well as the inconsistency with which minor collisions are recorded and reported, thereby inducing an issue of data quality into the equation. In modern times the ways in which collisions are reported has been standardised in a bid to reduce data inconsistencies (particularly with regard to collision severity, see Section 6.4), the human aspect of collision reporting (particularly with respect to

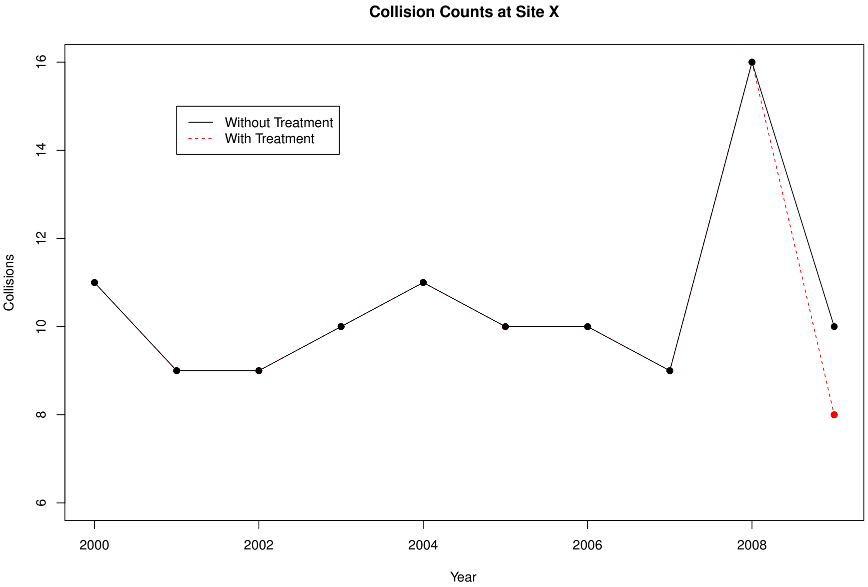


Figure 1.1: An illustrative example showing how reduction after treatment is applied (in 2008) could be (at least partially) explained by the reduction caused by an extreme value returning towards a baseline - the RTM effect

insurance premiums) means it is highly unlikely any dataset containing minor collisions will be perfect.

From a statistical perspective this variation can be thought of as making repeated observations from a Poisson distribution (since collision numbers are count data) with a fixed rate parameter λ over some fixed time interval, where successive observations will fluctuate (on average by $\sqrt{\lambda}$) despite there being no change in λ. Because of this fluctuation we occasionally see extreme high and low values, which, because there has been no change in the underlying rate, will almost always be followed by a less extreme value, i.e. a value closer to the true mean rate. It is this process of returning from an abnormal extreme value, i.e. regressing to the mean, that gives the RTM effect its name, and while the concept itself is simple severe problems can arise if it is not accounted for appropriately.

Figure 1.1 provides an illustrative toy example of this, in Figure 1.1 we observe a series of 10 yearly counts, from the years 2000-2009 at a site which has an underlying rate of 10 collisions per year. In the penultimate year, 2008, we observe an extreme count of 16 collisions, which assuming the data comes from a Poisson distribution with rate $\lambda = 10$, the probability of observing 16 or more collisions is $0.027 \approx 1$ in 37 chance. Considering it is not uncommon for practitioners to be monitoring networks made up of several hundreds of sites, the chance of extreme counts occurring on a network in any given year is certainly non-negligible. From the black line in Figure 1.1, we propose the scenario whereby the practitioner does not take action at the site in 2008, and we observe the RTM effect in the following observation in 2009, with the collision count returning to the mean value of 10. Problems arise when the abnormal observation is not acknowledged as such, and treatment is applied (or not applied, in the case of the abnormal observation being much lower than the mean rate) based on this extreme observation. Such a scenario is illustrated by the red dotted line in Figure 1.1, whereby after treatment, the accident count in 2009 reduces from 16 to 8, indicating an apparent treatment effect of –8 (8 fewer collisions due to treatment). However we know that the bulk of this reduction was due to the effect of the extreme value returning to a baseline level, the RTM effect, which accounts for –6 of the reduction, with only –2 being the true treatment effect. This is clearly a problem, since it indicates resources have been wasted on a site where treatment was not truly required (shown by the reduction without treatment from the black line in Figure 1.1), potentially meaning a site which was truly dangerous will now remain untreated in place of this site. Furthermore this would likely lead the practitioner to have a falsely inflated sense of the effectiveness of the treatment, thereby leading to it being implemented in the future, when in fact it may not be as effective as first believed, owing to this effect of RTM correction. While in the case of Figure 1.1, it may appear obvious that the 2008 count is an outlier, simply by examining the previous 8 points, it is not always possible to have this much data available. In practice this is often combined with public/political pressure on authorities to respond quickly to events, and so pragmatic approaches of waiting to collect more data before deciding on treatment are not always possible. While theoretically, randomised controlled trials of countermeasures are possible (and some small trials have taken place), it is often considered politically difficult (particularly given the often heated public reaction to increased presence of speed cameras for instance) to deploy countermeasures without thorough justification. Any controlled trial outcomes would also of course be subject to queries regarding the transferability of the results, between countries or sometimes even between areas within a country where the road safety system can be substanitally different. The collision counts may not provide enough information for the practitioner to be confident in their (non-)identification of an extreme value, particularly when data are limited, it is entirely possible the practitioner may be making judgements based on a single year's data, and so must appeal to additional sources of information in their analysis. It is important to note here that while in the example above the RTM effect leads to an reduction in collision count, the opposite case can also be true where a value is unusually low, and hence the RTM effect will cause an increase in collision count in the next time period. While this is not normally an issue in the context of scheme evaluation given that normally only locations with particularly high collision counts are treated, this is an issue for the problem of hotspot prediction, and this is discussed further in Chapter 5. The RTM effect is perhaps most commonly associated with the world of medicine, whereby many analogous examples to the road safety example described above can occur, for example if a patient presents with extreme symptoms, is prescribed a treatment and the symptoms reduce, how can we be certain how much the symptoms may have reduced without treatment? While this effect, often conflated with the placebo effect, can effectively be accounted for with a randomised control trial, this is not always possible in the case of extreme treatments (e.g. invasive surgeries which more frial patients may not be able to survive). For this reason statistical techniques are required to distinguish a treatment effect by accounting for confounding factors, and given the similarity between the problems of treatment effect calculation in medicine and road safety, it is unsurprising to see common approaches used in both settings, one key example of which is propensity scores as discussed in Chapter 4. Further examples detailing the RTM effect, and the potential pitfalls it causes, can be found in [García-Gallego et al., 2011].

Clearly effects such RTM can be much more easily identified when many years of data are available at each site, as the identification of unusual observations, and general patterns in time, becomes relatively simple. However, since as already stated, such data are not usually available to all practitioners, for this thesis we shall assume a “worst case scenario”, i.e. where only “before and after” data are available for scheme evaluation studies, and we have no expert prior information regarding our sites. The only area for which expert prior knowledge has been used, is in the formulation of some model parameters in the hotspot prediction model detailed in Chapter 5, however we believe these parameters to be applicable to all hotspot prediction datasets, and so the assumption of no dataset specific expert prior knowledge remains valid. Should additional information be available, this can easily be incorporated into the methods discussed throughout, however they are not a requirement, with this decision being made so as to make the methods proposed here useful for as many situations as possible. As such, for the purposes of model comparison, we shall focus chiefly on those which operate within a similar setup, that is models which have similar data requirements. This is again to keep in line with the principal aim of this research, to provide strong statistical methodologies for road safety analysis, which are as versatile, and thus as widely applicable, as possible.

1.3 Safety Performance Functions

As discussed in Section 1.2, in order to address the issue of identifying unusual collision counts in data, practitioners can instead appeal to additional sources of information, in the form of characteristic information regarding the sites on their network. The characteristic information can take any form (see [Wang et al., 2009], [Park and Abdel-Aty, 2016], and [Hanson et al., 2013] for studies using a large variety of characteristics). Common examples include:

- Average annual daily traffic (AADT)
- Speed Limit
- Average observed daily speed
- Percentage of drivers exceeding the speed limit
- Segment length (if sites are not designed to be of equal size)
- Road classification
- Road type
- Site location (e.g. Urban/Rural, Link/Intersection etc)
- Lane width

We can then use this characteristic information in order to determine what a “typical” collision count at a site displaying a given set of characteristics would be, by way of a regression model typically referred to as a safety performance function (SPF), or sometimes an accident prediction model (APM). There is extensive research in the literature detailing a variety of ways to construct SPFs in different scenarios, the simplest and most straightforward of which, is a standard log-linear GLM with P covariates, i.e.

$$\mu_i = \exp\left(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_P x_{i,P}\right) \tag{1.1}$$

where β_j is the fitted regression coefficient for characteristic, or covariate, j, β_0 is the fitted constant term, $x_{i,j}$ is the value of covariate j at site i, and μ_i is the fitted collision count obtained by our model for a site displaying the set of covariates $\boldsymbol{X}_i$. We note here that μ_i can be considered an estimate of λ_i provided by other sites via the SPF. When considering models over multiple time points we would expect the SPF estimate to vary also, be it due to a temporal trend effect built directly into the model (see Sections 4.5 and 5.4) or through varying covariate levels between time points. The covariates we shall analyse in this thesis are static, and so for notational convenience we do not include a time indicator on each covariate value, although it should be assumed in cases where variables vary in time, a time indicator should be added $x_{i,j,t}$ to denote the covariate value in time t. The implicit assumption in this form of modelling is that our errors will be independent conditional on this linear predictor, and so the inclusion of some trend effect is important when considering the potential temporal correlation present in such data. It is the fitted value, μ which provides valuable extra information in discerning the approximate underlying collision rate at a site, and thereby whether any recent observations are likely to be abnormal or not. However clearly since this is simply an estimate provided by a regression model, we expect there to be a degree of error involved, in addition we should not entirely discount data actually observed at our site simply because we think it may (at least partly) be abnormal. Methods for combining the two sources of information are discussed extensively in Chapter 2.

In order to maximise the versatility of the research carried out here, we shall, for the most part, retain a log-linear SPF (Equation (1.1)), so as to avoid placing any unnecessary data restrictions and allow the methods to be applied to as wide a variety of datasets as possible. However, as mentioned previously, there are a large variety of other SPFs available in the literature, many of which have been tailor-made to the area in which the study is taking place. Some bear a very close resemblance to the log-linear SPF described above, e.g. [Greibe, 2003] appear to place greater significance on the AADT covariate (seen as being the most important covariate in estimating road safety), and use an SPF of the form,

$$\mu_i = Q_i^{\beta} \exp\left(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_P x_{i,P}\right)$$

where Q is the AADT value, and β denotes an elasticity parameter, essentially incorporated to avoid the assumption of a linear relationship between collision rate and AADT. This model occasionally appears in epidemiology studies, where β is set to equal 1, leaving Q as an offset parameter, with the response variable providing the expected morbidity.mortality rate for a given disease. While this seems sensible, this can be shown to be exactly equivalent to the standard log-linear SPF given in Equation (1.1) by taking AADT on the log scale, i.e.

$$\begin{aligned}\mu_i &= Q_i^{\beta} \exp\left(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_P x_{i,P}\right) \\ &= \exp\left(\log\left(Q_i^{\beta}\right)\right) \exp\left(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_P x_{i,P}\right) \\ &= \exp\left(\beta \log\left(Q_i\right)\right) \exp\left(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_P x_{i,P}\right) \\ &= \exp\left(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_P x_{i,P} + \beta \log\left(Q_i\right)\right)\end{aligned}$$

which, subject to relabelling, is the same form as Equation (1.1). While AADT is often regarded as a highly important predictor variable, chiefly because it largely controls the risk exposure (clearly a location with a high AADT has more chance of a collision occurring than a location which has next to no vehicles passing it), it is often costly to measure AADT at all locations on a network. As such it is relatively rare for road safety practitioners (particularly in less affluent areas) to have access to AADT data for inclusion in their SPF. Traffic models are available which provide an estimated AADT which could be included, however there we introduce additional error into the model based on the accuracy of these simulations. In light of this it is common for other categorical variables (e.g. road class, urban/rural location etc) to act as proxies for AADT in an SPF, as these variables are often highly correlated with AADT anyway.

The SPFs described above are examples of fixed effects models, whereby the model coefficients are assumed constant across all sites being analysed. We can relax this assumption in a variety of ways. One approach would be to add a randomly distributed site specific term to the SPF to account for heterogeneity between sites, due to factors not included in the SPF i.e.,

$$\mu_i = \exp\left(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_P x_{i,P} + \sigma_i\right), \tag{1.2}$$

where σ_i is our site specific effect, independent of the regression coefficients $\boldsymbol{\beta}$. This model is sometimes referred to as a random parameters or random effects model [Chin and Quddus, 2003] [Hou et al., 2018] [Agbelie, 2016] or a random intercept model since the site-specific term is constant with respect to the other parameters and so simply affects the intercept of the link function, i.e.

$$\mu_i = \exp\left(\beta'_{i,0} + \beta_1 x_{i,1} + \cdots + \beta_P x_{i,P}\right),$$

where $\beta'_{0,i} = \beta_0 + \sigma_i$. Technically the model described in Equation (1.2) is a mixed effects model, since all regression parameters aside from the intercept, $\boldsymbol{\beta}_{-0} = (\beta_1, \ldots, \beta_P)$ are constant across all individuals, and hence are fixed effects. We can extend the generalisation of our model further by allowing not only the intercept to be site-specific, but the other model parameters also, giving rise to a fully random effects model,

$$\mu_i = \exp\left(\beta_{0,i} + \beta_{i,1} x_{i,1} + \cdots + \beta_{i,P} x_{i,P}\right) \tag{1.3}$$

There are a variety of mechanisms by which we can arrive at a random effects model. A common idea is to subset the collision dataset from which the SPF is obtained into separate components, usually with respect to binary categorical variables, (or occasionally small geographical areas), sometimes known as small area estimation [Ghosh et al., 1994]. After subsetting, SPFs are fitted independently to each subgroup of the dataset, using any remaining explanatory variables not used to form the subgroups of data. Hence each site i has SPF form given by Equation (1.3) where each element of the coefficient vector $\boldsymbol{\beta}_i = (\beta_{0,i}, \ldots, \beta_{P,i})$ is shared among all sites i in the same subgroup. Potential issues with this approach arise in the case where there are uneven splits in covariate distribution, leading to a subgroup(s) which does not contain enough data to properly fit a model. Further practical issues can include knowing which covariates to use to subset the data, and which to retain to be coefficients in the SPF. One particular advantage of considering a random effects model for μ_i is the potential for this to account for overdispersion in counts, an issue discussed later in Section 1.4, and is particularly prevalent when modelling casualty (as opposed to collision) counts, as discussed in Chapter 2.

Alternatively, should the dataset span a large area, the subsetting may take place geographically, as in distinct SPFs are fitted to sites grouped by their geographic region, as opposed to solely by categorical covariates, however the same procedure and warnings apply (see [Li et al., 2017]). Clearly subsetting in this manner can lead to modelling issues, since if there are a relatively small number of sites with a given combination of covariates (common in cases where there is a significantly uneven split in covariate makeup, or where the data has been subsetted too many times), there may not be sufficient data to reliably fit one or more of the SPFs. Because of this it is not straightforward to advocate general advice on data subsetting, a problem which is discussed further in Chapter 4. An alternative approach to deriving a random effects model based on geography, which does not require subsetting and hence avoids the issues outlined above, is geographically weighted regression.

The implicit assumption made when fitting models such as fixed effects models like Equation (1.1), or mixed/random effects models based on subsetting, is that the covariate effects are constant among all sites, or at least sites within the same subgroup. It is not difficult to conceive of scenarios where this assumption could prove false, for instance it would be realistic to suppose the safety level of a coastal town would be different to that of a large capital city, and so a single $Urban$ indicator variable, such as the one included in the Halle dataset in Chapter 5, would not be constant throughout the network.

To overcome this, we can loosen the restriction of a single, global covariate effect vector $\boldsymbol{\beta}$ across the network/subgroup, instead allowing each covariate effect to depend on spatial position, and so in effect replace equation (1.1) with,

$$\mu_i(t) = \exp\left(\beta_0\left(u_i, v_i\right) + \beta_1\left(u_i, v_i\right) x_{1,i} + \ldots + \beta_{n_p} x_{n_p} + \beta_t t\right). \tag{1.4}$$

where u_i and v_i are the respective longitudes and latitudes of site i, a model structure known as geographically weighted regression [Liu et al., 2017] [Li et al., 2013] [Gomes et al., 2017]. Clearly since we are attempting to implement a weighted regression analysis, we require a weight function to determine the weighting given to each observation when calculating the vector of regression coefficients, $\boldsymbol{\beta}(\boldsymbol{u}, \boldsymbol{v})$. A variety of inverse distance weight functions can be used, with the most common being the Gaussian,

$$w_{i,j} = \exp\left(-\frac{d_{ij}^2}{b}\right),$$

where $w_{i,j}$ is the weight given to data at site j in fitting with SPF at site i, d_{ij} is the distance between sites i and j, and b is the bandwidth parameter which needs to be estimated. In R the `spgwr` package calculates the bandwidth for a given dataset numerically by selecting the bandwidth value which minimised the cross validation score. The decision of how to define the distance d_{ij} can depend on the type of geographic effects expected in the dataset, where if the sites are in a location where general geographical/climatic features are likely to affect the data, the geographic distance between the two points may be most suitable. Conversely if the geographic conditions remain largely constant across the dataset, more local/infrastructure based changes may be better detected by defining d_{ij} to be the shortest link distance between sites i and j.

A key issue regarding the implementation of SPFs concerns the sites on which they are built. It is standard statistical practice when fitting regression models to try and avoid extrapolation, i.e. to attempt to ensure that the dataset on which a model is built is sufficiently "similar" to the data points on which it is then applied and inferences drawn. This issue is prevalent in road safety analyses, particularly scheme evaluation analyses (Chapter 2), although many authors/practitioners to not pay heed to it, either failing to verify the suitability of the sites from which they build their SPFs, referred to as "comparison" sites in the context of scheme evaluation, or use so called "off the shelf" SPFs (a common example being the CMF Clearinghouse used alongside the Highway Safety Manual (HSM) by practitioners in the USA [U.S. Department of Transportation,]), which provide no guarantee of relevance to the sites being analysed. Despite their prevelance among practitioners, problems with using approaches such as the HSM have already been discussed in the literature, see for example [Shirazi et al., 2016], [Park and Sahaji, 2013] and [Farid et al., 2016]. The exact nature of the danger posed by neglecting this requirement of site suitability when fitting SPFs is discussed further, and numerically demonstrated, in Chapter 3.

1.4 Modelling Overdispersion

Throughout this thesis we shall be attempting to model and predict road collision counts at various points along a road network. These will clearly be count data, with a widely used model for count data being the Poisson distribution, that is

$$Y \sim Po\left(\lambda\right), \qquad \lambda > 0.$$

A key feature of a Poisson random variable, is that it has mean equal to its variance,

$$E\left(Y\right) = \mathrm{Var}\left(Y\right) = \lambda,$$

which imposes a strong restriction on the dispersion of data we can use the Poisson distribution to model. In many cases, with road traffic collisions being among them, this assumption of equality between mean and variance is often not met, but rather we observe a situation where the variance is proportional to the mean,

$$\mathrm{Var}\,(Y) = \theta\lambda, \qquad \theta > 0,$$

with the most common case in road safety, being the case where $\theta > 1$, leading to overdispersion,

$$\theta = 1 + \gamma, \qquad \gamma > 0,$$

where γ clearly controls the severity of the excess variance, and so is typically referred to as the overdispersion parameter. There are cases where the data would be underdispersed, $\theta < 1$, typically in situations where there are excess zeroes in the data, however this may be accounted for using a zero-inflated Poisson model, which accounts for excess zeroes conditional on the linear predictor (in this case μ_i). The case where $\theta = 1$ returns the standard Poisson model.

One approach toward acknowledging overdispersion is through the Bayesian paradigm whereby instead of treating λ as a fixed value, we allow it to vary according to a specified prior distribution. We retain the same distribution for collision count Y, however in this case this distribution is conditional on λ,

$$Y|\lambda \sim Po(\lambda), \qquad \lambda > 0.$$

A convenient choice of prior distribution for λ is the Gamma distribution, since this is the conjugate prior for Poisson distributed data, thereby allowing an analytic posterior distribution for λ to be obtained. Hence we have

$$\lambda \sim Ga(\alpha, \beta),$$

which has density $g(\lambda)$ given by

$$g(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}\mathrm{e}^{-\beta\lambda}, \qquad \alpha, \beta > 0.$$

Hence we can obtain the unconditional distribution of Y, $f(Y)$,

$$\begin{aligned} f\,(y) &= \int_{\Lambda} f\,(y|\lambda)\, g\,(\lambda)\, \mathrm{d}\lambda, \qquad i = 1, \cdots, n, \\ &= \int_0^{\infty} \frac{\lambda^y}{y!}\mathrm{e}^{-\lambda}\frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}\mathrm{e}^{-\beta\lambda}\mathrm{d}\lambda, \\ &= \frac{\beta^{\alpha}}{\Gamma(\alpha)y!}\int_0^{\infty} \lambda^{y+\alpha-1}\mathrm{e}^{-(\beta+1)\lambda}\mathrm{d}\lambda \end{aligned}$$

By considering that if $X \sim Ga\,(y + \alpha, \beta + 1)$ then X has density given by,

$$f(X) = \frac{(\beta+1)^{(y+\alpha)}}{\Gamma\,(y+\alpha)}x^{y+\alpha-1}\mathrm{e}^{-(\beta+1)x}, \qquad x > 0,$$

and since $f(X)$ is a PDF we have $\int_0^\infty f(X)\mathrm{d}x = 1$ and so,

$$\begin{aligned}\int_0^\infty \lambda^{y+\alpha-1}\mathrm{e}^{-(\beta+1)\lambda}\mathrm{d}\lambda &= \int_0^\infty \frac{(\beta+1)^{(y+\alpha)}}{\Gamma(y+\alpha)}\frac{\Gamma(y+\alpha)}{(\beta+1)^{(y+\alpha)}}x^{y+\alpha-1}\mathrm{e}^{-(\beta+1)x}\mathrm{d}x,\\ &= \frac{\Gamma(y+\alpha)}{(\beta+1)^{(y+\alpha)}}\int_0^\infty \frac{(\beta+1)^{(y+\alpha)}}{\Gamma(y+\alpha)}x^{y+\alpha-1}\mathrm{e}^{-(\beta+1)x}\mathrm{d}x\\ &= \frac{\Gamma(y+\alpha)}{(\beta+1)^{(y+\alpha)}}\times 1\\ &= \frac{\Gamma(y+\alpha)}{(\beta+1)^{(y+\alpha)}}.\end{aligned}$$

Hence we have

$$\begin{aligned}f(y_i) &= \frac{\beta^\alpha}{\Gamma(\alpha)y!}\frac{\Gamma(y+\alpha)}{(\beta+1)^{(y+\alpha)}}\\ &= \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\Gamma(y+1)}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^y\\ &= \binom{y+\alpha-1}{y}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^y\end{aligned}$$

which we recognise as the probability mass function (PMF), $f(y) = \Pr(Y = y)$, of a Negative Binomial random variable and hence

$$Y \sim NB\left(r = \alpha, p = \frac{1}{\beta+1}\right).$$

We have free choice of hyperparameters α and β. Choosing $\alpha = \gamma$ and $\beta = \frac{\gamma}{\mu_i}$ gives rise to the Empirical Bayes methodology for safety scheme evaluation, discussed in Section 2.2.

1.5 Bayesian Inference

The majority of the statistical inference carried out in this thesis will be done within a Bayesian framework. Within the Bayesian paradigm, all modelling parameters and variables are considered equally to be examples of random variables, our beliefs and knowledge about which are described individually through probability distributions, and collectively through joint probability distributions. The choice to operate within the Bayesian framework provides many advantages, chiefly an enhanced natural framework for incorporating uncertainty in models and parameters, along with the potential for inclusion of expert prior information, a component which is crucial in small data problems, like those this thesis shall mainly focus on.

The main result of Bayesian statistics is Bayes Theorem which states that the posterior distribution of Θ, denoted $\pi(\theta|\boldsymbol{x})$, a vector of random variables, can be obtained by multiplying the prior distribution of Θ, $\pi(\theta)$, with the likelihood obtained by observing data $\boldsymbol{x}$, $\mathrm{L}\left(\boldsymbol{x}|\theta\right)$, and dividing by a normalising value constant with respect to Θ, $f(\boldsymbol{x})$. In formula form this is,

$$\pi\left(\theta|\boldsymbol{x}\right) = \frac{L\left(\boldsymbol{x}|\theta\right)\pi\left(\theta\right)}{f\left(\boldsymbol{x}\right)}.$$

Oftentimes the normalising constant is omitted from Bayes Theorem and it is sufficient to say that the posterior distribution is proportional to the prior distribution multiplied by the likelihood,

$$\pi\left(\theta|\boldsymbol{x}\right) \propto \mathrm{L}\left(\boldsymbol{x}|\theta\right)\pi\left(\theta\right). \tag{1.5}$$

1.5.1 Markov Chain Monte Carlo

When carrying out a Bayesian analysis, often the primary goal is to obtain a posterior distribution for the parameter(s) of interest, however unless the model in question is relatively simple, and/or certain restrictive modelling choices are made, obtaining an analytic form for the posterior distribution is not possible. Fortunately techniques have been developed for such cases whereby samples from the posterior distribution may be obtained numerically. Currently by far the most common mechanism for doing this is by implementing a Markov chain monte carlo (MCMC) algorithm, of which there are several kinds. For certain Bayesian modelling structures, such a conjugate models where the prior and posterior distributions are both from the same distribution family, the posterior distribution can be obtained analytically, such as the Poisson-Gamma model structure discussed in Section 2.2. Conjugate and semi-conjugate models can be sampled using a Gibbs Sampler [Casella and George, 1992], whereby successive posterior elements of θ are sampled from their respective fully conditional distributions (FCDs) – the analytic form of the posterior distribution $\pi(\theta|\boldsymbol{x})$ up to a constant of proportionality, obtained using Bayes Theorem (Equation (1.5)). Hence the general algorithm for a Gibbs Sampler with M iterations can be described as,

1. Initialise the chain at its initial value $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_n^{(0)}\right)$. Set counter $m = 1$.

2. Draw a sample for each element of θ from its FCD,

$$\begin{aligned}
\theta_1^{(m)} &\sim \pi\left(\left(\theta_2^{(m-1)}, \ldots, \theta_n^{(m-1)}\right)|\boldsymbol{x}\right)\\
\theta_2^{(m)} &\sim \pi\left(\left(\theta_1^{(m)}, \ldots, \theta_n^{(m-1)}\right)|\boldsymbol{x}\right)\\
&\vdots\\
\theta_n^{(m)} &\sim \pi\left(\left(\theta_1^{(m)}, \theta_2^{(m)}, \ldots, \theta_{n-1}^{(m)}\right)|\boldsymbol{x}\right)
\end{aligned}$$

3. If $m = M$ stop, else set $m = m + 1$ and go to step 2.

The resulting samples of $\boldsymbol{\theta}$ form a Markov chain which has stationary distribution equal to the posterior distribution $\Pi(\boldsymbol{\theta}|\boldsymbol{x})$. The initial value of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(0)}$, can be chosen arbitrarily (commonly the mean of the prior distribution $\pi(\boldsymbol{\theta})$ is chosen as the initial value) and so we do not expect this to be a sample from the posterior distribution. Hence the initial part of the chain which has not reached convergence to its stationary distribution (the posterior distribution) is often referred to as "burn-in" and is discarded from the final samples. Because of this it should not matter where the chain is initialised as it should always converge to the same stationary distribution. Checks of convergence can either be done numerically, or simply by initialising the chain at different points and ensuring the chains reach the same distribution by inspecting the traceplots of each element of the parameter vector $\boldsymbol{\theta}$.

We provide an example of a Gibbs sampler for a conjugate Normal-Normal model with unknown variance, which in addition to being a standard example of a Gibbs sampler, we shall also make use of this for an analysis in Chapter 6. In this context it is often convinient to parameterise a Normal distribution in terms of its precision ν which is simply the reciprocal of its variance, $\nu = \frac{1}{\sigma^2}$. Hence our conjugate model structure is,

$$\begin{aligned} y_i|\mu,\nu &\sim N\left(\mu, \frac{1}{\nu}\right), \qquad i = 1,\dots,n \\ \mu|\nu &\sim N\left(m_0, \frac{1}{p_0}\right), \\ \nu|\mu &\sim Ga\left(g_0, h_0\right). \end{aligned}$$

From this we derive the following FCDs for μ and ν,

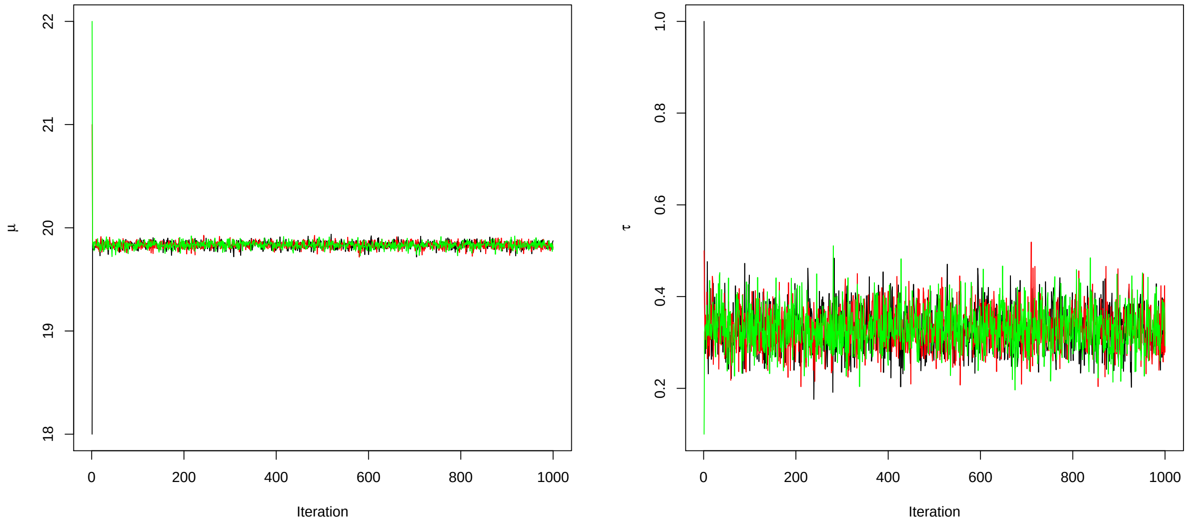
$$\begin{aligned} \mu|\nu &\sim N\left(\frac{m_0 p_0 + n\bar{y}\nu}{p_0 + n\nu}, \frac{1}{p_0 + n\nu}\right) \\ \nu|\mu &\sim Ga\left(g_0 + \frac{n}{2}, \frac{1}{2}\sum\left(y_i - \mu\right)^2 + h_0\right) \end{aligned}$$

We demonstrate this using example data,

$$y_i \sim N\left(20, 3^2\right), \qquad i = 1,\dots,n,$$

for $n = 100$ data points and model this using a conjugate Normal-Normal structure,

$$\begin{aligned} Y_i &\sim N\left(\mu, \frac{1}{\nu}\right), \qquad i = 1,\dots,n \\ \mu &\sim N\left(m_0, \frac{1}{p_0}\right) \\ \nu &\sim Ga\left(g_0, h_0\right) \end{aligned} \tag{1.6}$$



(a) Traceplots for the posterior samples for μ (b) Traceplots for the posterior samples for ν

Figure 1.2: Output for a Gibbs sampler for a Normal-Normal model using initialisation points $\mu^{(0)} = 18, 21, 22$ and $\nu^{(0)} = 1, 0.5, 0.1$ as the black, red and green lines respectively.

and specify vague prior information with the hyperparameter choices $m_0 = 0$, $p_0 = \frac{1}{100}$, $g_0 = h_0 = \frac{1}{100}$ We use varying initialisation points $\left(\mu^{(0)}, \nu^{(0)}\right) = (10, 1)$ and $\left(\mu^{(0)}, \nu^{(0)}\right) = (30, 0.01)$. Output from these Gibbs samplers is given in Figure 1.2.

Procedures such as Gibbs samplers are only possible to implement when specific combinations of prior and data distributions are used to give analytic FCDs, when this is not the case other methods must be used. One of the most common approaches in this case is to implement a Metropolis-Hastings (MH) algorithm ([Hastings, 1970]), which again produces a Markov Chain which has stationary distribution equal to the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$. The main difference between MH and Gibbs samplers is that whereas in the case of Gibbs we sample directly from the FCD, for MH we must propose posterior values for $\boldsymbol{\theta}$ which we then accept with a given probability determined by the observed data likelihood and the prior distribution for $\boldsymbol{\theta}$. Hence a typical MH algorithm would be:

1. Initialise the chain at its initial value $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_n^{(0)}\right)$. Set counter $m = 1$.

2. For each element θ_j:

   - Sample a proposal value for θ_j, θ_j^* from the proposal distribution of θ_j,

$$\theta_j^* \sim q\left(\cdot|\cdot\right)$$

- Set $\theta_j^{(m)} = \theta_j^*$ with probability α,

$$\alpha = \min\left(1, \frac{\pi\left(\theta_j\right) L\left(\boldsymbol{x}|\boldsymbol{\theta}\right) q\left(\theta^{(m-1)}|\theta^*\right)}{\pi\left(\theta^*\right) L\left(\boldsymbol{x}|\boldsymbol{\theta}^*\right) q\left(\theta^*|\theta^{(m-1)}\right)}\right)$$

  else set $\theta_j^{(m)} = \theta_j^{(m-1)}$

3. If $m = M$ stop, else set $m = m + 1$ and go to step 2.

Because of the additional accept/reject component, MH is objectively a less efficient sampler than the Gibbs sampler and so in situations where a Gibbs sampler can be used it is always preferable to use this over MH. There are many choices of proposal distribution $q(\cdot|\cdot)$ which can be used in an MH algorithm, the most common being a Normal distribution centred on the current value of θ_j known as a Metropolis random walk,

$$\theta_j^* \sim N\left(\theta_j^*, \epsilon_j\right).$$

In the case of a Metropolis random walk – as with any symmetric proposal distribution – we have $q\left(\theta^{(m-1)}|\theta^*\right) = q\left(\theta^*|\theta^{(m-1)}\right)$ and so the components relating to the proposal distribution cancel out in the formula for the acceptance probability α, meaning they do not need to be calculated. The parameter ϵ_j controls the magnitude of difference between the proposed values of $\boldsymbol{\theta}$ and the current state of the chain, and hence can determine the efficiency with which the Markov chain produced by the MH algorithm converges to, and explores and samples from, its stationary distribution which is the posterior distribution for $\boldsymbol{\theta}$. Values of ϵ_j which are too small will lead to proposal values of θ_j being very close to the current value, meaning that if the initial value $\theta_j^{(0)}$ is far from the stationary distribution, the chain will converge slowly and a large amount will be lost as burn-in, similarly it will lead to a chain with a high level of autocorrelation (correlation between successive values of the chain) in the chain indicating the posterior distribution is not being explored effectively. Equally, selecting values of ϵ_j that are too large will lead to proposed values far from the current accepted value of the chain, thereby increasing the likelihood that unsuitable values of θ_j will be proposed and hence the acceptance rate (the number of proposal values that were accepted divided by the number of iterations of the chain) will be low, again meaning we would not be effectively sampling from the posterior distribution. Selecting a good value of $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$, known as “tuning” the chain is important, and is often done manually via inspecting traceplots of the elements of $\boldsymbol{\theta}$ and by studying the acceptance rates of the chain, with rates in the range of 20% - 30% considered optimal.

We demonstrate this using the Normal-Normal model described in Equation 1.6 using a random walk Metropolis algorithm with varying choices of innovation parameter ϵ. Traceplots for the parameter μ are given in Figure 1.3.

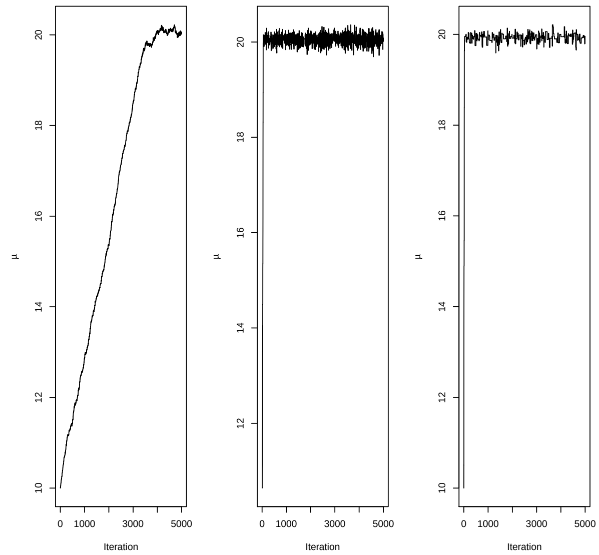


Figure 1.3: Traceplots for the Normal mean parameter μ from a Metropolis-Hastings algorithm with innovation parameter $\epsilon = 0.01, 0.5, 2$ respectively.

From Figure 1.3 we observe that for $\epsilon = 0.01$, the chain is too "hot", it updates very frequently, with an acceptance rate of 0.80, but the updates are very small, meaning the chain takes a long time to reach its stationary distribution, with successive samples being highly correlated, and hence the chain explores the posterior distribution inefficiently. For $\epsilon = 2$ the chain is too "cold", the update changes are relatively very large but there is a much higher rate of rejection, with an acceptance rate 0.06 meaning we reject 94% of all proposals, vastly reducing the effectiveness of the scheme. Finally for $\epsilon = 0.05$ we observe the chain converges quickly, but maintains a reasonable acceptance rate of 0.24, meaning we are less wasteful in terms of number of rejections, while maintaining a reasonable size of update, thereby reducing the autocorrelation of the chain and allowing us to explore the posterior distribution more effectively. This demonstrates the "Goldilocks principle" of parameter tuning, whereby we wish to select an innovation parameter which is not too small nor too large.

For complicated models MH sampling can be inefficient and computationally expensive, although there are adjustments which can be made to improve efficiency, for instance if posterior draws between different parameters are highly correlated, carrying out a joint update, whereby proposal values for the parameters in question are sampled simultaneously from a joint proposal distribution with a similar degree of correlation may improve the algorithm's performance. Recently newer algorithms such as Hamiltonian Monte Carlo (HMC) ([Neal, 2011]) have been developed with the aim of providing more efficient methods of sampling from posterior distributions. In some contexts, the model likelihood can be extremely computationally costly to calculate, or in some cases completely intractable, which has lead to a rise in likelihood-free inferential techniques such as Approximate Bayesian Computation (ABC) ([Diggle and Gratton, 1984]) being developed. Whilst these techniques are not necessary for the models developed in this thesis, we do take advantage of assumptions inherent in models we develop, mainly that we consider collision counts between sites to be conditionally independent of each other given the model parameters, to allow us to carry out MCMC updates and likelihood calculations in parallel, thereby reducing the computational runtime of the MCMC algorithms implemented.

1.5.2 Quantifying Uncertainty

A defining characteristic of a Bayesian approach to statistics is the interpretation of modelling parameters. Whereas in the non-Bayesian frequentist setting parameters are treated as having fixed, single values, with uncertainty on these estimates expressed via standard errors, within the Bayesian setting parameters are treated as having full distributions. One of the main areas this difference manifests itself is within the area of quantifying uncertainty on parameter estimates, via confidence intervals within the non-Bayesian setting, and credible intervals within the Bayesian framework. The purpose of an $\alpha\%$ confidence/credible is to provide values u and l such that

$$\Pr(l < \theta < u) = \alpha. \tag{1.7}$$

Under the frequentist paradigm, this is impossible, since as a parameter θ would take a fixed value and so the probability it has value between u and l is either 1 or 0 depending on whether $l < \theta < u$ is true or not. In the Bayesian setting since parameters have distributions, meeting the definition in Equation (1.7) is possible, simply by selecting values u and l between which $\alpha\%$ of the probability density of θ lies. This therefore provides an infinite number of possible $\alpha\%$ credible intervals (for $0 < \alpha < 100$) and so Bayesians often choose to work with the narrowest possible interval, which for a unimodal distribution will be the range of values containing the highest probability density, known as the highest density interval (HDI). Examples of credible intervals for a standard Normal distribution are given in Figure 1.4.

An additional commonly used statistical mechanism for quantifying uncertainty is the prediction interval, whereas a confidence interval provides a range of values which contain the mean of a particular parameter which probability α, a prediction interval provides a range of values between which a future observation will fall with probability α. Within the context of Bayesian inference we can obtain a full predictive distribution by obtaining the conditional probability distribution of y given parameter θ over the full parameter

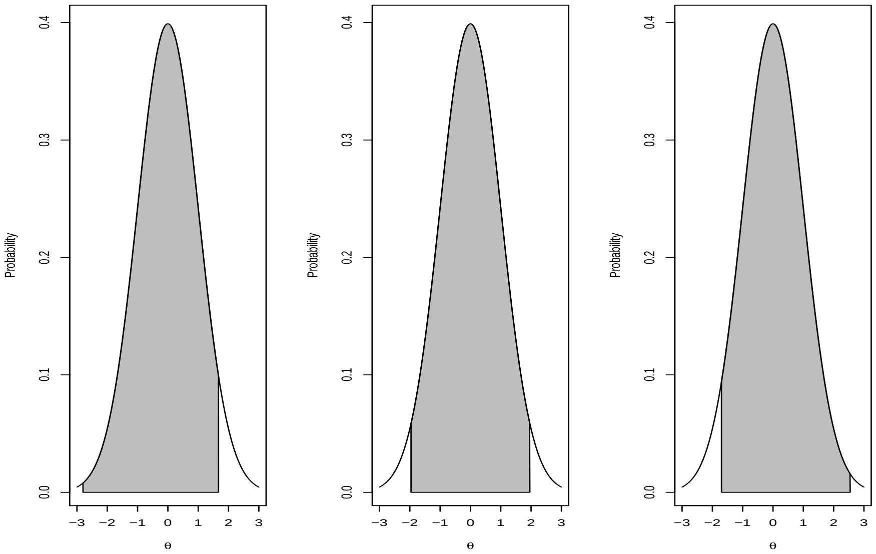


Figure 1.4: 3 plots showing possible 95% credible intervals for a standard Normal distribution. The central plot shows the HDI which coincides with the frequentist Wald confidence interval for a Normal distribution.

space for θ, denoted Θ. If θ is a continuous parameter we can obtain a distribution for a new value x' via,

$$f\left(x'|\boldsymbol{x}\right) = \int_{\Theta} f\left(x'|\theta, \boldsymbol{x}\right) g(\theta) d\theta \tag{1.8}$$

where $f\left(x'|\boldsymbol{x}\right)$ is the predictive distribution, $f\left(x'|\theta, \boldsymbol{x}\right)$ is the believed distribution of future values of $\boldsymbol{x}$ and $g\left(\theta\right)$ represents our beliefs regarding θ. If we have not observed any data $\boldsymbol{x}$ then our beliefs about θ will simply be the prior distribution for θ, $g(\theta) = \pi(\theta)$, and hence $f\left(x'\right)$ will be a prior predictive distribution. Conversely, if we have observed data then our beliefs about θ will be represented by its posterior distribution, i.e. $g(\theta) = \pi(\theta|\boldsymbol{x})$ and hence $f\left(x'|\boldsymbol{x}\right)$ will be a posterior predictive distribution, which is the predictive distribution we shall use in this thesis. If θ is a discrete parameter then the integral in Equation (1.8) is replaced by a summation

$$f\left(x'|\boldsymbol{x}\right) = \sum_{\Theta} f\left(x'|\theta, \boldsymbol{x}\right) g(\theta) d\theta.$$

From this distribution we can therefore obtain α% predictive intervals in the same way as for credible intervals, by obtaining limits u and l such that

$$\Pr(l < x' < u) = 0.95,$$

i.e. find limits between which 95% of the predictive distribution density lies. Again there are infinitely many possibilities for this (for $0 < \alpha < 100$) and so often we choose the HDI as the predictive interval. We note here the importance of using a full predictive interval rather than using a credible interval for the posterior distribution of θ, as this allows us to account for uncertainty in future observations of x. Similarly we should resist the temptation to generate a predictive distribution by taking θ as its posterior mean since this would fail to acknowledge uncertainty in our posterior beliefs regarding θ.

1.6 Road Safety Datasets

Throughout this thesis we shall make use of several real world datasets to demonstrate the techniques being discussed, and the results that can be obtained.

1.6.1 Northumbria Dataset

In the case of scheme evaluation techniques discussed in Chapter 2 we shall make use of data provided as part of a study into the effectiveness of mobile safety cameras, provided by the Northumbria Safety Camera Partnership (NSCP), now the Northumbria Safer Roads Initiative (NSRI) ([NSR,]). This data comprises of 67 non-treated comparison sites, from which we can obtain the SPF, and 56 sites which were treated by having a mobile safety camera stationed at the site. This data was collected over the years 2001-2003 (the before period) and 2004-2006 (the after period), and is slightly unusual in that it comprises casualty counts (i.e. the number of people requiring some form of medical treatment as a result of a collision) as opposed to collision counts which shall be the case for the other datasets, however casualty counts and collision counts are largely interchangeable with respect to the statistical modelling and so this will not be an issue. Covariate data were collected at each comparison and treated site for the purpose of training and fitting an SPF, comprising of continuous covariates which were averaged over their respective 3 year period, categorical covariates which remained constant throughout the data, and casualty counts which were aggregated. Hence the resulting dataset comprised of a single casualty count along with single observations for each covariate in the before period at each of the comparison and treated sites, along with a single casualty count in the after period at each of the treated sites. The covariates collected were:

- Speed Limit (5 levels: 30, 40, 50, 60, 70mph)
- Mean vehicle speed (mph)
- 85th percentile speed (mph)
- Percentage of drivers exceeding the speed limit

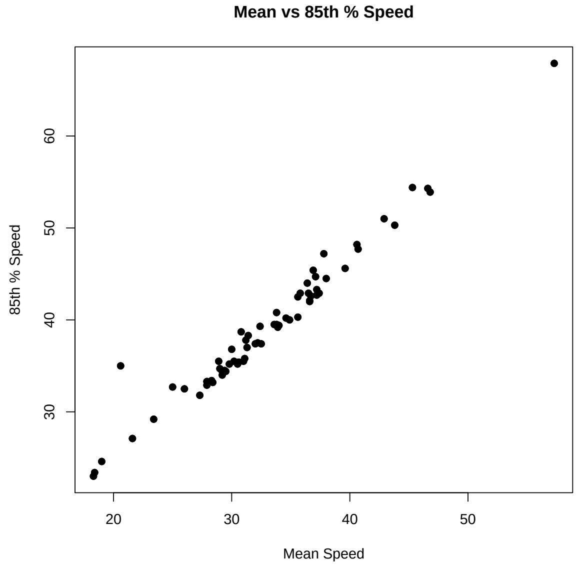


Figure 1.5: Plot of mean speed vs 85th percentile speed from the Northumbria dataset, showing a high degree of correlation

- Percentage of drivers 15mph over the speed limit
- Average daily traffic flow (in thousands)
- Categorical variable corresponding to the road classification at the site (4 levels: A, B, C and U)
- Indicator variable corresponding to whether the site was on a single or dual carriageway

Clearly from these covariates there is a strong likelihood of multicollinearity, particularly with respect to mean vehicle speed with 85th percentile speed. The Pearson correlation coefficient for this pair of covariates is $r = 0.97$ for the comparison data, with a plot given in Figure 1.5.

From this we see a high degree of multicollinearity between the covariates, and so remove the 85th percentile variable (this decision is largely arbitrary, however we chose to retain the mean over the quantile data since it is informed by all available speed data). It is commonly known that including highly correlated covariates can lead to hugely inflated standard errors in the regression coefficients for those covariates, thereby potentially vastly over inflating the posterior uncertainty in our analyses. This could further cause imprecise and potentially inaccurate regression coefficient estimates to be drawn, which in themselves are often useful for practitioners to discern the effect different covariates appear to have on collision counts on the network as a whole. For model

		x_1	x_2	x_3	x_{4A}	x_{4B}	x_{4C}	x_{4U}	y
Comparison	Mean/Prop.	32.95	35.73	9.13	0.37	0.16	0.30	0.16	4.28
	St. Dev.	6.89	29.63	6.48	–	–	–	–	4.77
Treated	Mean/Prop.	36.63	38.45	6.92	0.50	0.23	0.18	0.09	7.79
	St.Dev	9.90	21.67	6.48	–	–	–	–	5.38

Table 1.1: Table showing the mean and standard deviation of the continuous covariates, and proportion for the categorical covariates, from the Northumbria before-and-after dataset for the comparison and treated pools of sites

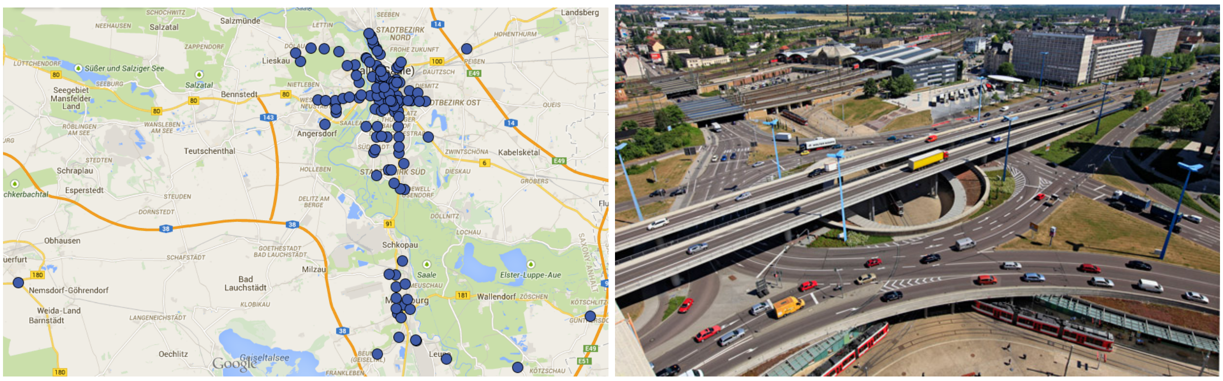
simplification purposes, as in Chapter 5, we elect to use frequentist variable selection procedures to identify a suitable subset of available potential SPF covariates by removing those deemed to be non-significant in modelling collision rates. Here we carry out a backwards stepwise regression to remove non-significant explanatory covariates we found that speed limit, percentage of drivers 15mph over the speed limit and whether the site was on a dual carriageway were non-significant $Pr(Z > |z|) > 0.05$ and so were not retained in the analysis. Hence we retain the following final covariates:

- Mean speed (x_1)
- Percentage exceeding the speed limit (x_2)
- Average daily flow (x_3)
- Road class ($x_{4A}, x_{4B}, x_{4C}, x_{4U}$)

Summary statistics for this dataset (along with casualty counts y) are provided in Table 1.1

1.6.2 Halle Dataset

The dataset used to train the initial hotspot prediction model described in Chapter 5 was provided by industrial partners at PTV Group [PTV Group, b] and consists of collision counts and covariates data, collected at 734 nodes over 9 years (2004-2012) in the city of Halle, Saxony-Anhalt in Germany. This dataset is available through an “Open Data Commons Open Database Licence”, available at “http://dx.doi.org/10.17634/154300-33” (contact Newcastle Research Data Service at rdm ncl.ac.uk for access). The covariates used to train the SPF were all categorical and constant across the four years with the exception of average daily traffic flow, which was taken from model estimates provided by PTV Group’s VISUM Safety software [PTV Group, a]. The covariates included are:



(a) The distribution of sites around the city of Halle

(b) The busiest intersection in terms of traffic volume in Germany - Riebeckplatz, Halle

- Indicator variable corresponding to if the node was in an urban location (x_1)
- Indicator variable corresponding to if the node was at an intersection (x_2)
- Indicator variable corresponding to if traffic signals were present at the node (x_3)
- Categorical variable corresponding to the speed limit at the node (6 levels: 30, 45, 50, 60, 70, 80km/h) ($x_{4A}, x_{4B}, x_{4C}, x_{4D}, x_{4E}, x_{4F}$)
- Indicator variable corresponding to if the node was on a major road (x_5)
- Indicator variable corresponding to if the node was at a major intersection (x_6)
- Indicator variable corresponding to if the node was at a four-legged intersection (x_7)
- Natural log of daily traffic flow from the major leg of the intersection (x_8)
- Natural log of daily traffic flow from the minor leg of the intersection (x_9)
- The year corresponding to each observation (t)

All variables were found to be statistically significant $Pr(Z > |z|) < 0.05$ and so were retained in the analysis. Summary statistics for this dataset are provided in Table 1.2

1.6.3 Halle Zonal Dataset

A second dataset from the city of Halle is used in Chapter 6 when modelling collision severities and causation factors. This dataset is taken over the same time period as the dataset described in 1.6.2, the 9 years from 2004-2012, but rather than be collision counts recorded at individual sites, this dataset consists of counts over 59 distinct regions of Halle. These collision counts are then disaggregated by month $s = 1, \ldots, 12$ and severity

	x_1	x_2	x_3	x_5	x_6	x_7	x_8	x_9
Mean/Prop.	0.91	0.86	0.27	0.63	0.20	0.24	6.70	3.87
St. Dev.	–	–	–	–	–	–	3.11	3.53

	x_{4A}	x_{4B}	x_{4C}	x_{4D}	x_{4E}	x_{4F}
Prop.	0.37	0.12	0.22	0.18	0.04	0.07

	y_{04}	y_{05}	y_{06}	y_{07}	y_{08}	y_{09}	y_{10}	y_{11}	y_{12}
Mean	3.65	3.73	3.57	3.71	3.55	3.64	3.29	3.11	2.97
St Dev.	4.63	4.95	5.07	4.98	4.51	4.80	4.40	4.40	4.57

Table 1.2: Tables showing the the mean and standard deviation of the continuous covariates, and proportion for the categorical covariates, as well as collision counts for all years for the Halle hotspot dataset

$k = 1, 2, 3$ (corresponding to collisions which: caused a fatality or serious injury, caused a slight injury, or caused no injury). The dataset includes the latitude and longitude of each regions centroid so as to enable a spatial analysis.

1.6.4 Florida Dataset

The dataset used to introduce the idea of seasonal and spatial effects in collision modelling in Chapter 6 was provided by the Florida Department of Transport [FDOT, 2019] and contains collision rates (collisions per vehicle kilometre travelled) at 52 traffic analysis zones (TAZs) across the state of Florida, U.S.A., over 46 years (1960-2015), although a significant amount of this data are missing, making any kind of longitudinal study impossible. The collision rates are disaggregated by month, allowing a seasonality component into the model, and the dataset contains the longitude and latitude of the centroid of each zone, enabling spatial effects to be investigated.

1.7 Thesis Structure

The main bulk of this thesis is devoted to examining two main tasks carried out by road safety practitioners: road safety scheme evaluation, and collision hotspot prediction. The aim of this is to investigate and develop methodologies which are as versatile and widely applicable as possible, to that end we shall focus on methods which:

- Require as little data as possible to achieve sensible results, making no assumptions or restrictions on the amount of data avilable to a practitioner

- Avoid specific data requirements with respect to data frequency nor covariate collection, meaning the methods here can be applied to as wide a variety of datasets as possible

- Are as autonomous as possible, thereby not making any assumptions regarding the availability of expert prior knowledge and hence making the models as data driven as possible.

Chapter 2 discusses and compares commonly used methods for scheme evaluation and their relative advantages and disadvantages, before carrying out a sensitivity analysis on the estimates of treatment effect. Chapter 3 describes issues arising from the approaches discussed in Chapter 2, providing a quantified example using simulated data, before Chapter 4 presents novel approaches to overcome them. Chapter 5 discusses the issue of hotspot prediction, before providing a novel approach to carrying this out, with applications to real data, and a comparison of other approaches toward hotspot identification. Chapter 6 gives extensions to the model described in Chapter 5, to allow for extra factors including seasonal and spatial trends, as well as the ability to account for collision severity. Finally Chapter 7 provides conclusions to the work carried out thus far, and discusses future avenues for the research in this project. All statistical analysis in this research was carried out using the software `R` [R Core Team, 2019], with models described in Chapters 2 and 5 using the `rjags` package [Plummer, 2003] to implement the MCMC algorithm.

The main research contributions of this thesis are: an investigation into the usage of the posterior predictive distribution in scheme evaluation studies (Chapter 2); a numerical demonstration of the effects of non-exchangeable comparison pools in before-and-after studies (Chapter 3); a new method developing bespoke SPFs using propensity score weighted regression (PSWR) to overcome issues of reference site selection, which then incorporates a data-driven parametric approach for trend estimation (Chapter 4); a new, robust method for hotspot prediction, which downweights observations further into the past deemed to be less informative, whilst accounting for network wide effects through an SPF with time component, but allowing site specific deviations from these (Chapter 5); an extended hotspot model which also allows for the inclusion of seasonal effects, as well as the incorporation of seasonal and spatial effects to allow information sharing between observations, with the additional potential to interpolate site specific effects spatially, as well as account for changes in collision severity and/or factors relating to collisions e.g. speed related collisions, collisions in a particular weather type etc (Chapter 6).

Chapter 2

Scheme Evaluation Studies

Notation

Below is a summary of the notation used to describe the statistical framework used in this chapter. This notation is retained throughout all future chapters.

Notation	Meaning
i	Site indicator, $i = 1, \ldots, T$ (for treated sites) $i = 1, \ldots, C$ (for comparison sites)
T/C	Number of sites in treated/comparison pools
j	Covariate indicator, $j = 1, \ldots, P$
P	Number of covariates in the SPF
$x_{i,j}$	The value of covariate j at site i
$y_{i,\text{BEF}}/y_{i,\text{AFT}}$	Observed collision counts in the before/after period at site i
$y'_{i,\text{AFT}}$	Unobserved collision count at site i in the after period in absence of treatment
$\lambda_{i,\text{BEF}}$	Underlying collision rate at site i in the before period
$\lambda_{i,\text{AFT}}$	Underlying collision rate at site i in the after period
λ_i	Underlying collision rate at site i
$\rho_{i,\text{BEF}}$	Deviation from underlying rate due to chance at site i in the before period
$\rho_{i,\text{AFT}}$	Deviation from underlying rate due to chance at site i in the after period
ρ_i	RTM effect at site i
κ_i	Trend effect at site i
τ_i	Treatment effect at site i
μ_i	Fitted estimate from the SPF for site i
β_j	SPF regression coefficient for covariate j
γ	Overdispersion parameter of the Negative Binomial error structure of the SPF
m/M	Iteration count/total number of iterations in an MCMC algorithm

2.1 Motivation

As discussed in Chapter 1, road safety is a topic which is garnering global attention as an area on which action must be taken. This is reflected in financial commitments by governments who ringfence significant amounts of funding for projects and schemes to improve road safety in their country, thereby contributing to the UN target discussed in Section 1.1. It is clearly important therefore to ensure that these schemes, where implemented, are efficient and successful in their task of improving road safety by reducing road traffic collisions. The procedure for investigating and analysing this effectiveness is carried out via scheme evaluation studies, the aim of which is to discern and quantify the effect that implementing treatment had on the collision rate, in areas where the treatment had been implemented.

2.1.1 Regression To the Mean

Clearly treatment effect elicitation would be a trivial exercise if we were to know the number of collisions which would have taken place at each treated site had treatment not been applied, since the treatment effect would simply be the difference between this and the collision counts at the site after treatment,

$$\tau_i = y\prime_{i,\mathrm{AFT}} - y_{i,\mathrm{AFT}} \tag{2.1}$$

where $y_{i,\mathrm{AFT}}$ is the collision count at site i after treatment, and $y\prime_{i,\mathrm{AFT}}$ is the unobserved collision count in the same period at site i. In reality of course we cannot know $y\prime_{i,\mathrm{AFT}}$ and so must estimate it in order to elicit a treatment effect. We can do this by performing inference to elicit a vaue for the underlying treatment effect at site i, λ_i. The most commonly used method for scheme evaluation is a so-called before and after study, whereby the same locations are analysed before and after treatment has been applied, removing between subject variability from the analysis, and hence simplifying the analysis while improving its precision (analogous to a crossover trial in medical studies [Jones and Kenward, 2003]). Perhaps the most obvious method for determining the effect of a treatment at a given site, would be to simply take the difference in collision counts between before and after the treatment was applied. For instance at a site i, if we set λ_i to be the underlying collision rate and τ_i to be the effect of treatment, we obtain our estimate of the treatment effect $\hat{\tau}_i$, via

$$\hat{\tau}_i = y_{i,\mathrm{BEF}} - y_{i\mathrm{AFT}} \tag{2.2}$$

where $y_{i,\text{BEF}}$ and $y_{i,\text{AFT}}$ are observations (or perhaps means of several observations) taken before and after the treatment was implemented, respectively. And hence we have,

$$\begin{aligned}\mathbb{E}[\hat{\tau}_i] &= E[y_{i,\text{BEF}} - y_{i,\text{AFT}}] \\ &= \mathbb{E}[y_{i,\text{BEF}}] - \mathbb{E}[y_{i,\text{AFT}}] \\ &= \lambda_i - (\lambda_i - \tau_i) \\ &= \tau_i\end{aligned}$$

and so provided our assumptions regarding the expected values of the collision numbers in the before and after periods, $y_{i,\text{BEF}}$ and $y_{i,\text{AFT}}$, are accurate, we have an unbiased estimator of the treatment effect, τ_i.

We can then combine these estimates of treatment effect across all n treated sites to give an overall estimate of the treatment effect, τ,

$$\hat{\tau} = \sum_{i=1}^{n} \hat{\tau}_i.$$

While this approach intuitively makes sense initially, it is considered a rather naive method for evaluating treatment effects since it fails to account for the possibility of RTM and trend in these calculations. RTM is of particular concern in the context of scheme evaluation studies since often the methods by which sites are selected for treatment provide a textbook example of where RTM can cause problems, namely that treatments are implemented where it is deemed they are most needed, rather than randomly, a clear example of selection bias. Because of this bias, we are therefore highly prone to treatments being applied to sites where the the collision count in the before period, $y_{i,\text{BEF}}$ was an unnaturally extreme occurence, which would return to a more usual value without intervention by the RTM effect. In essence we have,

$$\begin{aligned}y_{i,\text{BEF}} &= \lambda_i + \rho_{i,\text{BEF}} \\ y_{i,\text{AFT}} &= \lambda_i + \rho_{i,\text{AFT}} + \tau_i\end{aligned}$$

where λ_i is the true collision rate (our proxy for the level of safety), and $\rho_{i,\text{BEF}}$ and $\rho_{i,\text{AFT}}$ are the random deviations from this rate due to chance. Hence Equation (2.2) becomes

$$\begin{aligned}\hat{\tau}_i &= y_{i,\text{AFT}} - y_{i,\text{BEF}} \\ &= (\lambda_i + \rho_{i,\text{AFT}} + \tau_i) - (\lambda_i + \rho_{i,\text{BEF}}) \\ &= (\rho_{i,\text{AFT}} - \rho_{i,\text{BEF}}) + \tau_i, \\ &= \rho_i + \tau_i. \qquad (2.3)\end{aligned}$$

where $\rho_i = \rho_{i,\text{AFT}} - \rho_{i,\text{BEF}}$ is the RTM effect at site i. If treatments were allocated randomly we would have

$$\mathbb{E}\left(\rho_{i,\text{BEF}}\right) = \mathbb{E}\left(\rho_{i,\text{AFT}}\right) = 0$$

meaning $\mathbb{E}\left(\rho_i\right) = 0$ and hence $\hat{\tau}_i$ would be an unbiased estimator of τ_i,

$$\begin{aligned}\mathbb{E}\left(\hat{\tau}_i\right) &= 0 + \tau_i,\\ &= \tau_i.\end{aligned}$$

However often due to a mixture of social and political pressures (potentially also combined with a lack of awareness of effects such as RTM) sites are selected for treatment non-randomly, with sites with large collision counts in the before period almost always the ones chosen for treatment. Because of this non-random selection we therefore induce a bias whereby

$$\mathbb{E}\left(\rho_{i,\text{BEF}}\right) > 0$$

and since sites are selected before the after period, and so the observations in the after period are not subject to selection bias, and $\rho_{i,\text{BEF}}$ and $\rho_{i,\text{AFT}}$ are considered independent, we retain $\mathbb{E}\left(\rho_{i,\text{AFT}}\right) = 0$ and hence

$$\mathbb{E}\left(\rho_i\right) < 0 \tag{2.4}$$

and hence $\hat{\tau}_i$ as defined in Equation (2.2) is no longer an unbiased estimator of τ_i. As discussed in the illustrative example in Section 1.2 these extreme deviations from the underlying rate (i.e. large values of ρ_i) are not uncommon, and so it is important to account for these effects in any before-and-after analysis. Clearly if we were able to estimate ρ_i accurately, we could adjust our estimates for this and regain an unbiased estimator for the treatment effect, i.e. if $\hat{\tau}_i = y_{i,\text{BEF}} - y_{i,\text{AFT}} - \hat{\rho}_i$

$$\begin{aligned}\mathbb{E}\left(\hat{\tau}_i\right) &= \mathbb{E}\left(y_{i,\text{BEF}} - y_{i,\text{AFT}} - \hat{\rho}_i\right),\\ &= \rho_i + \tau_i - \rho_i,\\ &= \tau_i\end{aligned}$$

and so the challenge of before-and-after studies becomes in essence to identify the amount of observed change from before to after is due to RTM, and remove this in order to obtain the true treatment effect. We note here that whilst we would expect a reduction in collision counts due to the RTM effect (Equation (2.4)) it is still possible to have an increase in collision counts due to RTM, meaning we would underestimate the treatment effect should RTM not be accounted for. Since $\mathbb{E}\left(\rho_{i,\text{AFT}}\right) = 0$ we estimate the RTM effect via

$$\begin{aligned}\mathbb{E}\left(\rho_i\right) &= \mathbb{E}\left(\rho_{i,\mathrm{BEF}} - \rho_{i,\mathrm{AFT}}\right), \\ &= \mathbb{E}\left(\rho_{i,\mathrm{BEF}}\right), \\ &= y_{i,\mathrm{BEF}} - \mathbb{E}\left(\lambda_i\right),\end{aligned}$$

i.e. the RTM effect is only present where selection bias has been present. As discussed briefly in Section 1.2, RTM is by no means restricted to road safety, and is in fact hugely prevalent in studies of Biology (where it was first discovered/documented) and medicine. Whereas often medical studies will overcome the issue of selection bias by testing treatments through randomised clinical trials, where identical groups of patients are given and not given a treatment in order to determine its effects, this is generally not true for road safety studies. As discussed in Section 1.2, road safety practitioners operate within restricted budgets, meaning it is often not feasible to potentially waste funds implementing treatments in areas where they are clearly not needed, likewise they are often under considerable political and ethical pressure to respond to a dangerous area on the road network should it become apparent. Due to this we do not have the identical treated/non-treated groups available in the world of medicine, and so must respond in other ways to address the issue of RTM.

2.1.2 Temporal Trend

A clear weakness of the model described in Equation (2.3), is that it assumes the only changes in collision total between before and after treatment, are due to the treatment effect τ_i and RTM effect, ρ_i, with the underlying collision rate λ_i remaining constant. There are a variety of factors which may cause the underlying collision rate to change over time, with road safety awareness campaigns, improved vehicle safety features etc potentially decreasing the collision rate, or possibly pavement deterioration, increasing traffic volumes etc potentially causing increases. It is important to account for this in our model, since Equation (2.3) currently ascribes any change not due to RTM, to be due to the effect of treatment, meaning any significant change which is actually due to trend, will lead to a bias in the estimate of treatment effect. To account for the potential for change due to trend, we therefore no longer assume a constant collision rate λ_i, and instead specify a separate collision rate corresponding to the time period after treatment, $\lambda_{i,\mathrm{AFT}}$. We define the trend effect, denoted κ_i, to be the expected difference in collision counts between the before and after periods in absence of RTM and treatment effects,

$$\begin{aligned}\kappa_i &= \mathbb{E}\left(y_{i,\mathrm{BEF}} | \rho_{i,\mathrm{BEF}}\right) - \mathbb{E}\left(y_{i,\mathrm{AFT}} | \rho_{i,\mathrm{AFT}}, \tau_i\right) \\ &= \lambda_i - \lambda_{i,\mathrm{AFT}},\end{aligned}$$

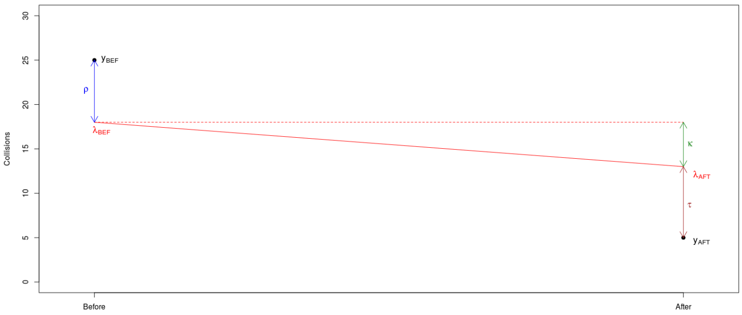


Figure 2.1: A toy example demonstrating the modelling structure for scheme evaluation studies. The change in collision counts (black) from before to after treatment is disaggregated into the RTM effect (blue line), trend effect (green line) and treatment effect (brown line), with the true underlying collision rate (red line) showing how these effects can be estimated.

hence,

$$\lambda_{i,\text{AFT}} = \lambda_i + \kappa_i$$

Hence the the expression in Equation (2.3) now becomes

$$\begin{aligned}\hat{\tau}_i &= y_{i,\text{AFT}} - y_{i,\text{BEF}}, \\ &= (\lambda_i + \kappa_i + \rho_{i,\text{AFT}} + \tau_i) - (\lambda_i + \rho_{i,\text{BEF}}) \\ &= \rho_i + \kappa_i + \tau_i\end{aligned}$$

and so in order to obtain an unbiased estimate of the treatment effect τ_i we must estimate both the RTM and trend effects. Given the restricted amount of data available in before-and-after studies it is not possible to distinguish between trend effects and treatment effects using treated site-specific data, and so we use data from comparison sites to obtain estimates of the trend effect. Here we shall assume a constant trend across all sites on the network $\kappa_1 = \kappa_2 = \cdots = \kappa_n = \kappa$ although improvements to this approach will be discussed in Section 4.5.

This structure can be seen visually in Figure 2.1 where a toy example has been constructed, demonstrating the change in collisions from before to after treatment deconstructed into RTM, trend and treatment effects.

2.2 Empirical Bayes and Full Bayes

In a before and after analysis we attempt to discern the true treatment effect of a road safety countermeasure, by removing any RTM and trend effects from the observed change in collision counts from before to after the treatment was implemented. Due to the fact we are often restrained to very limited data, potentially just a single observation before and after the treatment has been applied, we must appeal to other sources of data, in this case covariate information in the form of an SPF, to enable us to estimate the confounding effects. The most common way to do this is within a Bayesian framework, the most popular of which is the so-called Empirical Bayes (EB) method [Hauer, 1980] which is implemented as follows:

1. We assume our collision counts y_i follow a Poisson distribution conditional on rate parameter λ_i which takes a conjugate Gamma prior distribution (meaning unconditionally the collision counts will have a Negative Binomial distribution as shown in Section 1.4). Hence for a site i $(i = 1, \ldots, T)$ we have:

$$\begin{aligned} y_i|\lambda_i &\sim Po\left(\lambda_i\right), \\ \lambda_i|\mu_i &\sim Ga\left(\gamma, \frac{\gamma}{\mu_i}\right), \end{aligned} \tag{2.5}$$

   where γ is the reciprocal of the overdispersion parameter of the unconditional Negative Binomial distribution of the collision counts y_i. This choice of Gamma prior for λ_i is selected for conjugacy, in order to ensure an analytic posterior distribution, meaning a closed form expression for the posterior mean can be obtained. The specific parameterisation is chosen to allow the prior mean to be μ_i,

$$\begin{aligned} \mathbb{E}\left(\lambda_i|\mu_i\right) &= \frac{\gamma}{\frac{\gamma}{\mu_i}}, \\ &= \mu_i, \end{aligned}$$

   the number of collisions estimated by applying our fitted SPF, and the prior variance to be proportional to the relative overdisperion of the model estimate,

$$\begin{aligned} Var\left(\lambda_i|\mu_i\right) &= \frac{\gamma}{\left(\frac{\gamma}{\mu_i}\right)^2}, \\ &= \frac{\mu_i^2}{\gamma} \end{aligned}$$

   meaning our prior uncertainty is accurately represented by including the dispersion in the model fit, i.e. if there is a large degree of dispersion in the model, γ will be small and hence our prior uncertainty for λ_i will be large, and vice-versa.

2. We assign our SPF to be a Negative Binomial GLM, which we fit using maximum likelihood estimation from a set of exchangeable comparison sites with P covariates. From this we obtain our SPF coefficient vector $\boldsymbol{\beta}$ and an estimate of the inverse of the

Negative Binomial overdispersion parameter γ. We then apply the fitted coefficients to our treated sites to obtain an estimate of μ_i, the prior mean of λ_i:

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i,1} + ... + \beta_P x_{i,P}).$$

3. Combining the Gamma prior distribution with the Poisson likelihood gives a conjugate Gamma posterior for λ_i:

$$\lambda_i | Y_i = y_i \sim Ga\left(\gamma + y_i, \frac{\gamma}{\mu_i} + 1\right).$$

4. We take the mean of this posterior to be our point estimate of the untreated collision rate in the before period,

$$\begin{aligned} \hat{\lambda}_{i,\text{BEF}} &= E(\lambda_i | y_{i,\text{BEF}}, \boldsymbol{x}) && (2.6)\\ &= \frac{\gamma}{\gamma + \mu_i}\mu_i + \frac{\mu_i}{\gamma + \mu_i} y_{i,\text{BEF}} && (2.7)\\ &= w_i \mu_i + (1 - w_i) y_{i,\text{BEF}}. && (2.8) \end{aligned}$$

   And so we can observe that our estimate for the untreated collision rate at site i, is simply a weighted average of the observed collision count y_i, and the fitted collision rate from the SPF μ_i. We note also how this weighting w_i depends on the dispersion in the SPF, whereby if there is a high degree of dispersion, and hence γ is small, the posterior mean will give more weight to the observed collision count y_i and less to the SPF estimate, and vice-versa.

5. We can account for the effect of temporal trend by multiplying $E(\lambda_i | y_{i,\text{BEF}})$ (which informally can be thought of as our collision count in the before period once the effect of RTM has been removed), by a factor, κ_i. Here κ_i represents the percentage change in collisions expected due to trend (e.g. if it is believed there would be a 10% reduction in collisions due to trend we would take $\kappa_i = 0.9$), and should be chosen from a practitioner's expert prior beliefs and/or suitable reference data to inform trends. Often it will be difficult for experts to elicit site specific trend effects and so a network wide trend effect $\kappa_1 = \kappa_2 = \cdots = \kappa_n = \kappa$ will be assumed, improvements to this approach are outlined in Section 4.5. Should there not be suitable prior information, or if there is no reason to believe there will be a significant trend, we should set $\kappa = 1$. The result is therefore an estimator for the underlying collision rate in the period after treatment,

$$\hat{\lambda}_{i,\text{AFT}} = \kappa_i \mathbb{E}(\lambda_i | y_{i,\text{BEF}}, \boldsymbol{x})$$

An equivalent additive method for incorporating trend can be obtained where κ_i would be the expected change in collision counts due to trend at site i,

$$\hat{\lambda}_{i,\text{AFT}} = \mathbb{E}(\lambda_i | y_{i,\text{BEF}}, \boldsymbol{x}) + \kappa_i.$$

However, as discussed previously practitioners often prefer to specify trends on a network wide scale rather than a site-specific one, and hence the multiplicative framework is more appealing since this avoids any issues of specifying negative rates where the expected reduction due to trend is greater than the observed count in the before period.

6. We can therefore obtain estimates of ρ_i, κ_i and τ_i (the RTM, trend and treatment effects),

$$\begin{aligned} \hat{\rho}_i &= \hat{\lambda}_{i,\text{BEF}} - y_{i,\text{BEF}} \\ &= \mathbb{E}(\lambda_i | y_{i,\text{BEF}}, \boldsymbol{x}) - y_{i,\text{BEF}}, \\ \hat{\kappa}_i &= \hat{\lambda}_{i,\text{AFT}} - \hat{\lambda}_{i,\text{BEF}} \\ &= \kappa \mathbb{E}(\lambda_i | y_{i,\text{BEF}}, \boldsymbol{x}) - \mathbb{E}(\lambda_i | y_{i,\text{BEF}}, \boldsymbol{x}), \\ \hat{\tau}_i &= y_{i,\text{AFT}} - \hat{\lambda}_{i,\text{AFT}} \\ &= y_{i,\text{AFT}} - \kappa \mathbb{E}(\lambda_i | y_{i,\text{BEF}}, \boldsymbol{x}). \end{aligned} \tag{2.9}$$

The EB method has many desirable qualities, its conjugate structure allows for a closed form, formulaic estimator of the treatment effect, meaning minimal computing power is required, and point estimates (those of most use to practitioners) can be readily obtained. However there are also many restrictions which make EB suboptimal, namely the overoptimistic estimates of the uncertainty of our treatment effect (caused by its failure to retain uncertainty of the SPF estimate μ), and its modelling rigidity, where only the Poisson-Gamma structure can be used, even though alternative prior distributions may be superior [Fawcett and Thorpe, 2013].

We can overcome these restrictions by rejecting the Empirical Bayes framework in favour of a Full Bayes (FB) methodology [Yanmaz-Tuzel and Ozbay, 2010], [Kitali and Sando, 2017a] [Lan et al., 2009]. Here, as opposed to the formulaic method by which EB determines an estimate of the treatment effect thanks to an assumed conjugate structure, we instead compute everything via MCMC, thereby removing the constraint of conjugacy. Should we wish to retain a Poisson-Gamma structure our model structure may resemble:

$$\begin{aligned} y_i|\lambda_i &\sim Po(\lambda_i), \\ \lambda_i|\mu_i &\sim Ga\left(\gamma, \frac{\gamma}{\mu_i}\right), \\ \mu_i &= \exp\left(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_P x_{P,i}\right)), \\ \beta_j &\sim N(\mu_\beta, \sigma^2_{\beta_j}), \qquad j = 0, \ldots, P, \\ \log(\gamma) &\sim N\left(\mu_\gamma, \sigma^2_\gamma\right). \end{aligned} \tag{2.10}$$

where prior knowledge can be imparted into the model by choices of $\boldsymbol{\mu_\beta}$, μ_γ, $\boldsymbol{\sigma_{\beta_j}}$, σ_γ. Prior ignorance could be imparted by choosing for example, $\mu_{\beta_1} = \mu_{\beta_2} = \cdots = \mu_{\beta_P} = \mu_\gamma = 0$ and $\sigma_{\beta_1} = \sigma_{\beta_2} = \cdots = \sigma_{\beta_P} = \sigma_\gamma = 100$. We note here the

This model can be fitted in a Bayesian setting using a Metropolis-Hastings algorithm as described in Section 1.5, with the MCMC algorithm then being

1. Initialise the parameter vector $\boldsymbol{\beta}$ at its initial state, $\boldsymbol{\beta^{(0)}} = \left(\beta_0^{(0)}, \beta_1^{(0)}, \ldots, \beta_P^{(0)}\right)$. Hence each element i of $\boldsymbol{\lambda}$ is initialised at $\lambda_i^{(0)} = \exp\left(\beta_0^{(0)} + \beta_1^{(0)} x_{i,1} + \ldots + \beta^{(0)} x_{i,P}\right)$. Initialise counter at $m = 1$.

2. For each element β_j in $\boldsymbol{\beta}$,

   - Sample a proposal value $\beta_j^* \sim q(\cdot|\cdot)$
   - Hence for each comparison site $i = 1, \ldots, C$ obtain a proposal for μ_i,

     $$\mu_i^* = \exp\left(\beta_0^{(m)} + \beta_1^{(m)} x_{i,1} + \ldots + \beta_{j-1}^{(m)} x_{i,j-1} + \beta_j^* x_{i,j} + \beta_{j+1}^{(m-1)} + \ldots + \beta_P^{(m-1)}\right),$$

     sample a proposal for λ_i

     $$\lambda_i^* \sim Ga\left(\gamma^{(m-1)}, \frac{\gamma^{(m-1)}}{\mu_i^*}\right)$$

     calculate the site's likelihood contribution L_i,

     $$L_i\left(y_i|\lambda_i^*\right) = \frac{\lambda_i^{*y_i}}{y_i!} e^{-\lambda_i^*}$$

     and hence the overall proposed likelihood

     $$L\left(\boldsymbol{y}|\boldsymbol{\lambda^*}\right) = \prod_{i=1}^{C} L_i\left(y_i|\lambda_i^*\right)$$

   - Calculate the acceptance probability $v = \min\left(1, \frac{\pi\left(\beta_j^*\right) L(\boldsymbol{y}|\boldsymbol{\lambda^*}) q\left(\beta_j^{(m-1)}|\cdot\right)}{\pi\left(\beta_j^{(m-1)}\right) L\left(\boldsymbol{y}|\boldsymbol{\lambda^{(m-1)}}\right) q(\beta^*|\cdot)}\right)$

3. Set $\beta_j^{(m)} = \beta_j^*$ with probability υ else set $\beta_j^{(m)} = \beta_j^{(m-1)}$

4. Sample a proposal value for γ, $\gamma^* \sim q(\cdot|\cdot)$

5. For each comparison site $i = 1, \ldots, C$, obtain a fitted value for $\mu_i^{(m)}$

$$\mu_i^{(m)} = \exp\left(\beta_0^{(m)} + \beta_1^{(m)} x_{i,1} + \ldots + \beta_P^{(m)} x_{i,P}\right)$$

   and sample a proposal value for λ_i,

$$\lambda_i^* \sim Ga\left(\gamma^*, \frac{\gamma^*}{\mu_i^{(m)}}\right)$$

   and hence the site's likelihood contribution L_i,

$$L_i\left(y_i|\lambda_i^*\right) = \frac{\lambda_i^{*y_i}}{y_i!} e^{-\lambda_i^*}$$

   and hence the overall proposed likelihood

$$L\left(\boldsymbol{y}|\boldsymbol{\lambda}^*\right) = \prod_{i=1}^{C} L_i\left(y_i|\lambda_i^*\right)$$

6. Calculate the acceptance probability $\upsilon = \min\left(1, \frac{\pi(\gamma^*)L^{\mathrm{C}}(\boldsymbol{y}|\boldsymbol{\lambda}^*)q\left(\gamma^{(m-1)}|\cdot\right)}{\pi\left(\gamma^{(m-1)}\right)L\left(\boldsymbol{y}|\boldsymbol{\lambda}^{(\boldsymbol{m-1})}\right)q(\gamma^*|\cdot)}\right)$

7. Set $\gamma^{(m)} = \gamma^*$ with probability υ else set $\gamma^{(m)} = \gamma^{(m-1)}$

8. For each treated site $i = C+1, \ldots C+n$

   - Obtain the fitted value for $\mu_i^{(m)}$,

$$\mu_i^{(m)} = \exp\left(\beta_0^{(m)} + \beta_1^{(m)} x_{i,1} + \ldots + \beta_P^{(m)} x_{i,P}\right)$$

   - Generate a proposal value for λ_i,

$$\lambda_i^* \sim Ga\left(\gamma^{(m)}, \frac{\gamma^{(m)}}{\mu_i^{(m)}}\right)$$

   - Calculate the site's likelihood contribution L_i,

$$L_i\left(y_i|\lambda_i^*\right) = \frac{\lambda_i^{*y_i}}{y_i!} e^{-\lambda_i^*}$$

   - Hence obtain the overall proposed likelihood

$$L\left(\boldsymbol{y}|\boldsymbol{\lambda}^*\right) = \prod_{i=C+1}^{n} L_i\left(y_i|\lambda_i^*\right)$$

- Calculate the acceptance probability $\upsilon = \min\left(1, \frac{\pi(\lambda^*)L(\boldsymbol{y}|\boldsymbol{\lambda^*})q\left(\lambda^{(m-1)}|\cdot\right)}{\pi\left(\lambda^{(m-1)}\right)L\left(\boldsymbol{y}|\boldsymbol{\lambda^{(m-1)}}\right)q(\lambda^*|\cdot)}\right)$
- Set $\lambda_i^{(m)} = \lambda^*$ with probability υ else set $\lambda_i^{(m)} = \lambda_i^{(m-1)}$

9. If $m = M$ stop, else set $m = m + 1$ and go to step 2.

After discarding burn-in and thinning if necessary to remove significant autocorrelation in the chain, the resulting vector $\boldsymbol{\lambda_i}$ forms the posterior distribution for λ_i, $\pi\left(\lambda_i|y_{i,\text{BEF}}, \boldsymbol{x}\right)$. We can therefore obtain a distribution of estimated treatment effects by taking the difference between the observed collision count after treatment and the posterior distribution for λ_i,

$$f\left(\tau_i\right) = y_{i,\text{AFT}} - \pi\left(\lambda_i|y_{i,\text{BEF}}, \boldsymbol{x}\right), \tag{2.11}$$

with the point estimate of treatment effect often taken to be the mean of this distribution,

$$\begin{aligned}\hat{\tau}_i &= \mathbb{E}\left(\tau_i\right)\\ &= y_{i,\text{AFT}} - \mathbb{E}\left(\lambda_i|y_{i,\text{BEF}}, \boldsymbol{x}\right).\end{aligned}$$

It is the propogation of uncertainty in the comparison site SPF obtained in step 2 of the above algorithm, which is retained by using iterative samples of μ_i when sampling the posterior distribution of λ_i in step 8 as opposed to fixed values of the SPF as in the EB approach. We note that in the above algorithm, due to data at sites being treated as conditionally independent of data at other sites given model parameters, operations such as calculating site-specific likelihood contributions and carrying out updates of λ_i at the treated sites can be done in parallel. Often these calculations are relatively quick however, and so parallelisation should only be implemented if the number of sites is very large, so as to ensure improved computational efficiency in the face of overhead costs incurred from setting up parallel clusters etc.

Immediately it becomes clear that we now have much greater modelling flexibility than with the EB case, as the Gamma prior distribution can trivially be replaced by any other suitable prior distribution, common replacements being the Lognormal and Weibull distributions. There are additional options for the prior distribution for our regression coefficient $\boldsymbol{\beta}$ beyond the independent Normal distributions assigned here, one obvious alternative being to investigate correlation between regression covariates by fitting a single Multivariate Normal distribution, or even a Data Augmentation Prior to $\boldsymbol{\beta}$. Further to this, whereas in the EB case we simply fit a negative binomial GLM as our SPF to gain an estimate of μ using maximum likelihood estimation, now we fit our SPF in a fully Bayesian sense, i.e. with distributions for the regression coefficients, and retain this uncertainty in our distribution of λ_i. Because of this we therefore have a much more accurate representation of our uncertainty regarding the treatment effect τ_i. Previous reluctance to embrace

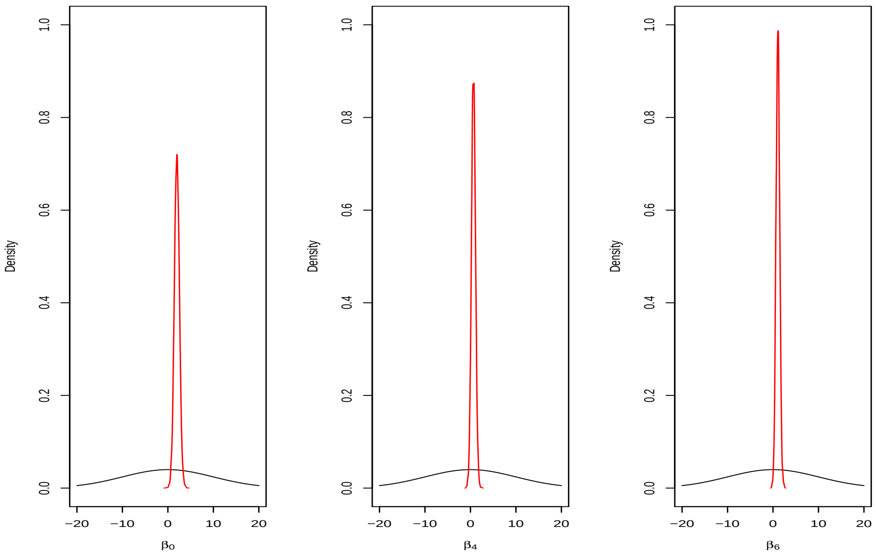


Figure 2.2: Plots showing the prior distributions (black) for β_0, β_4 and β_6 with posterior densities (red) overlaid.

an FB methodology in industry can be explained by the now enhanced level of computing and statistical ability required to implement these methods, which may not be accessible to some or even most road safety practitioners. Fortunately modern advances in computing coupled with software applications to tackle the statistical aspects of the technique have greatly reduced the significance of these drawbacks, and we are seeing FB become more and more commonplace ([El-Basyouny and Sayed, 2012], [Heydari et al., 2014]).

We demonstrate the differences in uncertainty estimation between the EB and FB approaches now using the Northumbria before-and-after dataset introduced in Section 1.6. Here we have 67 comparison sites from which to build the SPF, and 56 treated sites from which we wish to discern a treatment effect. We carry out a standard EB analysis against a Poisson-Gamma FB analysis with model described in (2.10) with vague priors on the regression coefficients,

$$\beta_j \sim N\left(0, 10^2\right), \qquad j = 1, \ldots, P$$

as in [Fawcett and Thorpe, 2013]. For simplicity (with respect to parameter tuning) we carry out the FB analysis in `rjags`, the model was run for 100,000 iterations and thinned by 10 to remove autocorrelation. Figure 2.2 shows prior and posterior densities for β_0, β_4 and β_6 highlighting the vague priors used, and the informativeness of the data demonstrated by a much more focussed posterior density. Posterior summaries for the SPF parameters are given in Table 2.1, with the uncertainties attached to these parameter estimates illustrating the difference in posterior uncertainties between the EB and FB

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	γ
Mean	1.89	-0.04	-0.01	0.46	0.69	0.87	1.09	2.18
St. Dev.	0.54	0.02	0.01	0.22	0.44	0.45	0.40	0.72

Table 2.1: Table showing posterior summaries for the SPF parameters

	Site 4	Site 20	Site 33	Site 56
Before	3	1	28	7
After	0	2	16	1

Method	$\lambda_4\|y_4$	$\tau_4\|\lambda_4$	$\lambda_{20}\|y_{20}$	$\tau_{20}\|\lambda_{20}$	$\lambda_{33}\|y_{33}$	$\tau_{33}\|\lambda_{33}$	$\lambda_{56}\|y_{56}$	$\tau_{56}\|\lambda_{56}$
EB	2.92	-2.92	1.79	0.21	19.85	-3.85	5.70	-4.70
	(0.368)		(0.368)		(0.370)		(0.370)	
FB	3.08	-3.08	1.70	0.30	19.73	-3.73	5.49	-4.49
	(1.415)		(1.000)		(4.418)		(1.920)	

Table 2.2: Table showing posterior means for λ and treatment effect $\tau = y_{\text{AFT}} - \lambda$ along with posterior standard deviation (in brackets) for sites in the Northumbria dataset

approaches.

We then compare posterior summaries for λ_i from within the EB and FB approaches. Within the EB framework we adopt a conjugate Gamma prior and so can obtain an analytic Gamma posterior distribution,

$$\lambda_i|y_i \sim Ga\left(\gamma + n\bar{y}, \frac{\gamma}{\mu_i} + n\right)$$

and hence we can obtain an analytic posterior standard deviation,

$$\begin{aligned} \text{Var}\left(\lambda_i|y_i\right) &= \frac{\gamma + n\bar{y}}{\left(\frac{\gamma}{\mu_i} + n\right)^2}, \\ \text{SD}\left(\lambda_i|y_i\right) &= \sqrt{\text{Var}\left(\lambda_i|y_i\right)}, \\ &= \frac{\sqrt{\gamma + n\bar{y}}}{\frac{\gamma}{\mu_i} + n} \end{aligned}$$

and so we can form comparisons between these and the posterior standard deviations obtained from the FB analysis. Results of this for selected sites from the Northumbria dataset are given in Table 2.2.

From Table 2.2 we can see that while the posterior means are similar between the EB and FB approaches, the posterior standard deviations are significantly higher for the FB

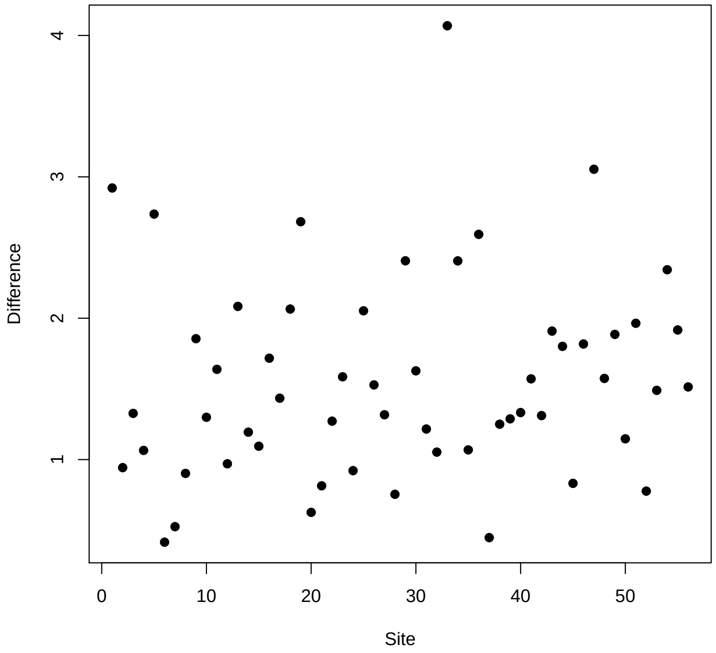


Figure 2.3: A plot of differences in posterior standard deviations of λ_i using the FB and EB approaches for sites in the Northumbria dataset.

approach, demonstrating the effect of propogating the uncertainty from the SPF through the analysis to be reflected in posterior uncertainty regarding λ and hence τ. Failure to do this leads to misleadingly overoptimistic estimates of posterior uncertainty, which can cause misleading conclusions to be drawn regarding the significance of the findings of the case study, which can directly affect future policy decisions with regards to scheme effectiveness. We demonstrate this for the entire Northumbria dataset in Figure 2.3 where we plot the difference in posterior standard deviations for λ_i at each site i between the FB and EB approaches.

Structurally the posterior standard deviations obtained from an FB approach must be higher than the corresponding posterior standard deviations from an EB approach. Figure 2.3 shows the difference in posterior standard deviation between the FB and EB approaches for the before-and-after analysis of the Northumbria dataset for each site, with a mean difference in standard deviation of 1.561. Despite FB methods being accepted in the literature for a while, with [Schlüter et al., 1997] providing support for a hierarchical model to replace EB in the 1990s, EB remains in common usage currently, with [Wang et al., 2017], [Høye, 2015], [Park and Abdel-Aty, 2015] providing examples of studies carried out in the last several years still relying on an EB methodology for in-

Prior	Gamma	Lognormal	Weibull	Beta-prime	Inv-Gamma
$\sum \mathbb{E}(\lambda_i \vert \boldsymbol{y})$	318	339	317	333	339
	(290,368)	(248,400)	(296,371)	(290,378)	(298,381)

Table 2.3: Expected numbers of total casualties (with 95% credible intervals) across the 56 sites of the Northumbria dataset when using different prior distributions for λ_i

ference. In light of this therefore, demonstrations of the improvements provided by FB relative to EB remain extremely pertinent to current discussion.

2.3 Prior Comparison

As discussed in Section 2.2 one of the main advantages of using an FB modelling framework over EB is the ability to choose non-conjugate prior distributions for λ_i in the Bayesian model. We now demonstrate this by fitting FB models with various prior distributions to the Northumbria dataset, in order to observe the sensitivity in estimates of λ_i to the choice of prior, and to carry out model comparison to determine the best fitting prior distribution to the Northumbria dataset as in [Hewett et al., 2019]. The rationale behind this is since our prior information is fully specified in the prior mean and variance, we still have freedom to specify exactly which distribution (with the specified mean and variance) our prior beliefs should follow. Comparing the fit of several candidate prior distributions allows us to ascertain which resulting posterior distribution provides the best fit to our data, while still including our specified prior knowledge. The prior distributions for λ_i we consider are the Gamma distribution, the Lognormal distribution, the Weibull distribution, the Beta-prime distribution and the Inverse-Gamma distribution ([Forbes et al., 2011]). We recall that in the EB case we had the prior mean and variance for λ_i to be,

$$\mathbb{E}(\lambda_i) = \mu_i \qquad\qquad \mathrm{Var}(\lambda_i) = \frac{\mu_i^2}{\gamma}$$

where μ_i is the fitted collision rate estmate from the SPF, and γ is the associated overdispersion parameter. For comparative purposes we retain these prior means and variances for our alternative prior distributions. Fitting these various prior distributions gives results shown in Table 2.3.

From Table 2.3 we observe a clear sensitivity to the choice of prior distribution when carrying out a scheme evaluation analysis on the Northumbria dataset. The observed total number of casualties across the 56 sites in the after period was 298, meaning that when using a Gamma prior distribution for λ_i we have a mean treatment effect of -20, whereas if we were to use a Lognormal or an Inverse-Gamma prior, the mean estimate of treatment effect would more than double to -41.

When we have several possible competing models to describe a dataset, as in this section where we have various possible prior distributions for λ_i, it is natural to wish to ascertain which model fits the data best in order to carry out model selection. A popular choice of goodness of fit test to determine which model describes the data best is the deviance information criterion (DIC) [Spiegelhalter et al., 2002]. For a given set of observations y_i $(i = 1, \ldots, n)$ with model parameters $\boldsymbol{\theta}$ we define the model deviance $D(\boldsymbol{\theta})$ to be,

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

hence for a Poisson model such as that in Section 2.2 we have the deviance to be,

$$D\left(\lambda\right) = -2 \times \left[n\bar{y}\log(\lambda) - n\lambda - \sum_{i=1}^{n}\log\left(y_i!\right)\right]$$

In Bayesian models where we sample from the posterior distribution(s) of our parameter(s) many times, we can calculate the deviance at each iteration m, and thereby obtain a distribution of posterior deviances,

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

with the mean of the distribution, denoted $\overline{D(\theta)}$, commonly taken as the measure of goodness of fit. We wish for our fitted model to have a high likelihood/log-likelihood, and so models which produce lower posterior mean deviances should be favoured. The issue with this when considering models with varying parameters, particularly nested models where the parameter vector of one model being compared is a subset of the parameter vector of another model being compared, is that more complex models will always fit the data better, and so have smaller posterior mean deviance. Since the aim in model selection is normally to find the most parsimonious model, the model which provides the greatest fit with the fewest parameters, we should include a component in our test criterion which penalises over-parameterised models. In order to do this we calculate an additional term, the effective number of parameters in the model, denoted p_D, defined as,

$$p_D = \overline{D(\theta)} - D(\bar{\theta}).$$

which increases as the number of parameters increases, thereby penalising overfitted models. We then define the deviance information criterion (DIC) to be

$$\text{DIC} = \overline{D(\theta)} + p_D$$

with lower values of DIC indicating a more parsimonious model.

Model	Gamma	Lognormal	Weibull	Beta-prime	Inv-Gamma
DIC	663.3	787.2	645.6	754.4	773.5

Table 2.4: Table showing the DIC values from using different prior distributions in an FB analysis of the Northumbria data

Calculating the DIC for the models fitted in Section 2.3 gives the results shown in Table 2.4 where we observe the Lognormal, Beta-prime and Inverse-Gamma prior distributions perform relatively poorly when modelling the Northumbria data, whilst the Weibull performed the best judging by its DIC score. This provides further weight to the argument for recommending an FB approach over an EB approach since here the flexibility of the FB approach allowed us to use a prior which modelled the data better than the Gamma prior enforced by the EB methodology.

There is an argument against using goodness-of-fit tests to summarise model quality in this way however. Since, as is the entire motivation for employing a Bayesian modelling structure as opposed to a naive before-and-after method, we suspect the possibility that elements of the observed data $\boldsymbol{y}$ will be anomalous, a model which fits to $\boldsymbol{y}$ well, may not necessarily best represent the collision rates at the treated sites. Of course in this, like many other modelling regards, we are heavily restricted by the small sample size at each location.

2.4 SPF-Free Approaches

While the EB and FB approaches mentioned in Section 2.2 are the most widely used methods in the literature for carrying out scheme evaluation studies, they are by no means the only methods used. There are several reasons why practitioners may be reluctant, or unable, to implement these approaches, including,

- Statistical complexity. Bayesian frameworks are complex, high-level statistical techniques, meaning many practitioners who do not have a background in statistics may be unable to make use of them. While the closed form, formulaic solution provided by the EB method makes it far more accessible, some practitioners are still reluctant to base decisions upon techniques which they are not fully familiar with the reasoning behind (which remains a problem since EB still makes use of a Bayesian hierarchical model). FB provides yet further difficulties to this, with its reliance on MCMC methods not only requiring more technical skill in developing an MCMC algorithm, but also computational cost and time in developing and running the algorithm for each analysis.

- Data requirements. Implementing EB and FB methods require the user to provide datasets which, while not particularly extensive, can prove troublesome for networks where not much data has been, or feasibly can be, collected. In particular the collection of usually highly important covariates such as average speed and AADT normally require specialist equipment to be set up in order to monitor these values, which may not be possible, and where it is still places a significant financial and time-consuming constraint on the practitioner. As discussed in Chapter 1, there is the possibility of deploying a traffic model to attempt to estimate these important parameters, however subscribing to a service to provide this modelling can also be expensive and so not feasible for some smaller local authorities, with unsophisticated modelling attempts likely to be fraught with error and so lead to erroneous results. Further to this the previously mentioned requirement of ensuring similarity between the sites used to build the SPF and the treated sites being monitored, mean EB and FB can quickly become unusable to any real degree in areas where data is scarce.
- Lack of intuitiveness regarding modelling assumptions. While the concept of comparing an observed collision count at a site, and comparing it with other similar sites to see how usual/expected it is makes sense on an intuitive level, the need to do so via a complex Bayesian framework has less natural intuitiveness. Hence practitioners may be reluctant to employ techniques which use formulae and algorithms, when it is difficult to discern exactly why such formulae should be used.

Due to these reasons (among others) there are a significant proportion of practitioners not making use of the EB and FB techniques, despite these methods being heavily championed in the literature. We therefore explore the relative merits of several of the more popular alternative methods in order to compare their effectiveness relative to the literature "gold standard" approaches of EB and FB.

2.4.1 Four Time Period Method

A perfect counterweight to the potentially technically strenuous Bayesian methods is the four time period method (FTP) developed by Dave Finney and Richard Allsopp. This approach provides a highly simple method for scheme evaluation studies by determining RTM and treatment effects. This approach solely requires collision counts at the treated site in question, although it does make use of several years of these. The approach works by splitting the observations at the treated site into four sections,

1. Before period (BP): These are the observations taken before the site was considered a potential candidate for treatment, and so will not have been used in any decision

making process regarding the decision to treat. There should be no treatment in place during this time period.

2. Site selection period (SSP): These are the observations taken when the site was considered a potential candidate for treatment, and so it should contain exactly each observation used in the decision making process.

3. After site selection but before installation of the camera (ASBiC) (this technique was developed for the analysis of speed camera effectiveness but can be equally applied to other road safety treatments): Since treatments often cannot be implemented immediately upon the decision to treat, there may be some observations taken at a treated site before it becomes treated, but will have no influence over the decision to apply treatment.

4. After period (AP): The observations taken after the treatment has been installed at the site.

The FTP approach suggests that the only time period during which any RTM effect can be present, is during the SSP, when the observations leading to the decision to treat are made. Therefore by discounting the SSP observations entirely and combining the BP and ASBiC into a single before period, and comparing with the AP we can determine the true treatment effect, i.e.

$$\begin{aligned}\mathrm{RTM}_j &= \frac{1}{n_{BP} + n_{ASBiC}} \left(y_{BP} + y_{ASBiC}\right) - \frac{1}{n_{SSP}} y_{SSP} \\ T_j &= \frac{1}{n_{BP} + n_{ASBiC}} \left(y_{BP} + y_{ASBiC}\right) - \frac{1}{n_{AP}} y_{AP}\end{aligned}$$

where n_{BP}, n_{ASBiC}, n_{SSP}, n_{AP}, and y_{BP}, y_{ASBiC}, y_{SSP}, y_{AP}, are the respective numbers of observations and collision counts for each of the respective time periods.

The appeals of the FTP approach are immediately obvious,

- Intuitive methodology. The logic underpinning the approach is straightforward and accessible to all practitioners. There is no reliance on high complexity statistics requiring an element of trust from anyone not fully versed in the subtleties of Bayesian statistics.

- Calculation speed. There is no requirement to run computationally intensive algorithms in order to obtain output. These formulae are more straightforward than those involved in EB calculations and can be done immediately with very little effort involved.

- Minimal data requirements. There is no need for (potentially difficult and expensive) covariate data collection with this approach, the only data required are collision counts. Furthermore there is no requirement for data collection at any site other than those being analysed, making the approach both efficient, and removing the need to ensure the similarity of a comparison pool of sites.

It is advantages such as these that have meant that whilst the FTP technique isn't widely reported in literature it has proved increasingly popular among practitioners who want quick, simple methodologies and are perhaps sceptical of any advantages provided by arduous statistical approaches. Unfortunately this method is not without its (quite considerable) drawbacks:

- Overestimation of RTM. The assumption that any reduction in the SSP is solely due to RTM is a very simplistic assumption and may not be correct, even if there is some RTM effect present in the change from SSP to AP, it does not mean that the entire change should be discounted and accredited to RTM. Such an assumption can therefore to be seen to be unfairly harsh on the effect of treatment.

- Restriction on blip location. While, as discussed, the assumption that all change in the SSP is due to RTM can be seen as being overly harsh, it is perhaps too lenient on other time periods where it is assumed that there is no RTM effect (i.e. no blips) present at all. This can be problematic, particularly if the BP or AP values are heavily influenced by blips since we base our treatment effect estimates almost entirely on these (with the ASBiC period usually very short, if any observations exist for it at all).

- No immediate method for prediction. Whilst the fact the technique is completely non-parametric boosts the simplicity and speed of the calculation, it does restrict the inference and thereby the information that can be gleamed from it. One such aspect the estimation and extrapolation of the true underlying of level of safety at the site (in the EB/FB case this would be given by the λ parameter) which can then be extrapolated to predict future collision counts at the site.

- No accounting for trend. As discussed in Section 2.2, it is not only the RTM effect which should be accounted for when determining the treatment effect, there is also the potential for temporal trend to affect the change from before to after a treatment is implemented, an effect which is particularly prevalent when considering data observed over many consecutive observations. Failing to account for this can lead to biases in the estimates of treatment effect, even if the strong assumptions regarding any RTM effects are correct.

- Places rigid boundaries on which time period an observation must fall. The underlying mechanism for the FTP method is based around the idea of splitting a sequence of observations at a site into four distinct time periods, based on the actions by the local authority at that time. However given that decisions are rarely made so rigidly over such a small time-frame (both in terms of the time taken to decide where to allocate treatment, and the window of data used to base that decision on), and so it can be difficult to implement this method by apportioning each observation to one of the four time periods, in an accurate and meaningful way.

- No method for estimating uncertainty. Again a drawback of the simplistic, non-parametric and quick nature of the analysis, there is no method for obtaining any measure of uncertainty regarding the estimate of treatment effect obtained from this method, with merely a point estimate of the estimated treatment effect being produced. Clearly this is over optimistic as it is tantamount to an analysis claiming 100% certainty in an estimate, which clearly cannot be the true.

2.4.2 Time Series Models

Time series models are occasionally used within the context of road safety analysis, where detailed analyses of changing trends are required ([Park et al., 2017a], [Quddus, 2008b]). In contrast to the models discussed in this thesis, time series models require no external data outside of the individual (in this case, site) being analysed, and hence do not require any form of SPF to be built. This is particularly useful when comparison data is difficult to obtain, of poor quality, or not suitably similar to the treated dataset (see Chapter 3), however the contrast to this is by definition, time series models require a significant amount of longitudinal data at each location in order to for these models to be fitted. Obtaining good quality longitudinal data for a sufficient Time series models are usually employed when we have a series of repeated observations in time, which we believe to be dependent. These observations can be either discrete or continuous, occur at regular or irregular intervals, and can be observations in any number of dimensions. Clearly in the case of collision counts, it is reasonable to assume the collision rate in year t is dependent on the rates in years $t-1$, $t-2$ etc, and so a univariate time series model may be suitable to model road safety data.

A common issue when considering time series models is that of stationarity. A temporal sequence of repeated observations $y_1, y_2, \ldots, y_n$ can be considered a single joint observation from $f(\boldsymbol{Y})$ where $f(\cdot)$ is a multivariate distribution. In order to implement time series models we usually consider these observations to be non-independent, and so the covariance matrix, Σ, of $f(\cdot)$ will be non-diagonal. For a strictly stationary time series, we require the joint distribution of observations $y_i, \dots, y_j$ to be the same as that of $y_{i+c}, \dots, y_{j+c}$ for all i, j and c, that is that we require the joint density $f(\boldsymbol{Y})$ to be constant with respect to time. Oftentimes this condition is considered too restrictive, and a more relaxed condition, known as weak stationarity, is sufficient for most time series models. For weak stationarity, rather than the full joint distribution needing to be constant with respect to time, we simply require the mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)$ and covariance matrix Σ to be constant with respect to time. In many applications, $f(\cdot)$ is chosen to be a Gaussian distribution, which since it is uniquely defined my its mean vector and covariance matrix, has the property the conditions for weak and strict stationarity are the same [Chatfield, 2016].

There are two main options for countermeasure effect estimation using time series modelling: we can either train a time series model on data before treatment, in order to estimate the collision count if no treatment had been applied, $y_{\hat{\text{AFT}}}$, and hence estimate the treatment effect as

$$\hat{\tau} = y_{\text{AFT}} - y_{\hat{\text{AFT}}};$$

or alternatively incorporate the existence of treatment as an extraneous variable which we include in our time series model (through for example ARMAX modelling), allowing the treatment effect to be explicitly included in the model parametrically. A relative drawback of this second approach is that it requires a good amount of data to be collected after treatment in order to accurately estimate the treatment effect, which may not be available and perhaps more pertinently, increases the wait time before treatments can be evaluated. Conversely the first approach may be implemented immediately following the after period, as with standard before-and-after approaches, however the estimate of treatment effect may be less accurate and confounded by heterogenous effects. An issue with all time series models however is they require a good amount of data in order for a model to be fitted at all, which we cannot assume and as stated in Chapter 1, the aim of this thesis is to develop methodologies which require as little data as possible, in the case of scheme evaluation just observations taken in a before period and after period, and so time series models are not suitable toward this aim.

2.5 Predictive Distributions

As discussed throughout this chapter, the goal of scheme evaluation studies is to estimate the treatment effect τ. Standard EB and FB methods do this by performing inference on the collision rate λ using data collected in the before period, with the distribution of treatment effect often taken to be the observed collision count in the after period minus the posterior distribution of λ, as described in Equation (2.11),

$$f(\tau) = y_{\mathrm{AFT}} - \pi\left(\lambda | y_{\mathrm{BEF},\boldsymbol{x}}\right),$$

with the estimate of treatment effect often taken at the mean of this distribution,

$$\hat{\tau} = y_{\mathrm{AFT}} - \mathbb{E}\left(\lambda | y_{\mathrm{BEF},\boldsymbol{x}}\right).$$

If however we consider the effect of treatment to be the difference between the observed collision count in the after period and the counterfactual count had treatment not been applied, the usage of posterior distribution to elicit this treatment effect can appear inadequate. If we consider the predictive distribution discussed in Section 1.5.2, we note that these distributions account for uncertainty in future observations which using the posterior distribution does not. Some discussion has taken place in the literature regarding the utility of predictive distributions in the context of SPFs [Wood, 2005], however these exist within a non-Bayesian paradigm, and so do not allow for the full characterisation of posterior beliefs, and hence predictive beliefs, in the form of a density as in the Bayesian paradigm. Hence in order to properly account for uncertainty in future observations, and hence uncertainty in any estimates of treatment effect, we propose using the posterior predictive distribution, which here takes the form,

$$f\left(y' | y_{\mathrm{BEF},\boldsymbol{x}}, \boldsymbol{x}\right) = \int_{\Lambda} f\left(y' | \lambda\right) \pi\left(\lambda | y_{BEF}, \boldsymbol{x}\right).$$

Here we note that, as defined in Section 1.5.2, $f\left(y' | \lambda\right)$ corresponds to the proposed distribution of future observations $\boldsymbol{y'}$. In a stationary system, this is often taken to be the same as the current data distribution $f(y)$. However as discussed in Section 2.2, there is the possibility of temporal trend effects being present, and so these will affect the distribution of future observations. Hence in the case of EB/FB for instance, rather than assuming future observations retain the distribution of current observations, i.e.

$$f\left(y' | \lambda\right) \sim Pois\left(\lambda\right)$$

we instead adjust the mean to account for any trend effects,

$$f\left(y' | \lambda, \kappa\right) \sim Pois\left(\lambda + \kappa\right)$$

(or equivalently the rate could be $\kappa\lambda$ depending on whether trend is being accounted for additively or multiplicatively. We note here the inclusion of trend in the posterior predictive distribution prevents the need for its inclusion when determining the posterior distribution of λ, an altogether more coherent approach. The trend parameter κ can be provided either through expert prior information in which case it may be a fixed constant, or via data driven methods such as those included in Chapter 4.

Here we demonstrate the effect of using the posterior predictive distribution to estimate and quantify uncertainty on the treatment effect τ using the before-and-after dataset from Northumbria described in Section 1.6.

Chapter 3

The Effect of Comparison Pool Exchangeability on RTM Effect Bias

Notation

Below is a summary of the notation used to describe the statistical framework used in this chapter. All notation used in other chapters carries the same meaning as in this chapter.

Notation	Meaning
i	Site indicator, $i = 1, \dots, T$
j	Covariate indicator, $j = 1, \dots, P$
c	Comparison pool indicator $c = 1, \dots, C$
n	Comparison pool size
N	Number of replications used in the simulation study
y_i	Collision count at site i
ρ_i	RTM effect for treated site i
$\delta_{i,c}$	Error in estimate of RTM effect at site i using comparison pool c
$x_{j,\mathrm{T}}/x_{j,c}$	Covariate j for the treated pool/comparison pool c
γ_i	Overdispersion parameter of the Negative Binomial distribution for site i

3.1 Introduction

In this chapter we shall examine further the claims made in Chapter 2, regarding the need for exchangeability between treated groups and comparison groups, when carrying out an EB/FB based before and after study. It is claimed that failing to ensure suitably exchangeable comparison sites will lead to biased estimates of the RTM effect, $\hat{\rho}_i$ and hence of the treatment effect, $\hat{\tau}_i$. It is clear how the comparison sites have an explicit effect on $\hat{\rho}_i$, given the direct relationship between $\hat{\rho}_i$ and the posterior mean of the collision rate $\mathbb{E}\left(\lambda_i|y_{B,i}\right)$ (via Equation (2.9)) with the comparison pool based SPF providing μ_i, the prior mean for λ_i (given in Equations (1.1) and (2.5) respectively). From this it makes intuitive sense that having poor comparison data could lead to an inaccurate estimate of μ_i which in turn would lead to an inaccurate $\mathbb{E}\left(\lambda_i|y_{B,i}\right)$ and hence an inaccurate $\hat{\rho}_i$, and this wisdom has been accepted by some authors ([De Pauw et al., 2014], [Hauer, 1991] [Persaud and Lyon, 2007]) but never numerically verified. It is the purpose of this chapter to numerically demonstrate a decrease in accuracy, henceforth referred to as an increase in bias, of $\hat{\rho}_i$, as the comparison pool used becomes increasingly dis-similar to the treated pool being analysed, via a simulation study. The contribution provided here being a demonstration of the potential severity of the problem caused by non-exchangeable reference data and hence further motivate the need for methods to account for these biases (which we then attempt to do in Chapter 4).

The details of the simulation study design are given in Section 3.2, the results as it pertains to RTM bias are given in Section 3.3, and examples of statistical tests for the exchangeability of treated and comparison pools are demonstrated using the simulated data in Section 3.4.

3.2 Simulation Study Outline

We wish to numerically demonstrate an increase in bias in estimates of the RTM effect ρ_i, as the comparison pool of sites from which the SPF is obtained becomes less similar to the treated pool. Clearly in order to enumerate the bias of an estimator, we must know the true value of the parameter it is estimating, in this case the true RTM effect ρ_i for which we require the true underlying collision rate λ_i. Clearly it is impossible to know this in reality, and limited datasets make even reliable estimates impossible, and so we must simulate a road safety dataset, where the parameters driving the data can be known exactly. We can then carry out a before and after study on the simulated data, using a variety of possible comparison pools, and obtain a $\hat{\rho}_i$ which can then be compared with the true ρ_i in order to obtain an estimate of the bias. We also wish to investigate the effect of varying the prior distribution for λ_i in an FB analysis on the RTM bias, and so shall repeat our simulation study with different commonly used priors for λ_i, namely the Gamma, Lognormal and Weibull distributions in order to make the findings of the simulation study as thorough and expansive as possible.

The general methodology of the simulation study is:

1. We simulate covariates $x_1, \ldots, x_P$ at T sites using specified covariate generating distributions $f_1(\cdot), \ldots, f_P(\cdot)$, where the covariate generating distributions have means which are monotonic functions of an input c, which we use to create dissimilarity between treated and comparison pools. We use the case $c = 0$ to generate covariates for the "treated" pool of sites.

2. Using a fixed SPF structure we convert these covariates into values of μ_i, the prior mean at each treated site i,
$$\mu_i = g\left(\boldsymbol{x}_i\right)$$
and hence generate a collision rate λ_i at each treated site from a prior distribution with mean μ_i
$$\lambda_i \sim h\left(\text{mean} = \mu_i\right)$$
and finally a collision count for the "before" period (although for the purposes of RTM estimation we do not need to consider the after period)
$$y_i \sim Pois\left(\lambda_i\right)$$

3. Hence we calculate the true RTM effect at each treated site,
$$\rho_i = \lambda_i - y_i.$$

4. We then simulate covariate data $x_1, \ldots, x_P$ to form C comparison pools of n sites. The comparison pool index c, $c = 1, \ldots, C$ determines the mean of the covariate generating functions $f_1(\cdot), \ldots, f_P(\cdot)$ used to simulate the covariate values.

5. We use the same SPF structure as for the treated sites to generate the prior mean μ_i at each comparison site in each comparison pool,
$$\mu_i = g\left(\boldsymbol{x}_i\right)$$
and hence generate a collision rate λ_i
$$\lambda_i \sim h\left(\text{mean} = \mu_i\right)$$
and finally a collision count,
$$y_i \sim Pois\left(\lambda_i\right)$$

6. For each of comparison pool c we carry out an FB analysis to obtain the posterior mean at each treated site i, $\mathbb{E}\left(\lambda_i|y_i, \boldsymbol{x}_c\right)$, and hence obtain an estimate of the RTM effect at each treated site,
$$\hat{\rho}_{i,c} = \mathbb{E}\left(\lambda_i|y_i, \boldsymbol{x}_c\right) - y_i.$$
7. We then obtain the absolute error in estimate of RTM effect for each treated site, $\delta_{i,c} = |\hat{\rho}_{i,c} - \rho_i|$ and hence the overall RTM error for comparison pool c, $\delta_c = \sum_{i=1}^{T} \delta_{i,c}$.
8. Repeat steps 4 to 7 N times in order to obtain a distribution of δ_c for each comparison pool c, of which we take the mean as the point estimate of RTM error for pool c.

For our study we choose to simulate data for $T = 50$ treated sites, with $C = 100$ comparison pools used each comprising $n = 50$ sites, and repeat the study $N = 1000$ times. We simulate two covariates, average observed speed (x_1) and an urban/rural indicator variable (x_2). The covariate generating distributions are:

$$x_1 \sim N\left(30 + \frac{40}{C}c, 1\right)$$
$$x_2 \sim Bern\left(0.7 - \frac{0.4}{C}c\right)$$

indicating in our study, the treated pool comprises of low speed, urban locations, with the comparison pools having gradually higher average speed and a great proportion of rural locations. We choose to retain the log-linear SPF structure

$$\mu_i = \exp\left(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_P x_{i,P}\right)$$

with coefficients $\beta_0 = 3$, $\beta_1 = -0.05$ and $\beta_2 = 0.8$. Finally we choose to employ a standard EB Gamma prior for the collision rate λ_i,

$$\lambda_i \sim Ga\left(\gamma, \frac{\gamma}{\mu_i}\right),$$

and choose dispersion parameter $\gamma = 1$.

3.3 Simulation Study Results

Simulating data in this way generates data such as that shown in Table 3.1 which shows that as c increases, the mean average speed (x_1) in the group increases and the proportion of urban sites decreases (x_2) along with collision count (y), showing clearly that as c increases, the data becomes increasingly dis-similar to the treated pool.

	x_1	x_2	y
Treated	30.02	0.7	37.72
$c = 1$	30.40	0.7	32.91
$c = 10$	34.00	0.66	26.82
$c = 100$	67.00	0.30	4.51

Table 3.1: Table showing covariate and collision count means for the treated pool as well as comparison pools $c = 1, 10 \text{and} 100$

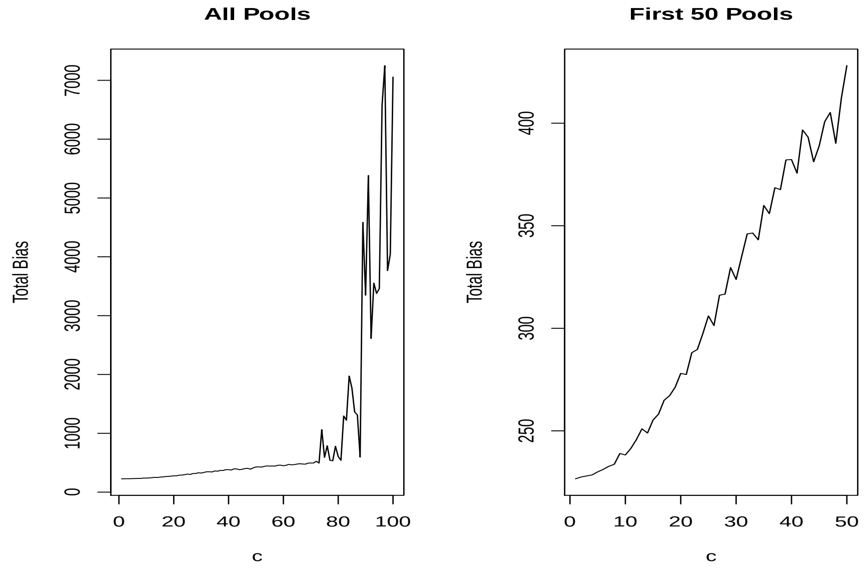
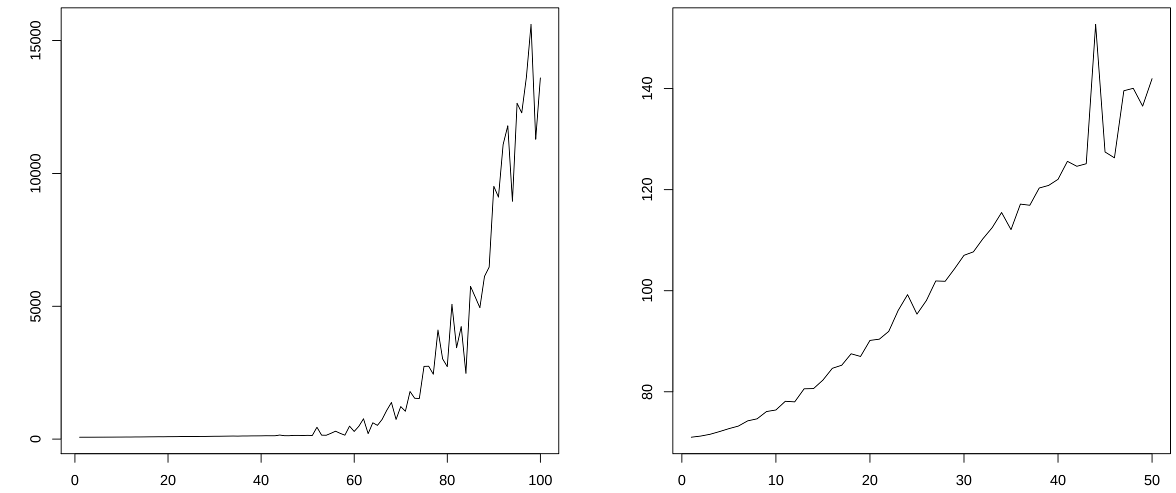


Figure 3.1: A plot showing mean total RTM error against comparison group c. The plot on the left shows all $C = 100$ comparison pools, the plot on the right focuses on the first 50.

Carrying out the simulation study as described above gives results shown in Figure 3.1 where there is a clear increase in overall RTM error δ is c increases. This demonstrates clearly that as the comparison group becomes less and less similar to the treated pool, we greatly increase the risk of obtaining inaccurate results in terms of estimates of the RTM effect which directly influences estimates of the treatment effect.

To validate our simulation study we repeat the study with different covariates and SPF coefficients. For our repeated study we take SPF equation,

$$\mu_i = \exp\left(-0.8 + 0.06x_1 - 0.7x_2\right),$$



(a) Comparison pool $c = 1, \ldots, 100$

(b) Comparison pool $c = 1, \ldots, 50$

Figure 3.2: Plots of RTM error against comparison pool for $c = 1, \ldots, 100$ (Fig. 3.2a) and for the first 50 comparison pools (Fig 3.2b).

and take covariate generating distributions for the treated pool to be

$$\begin{aligned} x_1 &\sim N(70, 1) \\ x_2 &\sim Bern(0.25) \end{aligned}$$

and for covariate pool c,

$$\begin{aligned} x_1 &\sim N(70 - \frac{40}{C}c, 1) \\ x_2 &\sim Bern(0.25 + \frac{0.5}{C}c). \end{aligned}$$

As with the first study we simulate a pool of $T = 50$ treated sites, along with $C = 100$ comparison pools, each containing 50 sites, for $N = 1000$ simulations. Plots of mean total RTM error against comparison pool c are given in Figure 3.2 where we can see again a clear increase in RTM error with decreasing comparison pool similarity, demonstrating the need for exchangeable comparison pools in order to make valid inferences regarding RTM, and thus treatment, effects.

3.4 Post-hoc Testing for Exchangeability

There are metrics by which we can test to see how exchangeable a given comparison pool is with our treated pool, and therefore whether it would be appropriate to use in a before and after analysis. One such method is the permutation test, where we investigate the exchangeability of each of our covariates in turn. We do this for covariate j by obtaining the absolute mean difference between the comparison ($x_{j,C}$) and treated ($x_{j,T}$) observations for this covariate,

$$d_j = |\bar{x}_{j,C} - \bar{x}_{j,T}| \qquad j = 1, \ldots, n_p.$$

If our groups are exchangeable with respect to this covariate then the differences between the groups should not vary significantly if we were to randomly permute sites between the two groups. We can carry out a hypothesis test of this by repeatedly permuting the allocation of the groups (to groups say, A and B of the same size as the comparison and treated pools) and taking the absolute mean difference,

$$d'_j = |\bar{x}_{j,\mathrm{A}} - \bar{x}_{j,\mathrm{B}}|.$$

We can then test the null hypothesis that the pools are exchangeable by observing the proportion of times d'_j is greater than d_j, which we take to be the p-value of our hypothesis test, i.e.

$$I_i = \begin{cases} 1, & \text{if } d_j \geq d'_j \\ 0, & \text{otherwise,} \end{cases}$$

$$p = \sum_{i=1}^{N} \frac{I_i}{N}.$$

Whilst permutation tests provide a valuable insight into the exchangeability of specific covariates, it may be preferable to simply have a single, overall summary of the exchangeability of the two groups. We can do this by instead calculating the Mahalanobis distance [Mahalanobis, 1936], which provides a single overall summary of exchangeability. It is calculated for treated site j as,

$$D_j = \sqrt{\left(\mathbf{X}_j^{\mathrm{T}} - \overline{\mathbf{M}}^{\mathrm{C}}\right)^T \Sigma^{-1} \left(\mathbf{X}_j^{\mathrm{T}} - \overline{\mathbf{M}}^{\mathrm{C}}\right)}$$

where $\boldsymbol{X}_j^{\mathrm{T}}$ is the jth row of the covariate matrix for the treated group, $\overline{\mathbf{M}}^{\mathrm{C}}$ is the vector of covariate means for the comparison group, and Σ is the covariance matrix of the comparison group, i.e.

$$\Sigma_{i,j} = \mathrm{cov}(x_i^{\mathrm{C}}, x_j^{\mathrm{C}}).$$

We can evaluate the exchangeability of a given comparison site by considering its mean Mahalanobis distance over all $T = 50$ treated sites in our treated group, i.e.

$$\overline{D} = \frac{1}{50} \sum_{j=1}^{T} \sqrt{\left(\mathbf{X}_j^{\mathrm{T}} - \overline{\mathbf{M}}^{\mathrm{C}}\right)^T \Sigma^{-1} \left(\mathbf{X}_j^{\mathrm{T}} - \overline{\mathbf{M}}^{\mathrm{C}}\right)}$$

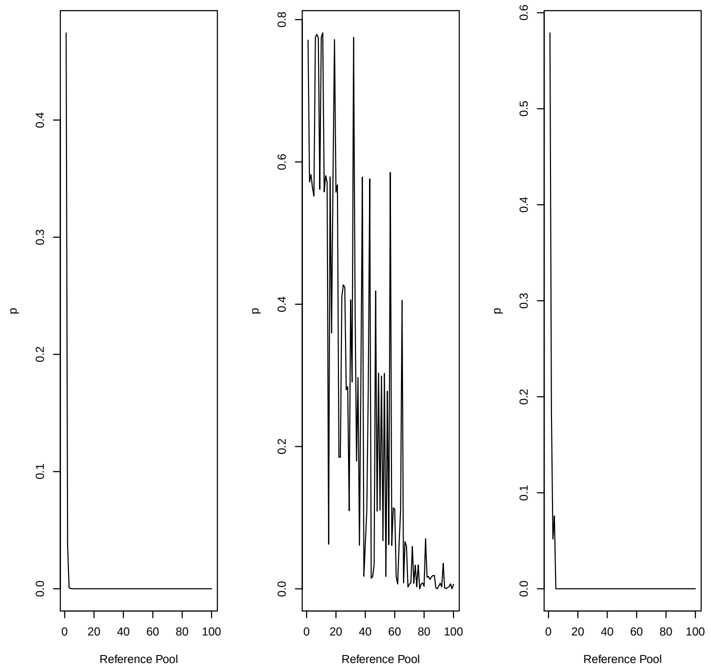


Figure 3.3: Permutation tests for x_1, x_2 and the Mahalanobis distance against comparison pool c

Computing these p-values exactly would require taking all $\binom{N=100}{100} \approx 1 \times 10^{29}$ possible combinations into the two possible groups A and B which is clearly computationally infeasible to compute exactly, and so we approximate these p-values by taking N large, in this case $N = 10000$. Here we compute the p-values for x_1, x_2 and the Mahalanobis distance, denoted p_1, p_2, p_M respectively, for each comparison pool $c = 1, \ldots, C$ used in the simulation study in Section 3.2, with results given in Figure 3.3.

Figure 3.3 shows a clear decrease in p-value as c increases, and hence as the comparison pools become less exchangeable with the treated pool, which the simulation study in Section 3.2 shows corresponds directly to an increase in RTM effect estimate error. Hence we can observe that low p-values for the permutation tests correspond to increased potential for error in RTM estimation. We further validate the post-hoc tests by carrying them out for data generated in the second simulation study with changed parameters, with output given in Figure 3.4, where again we see a clear decrease in p-value with increase in c, and hnce decrease in pool exchangeability.

We can apply the permutation tests to the Northumbria dataset analysed in Chapter 2. We recall for this dataset we have covariates:

- Average speed (x_1)

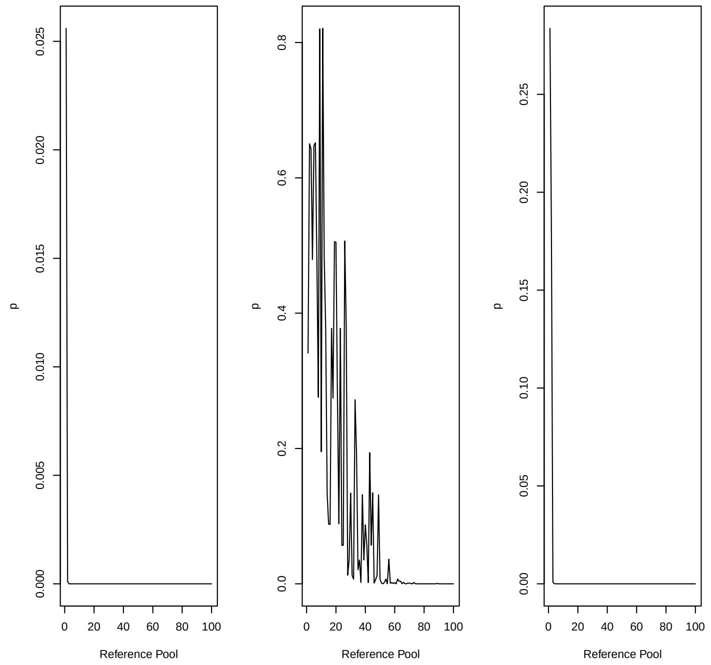


Figure 3.4: Permutation tests for x_1, x_2 and the Mahalanobis distance against comparison pool c

	x_1	x_2	x_3	x_{4A}	x_{4B}	x_{4C}	Mahal.
p-value	0.02	0.57	0.03	0.15	0.25	0.09	0.24

Table 3.2: Table showing p-values for permutation tests from the Northumbria before-and-after dataset

- Percentage exceeding the speed limit (x_2)
- Average daily flow (in thousands) (x_3)
- Road class (3 levels: x_{4A}, x_{4B}, x_{4C})

Computing the exact p-values would require $\binom{123}{67} \approx 4.68 \times 10^{35}$ simulations, and so again we obtain approximate p- values using $N = 100000$ simulations. Hence we obtain the following p-values with output given in Table 3.2. From Table 3.2 we see that we have significant ($p < 0.05$) results for x_1 (average speed) and x_3 (flow), indicating the comparison sites are not exchangeable with the treated sites with respect to these variables. While the p-value corresponding to the Mahalanobis distance remains non-significant and so the comparison site as a whole should not be discarded, it may be prudent to consider removing x_1 and x_3 from the analysis and/or attempting to use other more exchangeable covariates in their place.

Chapter 4

Accounting for Comparison Site Non-Exchangeability and Temporal Trend

4.1 Notation

Notation	Meaning
i	Treated site indicator, $i = 1, \dots, n$
c	Comparison pool indicator $c = 1, \dots, C$
j	Covariate indicator, $j = 1, \dots, P$
y_i	Collision count in the before period for treated site i
ρ_i	RTM effect for treated site i
$\bar{x}_{j,\mathrm{T}}/\bar{x}_{j,\mathrm{c}}$	The mean value of covariate j for the treated pool/comparison pool c

4.2 Introduction

As discussed in Chapter 3 there is a clear and demonstrable risk of bias in estimates of RTM and thus, treatment effect when non-exchangeable pools of comparison sites are used in an analysis, regardless of the analytic framework adopted. It is clearly important therefore to ensure, when selecting sites to form a comparison pool, that these sites are sufficiently exchangeable with the treated sites we wish to analyse. In Section 3.4 we discuss various measures for retrospectively determining the exchangeability of a given comparison pool, although this is suboptimal for several reasons. Firstly, it is extremely inefficient to construct comparison pools prior to determining their suitability, and thus leave open the possibility of having to repeat the process. This is further hindered by the retrospective approach not providing much useful information regarding how best to select future sites so as to improve comparison pool exchangeability. Clearly therefore an improvement on this approach would be a method which identifies, from a large group of candidate comparison sites, the most appropriate sites from which an exchangeable comparison pool can be formed. While this can be done to a certain degree without any form of computational algorithm, selecting sites of a roughly similar makeup of variables from the same geographic location as the treated sites for instance, it would be preferable to have a method which provides an objective result to determine which sites are most suitable. One method which is gaining popularity (see [Li and Graham, 2016], [Wood and Donnell, 2016], [Wood et al., 2015]) is the usage of the propensity score matching approach.

4.3 Propensity Score Matching

The Propensity Score Matching (PSM) method was originally implemented in the field of medicine, for a similar purpose - the elicitation of a treatment effect via the removal of confounding effects - to that of a scheme evaluation study in road safety. The method's primary purpose is to determine similarity between treated and non-treated individuals by using covariates describing the individuals to determine the probability that each would be selected for treatment. This makes the approach a good fit for comparison site elicitation since these covariates will have already been measured in order to develop the SPF. Once the probability of each individual, both those who are treated and non-treated, being selected for treatment - known as the individual's propensity score - has been determined, these scores are then compared with each other, with the similarity of the scores corresponding to the similarity of the corresponding individuals. This algorithm can be described more formally as having two key stages.

- Propensity Score Estimation. Here our response variable, Y_i, $i = 1, \ldots, n$, is a binary response variable indicating whether individual i, was selected for treatment, i.e.

$$Y_i = \begin{cases} 1, & \text{Individual } i \text{ was selected for treatment,} \\ 0, & \text{otherwise.} \end{cases}$$

  We estimate $p_i = Pr(Y_i = 1)$ by carrying out a logistic regression on Y_i with set of covariates X. This vector of covariates must be chosen carefully however, so as to satisfy the two key assumptions underlying the PSM approach: the conditional independence assumption, also known as the unconfoundedness condition; and the common support condition, also known as the overlap condition.

  1. The conditional independence assumption (CIA) states the value of Y_i must dependent only on the covariates included in X and after conditioning on X,

the value of Y_i is purely random, i.e.

$$Y_i|X \sim Bern(p)$$

where p is the proportion of sites to be treated.

2. The common support condition (CSC) states that all individuals i should have positive probability of being selected for treatment, i.e.

$$0 < p_i < 1, \forall i.$$

This is to ensure a good mixing of propensity scores between the individuals, and hence adequate matching at the next stage.

There are a variety of link functions we could use to carry out this logistic regression, with the logit,

$$p_i = \frac{\exp\left(\beta_0 + \beta_1 + \ldots + \beta_{n_p}\right)}{1 + \exp\left(\beta_0 + \beta_1 + \ldots + \beta_{n_p}\right)},$$

and probit,

$$p_i = \Phi\left(\beta_0 + \beta_1 + \ldots + \beta_{n_p}\right)$$

being the most commonly used. Results are typically very similar between the two, and so it is tantamount to an arbitrary choice as to which link function to use, in this case we opt for a logit link function. We then carry out the logistic regression to obtain the maximum likelihood estimate of p_i, denoted $\hat{p}_i$, which is the propensity score for individual i.

- Matching algorithm selection. Clearly there are a multitude of algorithms which could be employed to "match" a propensity score with its closest neighbours, however the majority of these have clear and immediate drawbacks.

1. Nearest neighbour matching (NNM) obtains the vector of absolute differences between a given propensity score and the others,

$$\boldsymbol{d}_i = \left(d_{i1}, \ldots, d_{in}\right),$$
$$d_{ij} = |\hat{p}_i - \hat{p}_j|,$$

and rearranges it into a vector $\boldsymbol{d'}_i = \left(d'_{i,1}, \ldots, d'_{i,n}\right)$, such that $d'_{i,1} \leq d'_{i,2} \leq \ldots \leq d'_{i,n}$. Hence $\boldsymbol{d'}_i$ is the vector of "nearest neighbours" for individual i. From this we then select the first n_{C} individuals (not including d'_{i1} which will correspond to individual i itself) to form the comparison pool for individual i. The advantages to this approach are clear, only the most similar individuals are selected and so we can be certain we have taken the best subset possible of the treated pool. However there are equally apparent disadvantages to this approach, firstly the seemingly arbitrary choice of n_{C} although this is clearly bounded above by n, the total number of individuals from which to choose, and the lower bound could depend on P, the number of covariates being included in the model. Secondly there is the lack of information regarding the values of d for the first n_{C} individuals, meaning there is no way to observe how similar the chosen comparison pool is to the treated individual. If we consider the scenario where $d'_{i1}, \ldots, d'_{in_{\mathrm{C}-4}}$ were small, but there is a large difference to $d'_{in_{\mathrm{C}-3}}, d'_{in_{\mathrm{C}-3}}, \ldots$, it would clearly make much more sense to only include the first $n_{\mathrm{C}-4}$ individuals, and exclude the remaining 4 which are significantly less similar to the treated individual, which may lessen the accuracy of the analysis. Likewise if the first $n_{\mathrm{C}+3}$ individuals all had low, similar values of d before a large jump for the $n_{\mathrm{C}+4}$th individual, it would make sense to extend the comparison pool size to incorporate these extra individuals, as otherwise we would be needlessly discarding useful information. Attempting to solve this manually by inspecting $\boldsymbol{d}_i$ for each treated individual will be incredibly cumbersome if n is large, almost defeating the point of using a matching algorithm in the first place, and can lead to issues of deciding at what point d is sufficiently large that an individual should no longer be considered sufficiently “similar”, meaning the result is no longer the objective, algorithmic solution we desire.

2. Radial matching (RM) overcomes the issues of NNM, by matching based on the actual values of $\boldsymbol{d'}$ rather than merely the position of the values in the vector. Here a radius r is fixed, and the selected subgroup for treated individual i is the subset of $\boldsymbol{d}_i$ corresponding to all elements less than r. This provides the advantage that we can be certain that all individuals within each subgroup have at least a certain degree of similarity to the treated individual. However the obvious drawback to this is we cannot guarantee the number of individuals included in each subgroup, which can then cause issues regarding model fit later. As mentioned previously, the minimum number of individuals in each subgroup depends largely on the number of covariates included in the model, and should be substantially greater than this, that is $n >> n_p$ to ensure (a good) model fit. We can adapt the RM approach by specifying a minimum threshold number of individuals per subgroup, and setting r to be the minimum value which reaches this threshold for all i. However this raises several issues, firstly it returns to the main issue with the NNM approach, in that it requires the specification of a minimum subgroup size, and hence devalues one of the

main appeals of adopting RM in the first place. Furthermore this devalues the concept of selecting the radial value r, which conceptually should be a value of similarity we wish to ensure, as opposed to largely being decided in order to accommodate the data.

As discussed previously, there are significant drawbacks to the implementation of step 2 of the PSM algorithm, namely the component which matches the propensity scores to a suitably similar subgroup. A further issue which underpinning the entire concept of finding a subset of potential comparison individuals, is the binary outcome of whether an individual is included or not. This has the rather unwelcome consequence, that all individuals which meet the criteria for selection, are deemed to provide an equal amount of information to the analysis, and those individuals which do not are deemed to provide no information at all. These rigid structure clearly will not give the most efficient use of a candidate pool of individuals, since even within the accepted group some individuals will be more similar, and thus appropriate, for a given treated individual, and likewise those which are not selected do not (necessarily) provide no information at all. This idea of discarding individuals which do not meet the selection criteria is also not optimal as this is simply a waste of data, which becomes especially problematic if the initial set of individuals we are selecting from is small to begin with. Clearly the most efficient use of the available data would be to include all candidate individuals in the analysis, but weight their contributions by their similarity to the treated individual in question. It is this usage of the kernel matching (KM) approach which gives rise to a solution to the issue of comparison site selection in road safety analyses: propensity score weighted regression.

4.4 Propensity Score Weighted Regression

As discussed in Section 1.3, comparison sites are used in road safety analyses, particularly in scheme evaluation analyses, to overcome the danger of effects of RTM and trend causing biased results when little data are available. Commonly this is done via the formation of an SPF which uses collision and covariate data from a pool of comparison sites to train a model which provides an estimate of the collision count at a site, for a given set of input covariates. In Chapter 3, we showed how this approach can lead to biased estimates when the comparison group of sites were not sufficiently similar, or exchangeable, to the treated pool of sites. Section 4.3 described how the concept of propensity scores can be used to quantify the similarity between a treated site and candidate comparison sites, however there are significant drawbacks with the most commonly applied matching algorithms, making an entirely PSM-based approach undesirable. One drawback of the PSM mechanism, that of treating all selected individuals uniformly, can also be applied to the standard approach for applying SPFs. We recall from equation 1.1 that normally SPFs are fitted using a standard, non-weighted Negative Binomial generalised linear model, however as discussed in this chapter and Chapter 3, not all data points (comparison sites) are equally informative in the regression model, as some sites will be more similar to the treated site, and so provide more information. Clearly therefore our regression should reflect this, and so a weighted regression should be used to form the SPF, with the data points weighted by their similarity to the treated site in question.

4.4.1 Weighted Regression

We are familiar with the concept of ordinary least squares regression, whereby the general model structure is of the form,

$$\begin{aligned} y_i &= \eta_i + \epsilon, \\ \eta_i &= \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_P x_{P,i}, \\ \epsilon_i &\sim f(\cdot, \sigma^2) \end{aligned}$$

where y is the response variable, η_i is the linear predictor, and ϵ is the error term with distribution f depending on the form of linear model being fitted. The key feature of this is that the error terms are drawn from the same distribution, and thus ordinary least squares invokes the assumption of homoscedasticity, i.e. constant variance σ^2 on the error terms. This in effect means that the process of minimising the residual sum of squares in order to obtain the maximum likelihood estimates for the regression coefficients β_j gives equal weighting to all data points, i.e. all error terms ϵ_i have the same priority in terms of being minimised. However this assumption of equal importance among the data points may not always be appropriate, as we may wish to give some data points greater influence over our fitted model than others. In this case we should relax the homoscedastic assumption on the error terms, and instead fit a heteroscedastic structure to them,

$$\begin{aligned} y_i &= \eta_i + \epsilon_i, \\ \eta_i &= \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_P x_{P,i}, \\ \epsilon_i &\sim f\left(\cdot, \frac{\sigma^2}{w_i}\right). \end{aligned} \tag{4.1}$$

Equation (4.1) now provides a weighted least squares structure, whereby the error terms are no longer identically distributed, and the variance of each error, ϵ_i depends on its weighting, w_i. This therefore induces a hierarchy of importance into the data points, since data points which have error terms with large variances will influence the resulting estimates of the coefficient parameters β_j, less than those with a smaller variance. Hence we can state that the influence of data point i is proportional to its weight w_i, and so for data points we wish to prioritise we should allocate them a large value of w_i and vice versa for those we wish to assign a low importance to. We note here that it is common to standardise the weights such that the weight vector $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)$ sums to 1, and so $0 < w_i < 1, \forall i$, and hence data points of high influence will have values of w close to 1, and low priority data points will have w values closer to 0. Weighted regression can be carried out easily in `R` via the `weights` argument in either the `lm` or `glm` commands (or in this case the `glm.nb` command).

4.4.2 Similarity Weighted SPFs

As discussed in Section 4.4, not all data points contributing to the SPF have equal importance, with some being more exchangeable to the treated site, and thus providing more useful information than others. We wish to carry out a weighted regression in place of an ordinary least squares regression when forming the SPF in order to reflect this, however in order to do so we must first construct a vector of weights, $\boldsymbol{w}_i$ which reflects the similarity between each comparison site and the treated site i (and hence the importance we wish to assign to each comparison site). Fortunately we already have a mechanism to provide this, as shown in Section 4.3, the vector $\boldsymbol{d}_i$, the difference between the propensity scores of the candidate comparison sites and treated site i, provides a numerical summary of the suitability of each candidate comparison site to be used in the SPF for treated site i. Since the importance of a candidate comparison site j in the regression is inversely proportional to the value of d_{ij} (the larger the value of d_{ij}, the less similar candidate comparison site j is to treated site i, and hence the less importance we wish to place on it in the regression), and the size of the value of w_{ij} is proportional to the weighting given to site j in the regression, we wish for the value of d_{ij} to be inversely proportional to the value of w_{ij} (i.e. the larger the value of d_{ij}, the smaller the value of w_{ij} and vice-versa). There are a variety of kernels by which the propensity score differences can be converted to weights, generally the results do not vary strongly between them [Yu et al., 2014a] and so the choice is largely arbitrary. Here we elect to make use of a Gaussian weight function,

$$w_{ij} = \exp\left(-\frac{d_{ij}}{b^2}\right),$$

where d_{ij} is the difference in propensity scores between treated site i and comparison site j, w_{ij} is the corresponding weight, and b is a bandwidth parameter to be chosen. We note here that the Gaussian kernel successfully maps values of d to the range [0, 1], with larger values of d_{ij} resulting in smaller values of w_{ij} and vice-versa, as desired. This is essentially the same as a kernel matching algorithm, sometimes used in a standard PSM approach in step 2 of the algorithm given in Section 4.3, except here we have embedded the weighting within the regression used to form the SPF for each treated site. The result of this is a much more efficient use of the available data, as it avoids the need to throw information away, whilst respecting the fact the utility of each comparison site depends on its similarity to the treated site in question. As for the PSM approaches discussed in Section 4.3, this then gives rise to site-specific SPFs, where the vector of weights $\boldsymbol{w}_i$ is clearly specific to each treated site, and hence the resulting SPF regression equation shall also be site specific. This provides an additional advantage to a PSWR approach rather than the standard SPF approach discussed in Section 1.3, since it is clearly illogical, particularly where the number of treated sites is large and/or diverse, to expect the same SPF equation to govern all treated sites equally well. We can therefore view PSWR as being a more general case of SPF fitting, with the standard global SPF structure being recovered when the weights are all equal for all treated sites, i.e. when all treated sites and comparison sites have identical propensity scores with each other, which is clearly unrealistic and so further reason to support a PSWR approach.

4.4.3 PSWR Demonstration

We demonstrate the PSWR approach using the simulated data from Chapter 3. We recall the mechanism for the simulation study whereby we simulate data to form an analysis pool of sites, and then simulate a series of comparison pools from which we can form an SPF to better inform our analyses. The simulation study in Chapter 3 demonstrated that as the comparison sites became increasingly less exchangeable with the treated pool, the error in RTM effect estimation induced in the analysis significantly grew. We therefore concluded that we should avoid the inclusion of non-exchangeable sites in our comparison pool to train the SPF, so as to best reduce errors in RTM effect estimation. Within the PSWR framework we avoid discarding data altogether, preferring to weight data points by their utility, here exchangability to the treated pools. Hence in order for our algorithm to be effective, we desire for it to identify suitably exchangeable candidate sites within a pool and give them increased weighting in the analysis, while giving less weight to non-exchangeable sites within a pool. Our methodology to verify PSWR as effective therefore becomes:

1. Simulate covariate and collision data to form our analysis pool.
2. Simulate a diverse pool of candidate comparison sites with varying degrees of exchangability with the analysis pool
3. Run the PSWR algorithm on the treated pool and the entire pool of candidate comparison sites

4. Check the weightings given to each candidate site relative to their exchangeability (and thus likelihood of contributing an error to the analysis) with the analysis sites

We retain the covariate generating distributions for our analysis sites as used in the simulation study in Chapter 3 and so we simulate two covariates: average speed (x_1) and a binary indicator variable corresponding to whether the site was in an urban location (x_2). We use the following SPF to generate estimates of μ_i from the covariates,

$$\mu_i = \exp\left(3 - 0.05x_{i,1} + 0.8x_{i,2}\right) \qquad i = 1, \ldots, n \tag{4.2}$$

and use the following covariate generating functions to produce covariates for the analysis sites,

$$\begin{aligned} x_1 &\sim N\left(30, 1^2\right), \\ x_2 &\sim Bern(0.7). \end{aligned}$$

We shall retain a classical Empirical Bayes structure to generate collision rates,

$$\lambda_i \sim Ga\left(\gamma, \frac{\gamma}{\mu_i}\right)$$

where we take overdispersion parameter $\gamma = 1$, and hence we generate collision counts for the before period

$$y_i \sim Pois\left(\lambda_i\right).$$

As in the simulation study we make the mean of the covariate generating functions for the comparison data to be a function of the group number $c = 1, \ldots, C$ so as to ensure increasing dissimilarity between analysis and comparison groups as c gets bigger. In this case in order to generate a heterogenous pool of candidate comparison sites with varying exchangeability to the treated pool, we sample a small number of sites from each pool c to form a single comparison pool, where clearly we would wish to give more weight to sites sampled from groups where c is small compared with large. Here we simulate from $C = 20$ pools, sampling 10 sites from each to generate a pool of 200 candidate comparison sites along with $n = 100$ sites to form the analysis pool. The covariate generating distributions for comparison group c are

$$\begin{aligned} x_1(c) &\sim N\left(30 + \frac{40}{C}c, 1\right), \\ x_2(c) &\sim Bern\left(0.7 - \frac{0.4}{C}c\right), \end{aligned}$$

for which we use the same SPF in Equation (4.2) to convert the covariates into elements of $\boldsymbol{\mu}$ which we then use the same EB structure with overdispersion parameter $\gamma = 1$ to generate collision rates λ_i and thus collision counts y_i for each comparison site.

With this data we are now in a position to carry out a PSWR analysis, we combine the covariate data for the analysis and comparison datasets and assign a binary indicator to each data point to correspond to whether the datapoint belongs to an analysis site or not ($T_i = 1$ if an analysis site, 0 otherwise). We then carry out a logistic regression on T_i using the simulated covariates x_1 and x_2 as explanatory variables in order to obtain $\hat{p}_i$, the fitted probability of site i being treated, which in this context is the site's propensity score. Hence for each treated site we compute the vector of differences in propensity score between treated site i, and each candidate comparison site $j = 1, \ldots, C$

$$\boldsymbol{d}_i = (d_{i1}, \ldots, d_{iC}), \qquad\qquad = (|\hat{p}_i - \hat{p}_1|, \ldots, |\hat{p}_i - \hat{p}_C|).$$

From each bespoke difference vector $\boldsymbol{d}_i$, we obtain a bespoke vector of weights, $\boldsymbol{d}_i$, subject to an inverse distance weight function,

$$w_{ij} = f(d_{ij})$$

where $f(\cdot)$ is a monotonically decreasing function. In this case we choose a simple inverse distance metric,

$$w_{ij} = \frac{1}{|d_{ij}|}.$$

This choice of metric is made for simplicity, removing the need for any parameter specification, thereby retaining the method's lack of need for expert involvement to be implemented. Computing the mean propensity score difference for each candidate comparison site, $\bar{d}_j = \frac{1}{n}\sum_{i=1}^{n} d_{ij}$, and hence the mean weighting $\bar{w}_j = \frac{1}{n}\sum_{i=1}^{n} w_{ij}$ and plotting against the comparison pool from which the site was taken gives plots shown in Figure 4.1. From Figure 4.1 we see a clear increase in propensity score difference, and hence decrease in SPF weighting with increasing comparison pool c, demonstrating that PSWR gives more weight to more exchangeable comparison sites and downweights non-exchangeable comparison sites, thereby making the SPF more closely resemble a form which would give lower error in RTM effect estimate. We highlight here that PSWR was able to identify the desirable datapoints from a heterogenous dataset completely autonomously without any need for external input, thereby making it a highly useful tool for practitioners, particularly those without statistical training.

4.4.4 Bayesian Propensity Score Weighted Regression

From Section 4.4.3 we can be satisfied that PSWR provides a useful method for incorporating comparison sites into a before-and-after scheme evaluation analysis. However the approach to PSWR outlined thus far in Section 4.4 is not without its drawbacks, mostly akin to those with the EB method discussed in Section 2.2. As with the EB method, the

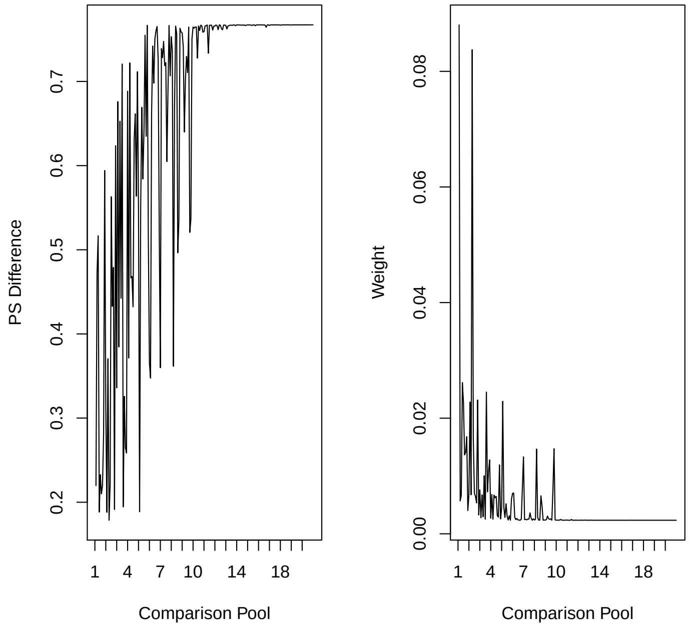


Figure 4.1: A plot of mean propensity score difference and mean SPF weighting against comparison pool c for each candidate comparison site

PSWR procedure we have outlined is based in a non-Bayesian, frequentist paradigm, and as such experiences the familiar drawbacks of non-Bayesian methods when compared to Bayesian approaches. The chief limitations of the above PSWR procedure are:

- **Lack of uncertainty acknowledgement on the weightings**. Because the logistic regression used to obtain the propensity scores is carried out in a non-Bayesian paradigm, the obtained maximum likelihood estimates for the regression coefficients are treated as "exact", with no uncertainty on them propogating through the analysis. The result of this is then a lack of measure of uncertainty regarding the weights

- **No scope for incorporating prior knowledge**. While, as with all other methods discussed in this thesis, we make no assumption of the availability of expert prior knowledge, where it is available it should be incorporated into an analysis so as to avoid wasting information. This is of particular use in a PSWR analysis where there is potential for the inclusion of prior information both at the logistic regression stage and in the SPF stage. It seems sensible that experienced practitioners will have insights as to the impact of a covariate on the likelihood of a location being selected for treatment, and additionally the anticipated effect a collision will have on collision counts in the SPF.

The Bayesian PSWR (B-PSWR) algorithm can be broken down into 3 main parts:

- Obtain a sample of the coefficient vector $\boldsymbol{\zeta} = (\zeta_0, \ldots, \zeta_L)$ from the logistic regression of treatment indicator on covariates.
- Using the obtained coefficient vector, calculate the fitted probability of treatment, i.e. the propensity score, for each site, $\boldsymbol{p}_i$ and hence the vector of differences $\boldsymbol{d}_i = (d_{i1}, \ldots, d_{iC})$ where $d_{ij} = |p_i - p_j|$ for treated sites i and comparison sites j.
- From the difference vector $\boldsymbol{d}_i$, obtain a vector of weights, $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iC})$ using weight function f, $w_{ij} = f(d_{ij})$ and hence obtain a sample of the SPF coefficient vector $\boldsymbol{\beta}_i = (\beta_{0,i}, \ldots, \beta_{P,i})$ by caryying out a weighted Negative Binomial GLM regression using comparison sites $1, \ldots, C$ with corresponding weights $w_{i1}, \ldots, w_{iC}$.

The algorithm for carrying out B-PSWR is therefore:

1. Initialise $\boldsymbol{\zeta} = (\zeta_0, \ldots, \zeta_L)$ at its initial value $\boldsymbol{\zeta}^{(0)} = \zeta_0^{(0)}, \ldots, \zeta_L^{(0)}$ and $\boldsymbol{\beta}_i$ at $\boldsymbol{\beta}_i^{(0)} = \left(\beta_{0,i}^{(0)}, \ldots, \beta_{P,i}^{(0)}\right)$ for $i = 1, \ldots, n$. Initialise counter $m = 1$.

2. For each element $1, \ldots, L$ of $\boldsymbol{\zeta}$, sample a proposal value ζ_l^* using a Metropolis random walk centred at the current value $\zeta_l^{(m-1)}$,

$$\zeta_l^* | \zeta_l^{(m-1)} \sim N\left(\zeta_l^{(m-1)}, e_\zeta\right) \qquad l = 1, \ldots, L$$

   and set $\zeta_l^{(m)} = \zeta_l^*$ with probability α

$$\alpha = \min\left(1, \frac{\pi\left(\zeta_l^*\right) L\left(\boldsymbol{x}|\boldsymbol{\zeta}_l^*\right)}{\pi\left(\boldsymbol{\zeta}_l^{(m-1)}\right) L\left(\boldsymbol{x}|\boldsymbol{\zeta}^{(m-1)}\right)}\right),$$

   else set $\zeta_l^{(m)} = \zeta_l^{(m-1)}$.

3. Using the sample $\boldsymbol{\zeta}^{(m)}$ obtain fitted values of the propensity scores $\boldsymbol{p}^{(m)}$ for the treated sites $1, \ldots, n$ and comparison sites $1, \ldots, C$.

4. For treated site $i = 1, \ldots, n$ obtain the vector of differences in propensity score between it and each comparison site j, $\left(d_{i1}^{(m)}, \ldots, d_{iC}^{(m)}\right)$, where $d_{ij}^{(m)} = |p_i^{(m)} - p_j^{(m)}|$.

5. Calculate the weight vector for treated site i, $\boldsymbol{w}_i^{(m)} = \left(w_{i1}^{(m)}, \ldots, w_{iC}^{(m)}\right)$, where $w_{ij}^{(m)} = f\left(d_{ij}^{(m)}\right)$ and $f(\cdot)$ is a non-zero monotonically decreasing function.

6. For each treated site $i = 1, \ldots, n$, and for each element of $k = 1, \ldots, P$, sample a proposal value β_k^* using a Metropolis random walk centred at the current value $\beta_k^{(m-1)}$,

$$\beta_{k,i}^* | \beta_{k,i}^{(m-1)} \sim N\left(\beta_{k,i}^{(m-1)}, e_\beta\right) \qquad k = 1, \ldots, P.$$

   and set $\beta_{k,i}^m = \beta_{k,i}^*$ with probability α

$$\alpha = \min\left(1, \frac{\pi\left(\beta_{k,i}^*\right) L\left(\boldsymbol{x}_i | \boldsymbol{\beta}_i^*, \boldsymbol{w}_i^{(m)}\right)}{\pi\left(\beta_i^{(m-1)}\right) L\left(\boldsymbol{x}_i | \boldsymbol{\beta}_i^{(m-1)}, \boldsymbol{w}^{(m)}\right)}\right)$$

   else set $\beta_{k,i}^{(m)} = \beta_{k,i}^{(m-1)}$.

7. Set $m = m + 1$. Go to step 2.

While adopting a B-PSWR approach provides a mechanism for accounting for uncertainty in the SPF weightings as well as incorporating expert prior information, it does lead to a significant computational increase, with the resulting model containing $nP + P'$ parameters (where P' is the number of covariates in the logistic regression model used to obtain the PSWR weights) and so may not be applicable to larger datasets.

4.5 Accounting for Trend

As discussed in Section 2.1.1, the change between counts before and after treatment at a site can be represented as a combination of the RTM, trend and treatment effects,

$$y_{i,\text{BEF}} = \lambda_{i,\text{BEF}} - \rho_i, \tag{4.3}$$

$$y_{i,\text{AFT}} = \lambda_{i,\text{AFT}} + \tau_i, \tag{4.4}$$

$$y_{i,\text{AFT}} - y_{i,\text{BEF}} = \rho_i + \alpha_i + \tau_i, \tag{4.5}$$

where $y_{i,\text{BEF}}$ and $y_{i,\text{AFT}}$ are the collision counts before and after treatment, and λ_i, ρ_i, α_i and τ_i are the underlying collision rate, the RTM effect, trend effect and treatment effect respectively. We have discussed extensively methods for estimating the RTM effect ρ_i, through the use of comparison sites via safety performance functions. However the estimation of the trend effect α_i has been largely neglected, and there appears to be no clear method for estimating this when we are restricted to just before and after data (although extensive research has been done for the case when there are many years of data via time series methods, see for example [Carnis and Blais, 2013], [Park et al., 2017b], [Sacchi et al., 2014]). In Chapter 2 we discussed a method used by [Fawcett and Thorpe, 2013] which is dependent on expert prior knowledge (or reliable prior data) to inform the trend component of the analysis. Whilst such an approach may be suitable should such knowledge/information be available and accurate, this is a strong assumption to make, and can be cumbersome for an expert to provide beliefs regarding their beliefs regarding trend at each site, and potentially too strong of an assumption to assume a uniform trend across all treated sites. A clear improvement would be to obtain a data driven, objective estimate for the trend effect at each individual site, without having to rely on large amounts of data.

A simple approach to this, which we advocate here, is repeating the data collection process for all sites for the after period, and combining these to form a single SPF, with additional indicator variable denoting which period the data was collected in,

$$\alpha_i = \begin{cases} 0, & \text{Observation } i \text{ was taken in the before period} \\ 1, & \text{Observation } i \text{ was taken in the after period.} \end{cases}$$

Hence we obtain,

$$\mu_i = \exp\left(\beta_0, \beta_1 x_1 + \ldots + \beta_P x_P + \tau_i \alpha_i\right) \qquad i = 1, \ldots, 2n$$

Here we assume the same time distance between before and after periods, thereby α_i need only be a binary indicator variable to denote the effect of being in the after period. This assumption could easily be relaxed with α_i becoming the effect of unit time passing, and so would have SPF coefficient equal to the number of time periods between the before and after observations. This approach provides a solely data-oriented approach to trend estimation, although should the SPF be fitted within the Bayesian paradigm, there is also the possibility to include expert beliefs regarding α_i in its prior distribution. Furthermore this approach removes the implicit assumption otherwise present in EB/FB approaches, that the covariate values at treated sites remain the same between before and after periods. This is clearly a restrictive assumption and leave the model susceptible to many flaws, for instance the possibility of an extreme covariate value (e.g. if a special event had happened near a site meaning there was an unusually large traffic flow there) meaning there is the possibility of an RTM effect in covariate values which is otherwise left unaccounted for. By combining the before and after data into a single dataset to fit the SPF, we are making the implicit assumption that the covariate effects are constant between before and after periods. While we believe this to be a reasonable assumption when there is not much temporal distance between before and after periods, and hence not much time for covariate effects to change, if there is a significant gap between periods this assumption may be less valid. In this case we advocate forming dual SPFs for the before and after periods respectively, to allow for possible changes in covariate effects,

$$\begin{aligned} \mu_{i,\text{BEF}} &= \exp\left(\beta_{0,\text{BEF}} + \beta_{1,\text{BEF}} x_{i,1,\text{BEF}}, \ldots, \beta_{i,P,\text{BEF}}\right) \\ \mu_{i,\text{AFT}} &= \exp\left(\beta_{0,\text{AFT}} + \beta_{1,\text{AFT}} x_{i,1,\text{AFT}}, \ldots, \beta_{i,P,\text{AFT}} + \tau_i \alpha_i\right) \end{aligned}$$

While this approach does require more data than say, an FB analysis with expert provided trend multiplication factor, we feel these data requirements are minimal, as it only assumes a single additional datapoint at each site, and the covariates required are the same as those for the standard SPF and so should be easily available. Furthermore by doubling the number of datapoints used to fit the model, at the cost of just a single extra parameter, we increase the stability of the model with regards to providing coefficient estimates. Hence we believe this approach provides a good, accessible means for estimating trend without the assumption of expert prior information. This approach coupled with a standard global SPF still provides an improvement, however as for the covariates effects, fitting a global SPF fails to account for heterogeneity among treated and comparison sites, assuming the effects are the same among treated sites, and all comparison sites are equally informative as to the value of these effects. Hence we again make the case that making use of PSWR, in conjunction with the dual SPF approach, allows for a site-specific trend estimate, with priority given to the most exchangeable comparison sites when eliciting this estimate.

Chapter 5

Hotspot Prediction

Notation

Below is a summary of the notation used to describe the statistical framework used in this chapter. Much of the notation carries over to Chapter 6 and has the same meaning as here.

Notation	Meaning
i	Site indicator, $i = 1, \ldots, n$
j	Covariate indicator, $j = 1, \ldots, n_p$
t	Time period indicator, $t = n_y - 1, \ldots, 0$
n	Number of sites to be analysed
n_p	Number of covariates in the SPF
n_y	Number of observations at each site in the dataset
$x_{j,i}$	The value of covariate j at site i
$y_{i,t}$	Observed collision count in time period t at site i
λ_i	Underlying collision rate at site i
σ_i	Site effect at site i
α_i	Site specific trend effect at site i
α_{N_i}	Zero-inflation component of α_i
α_{Z_i}	The Normal distribution component of α_i
τ_i	Variance inflation parameter for site i
μ_i	Fitted estimate from the SPF for site i
$\beta_{j(,i)}$	SPF regression coefficient for covariate j (at site i)
$\boldsymbol{\Psi}$	Vector of all parameters in the hotspot model

5.1 Introduction

As discussed in Chapter 1, one of the primary duties of road safety practitioners is analysing the road network, so as to discern locations to apply road safety treatments. Given limited budgets (and as discussed in Section 2.2, the potential for treatments to sometimes cause an increase in the number of casualties) treatments cannot be applied everywhere, and so some degree of selection must take place. Logically it would make sense for locations to be prioritised for treatment based on the potential for improvement at the location, that is the number of collisions we would expect the treatment to remove, we shall refer to any such locations as "hotspots". However as outlined in Section 2.1.1, we must be wary of any blip and trend effects present in the data, and must account for these so as to avoid any RTM and trend effects misleading our conclusions. Using the notation described in Chapter 2 we can consider the number of preventable collisions as a comparison between λ, the underlying collision rate at a location, and μ, the collision rate we would expect at a typical location of this nature, and hence obtain the potential for safety improvement (PSI) [Jiang et al., 2014],

$$\text{PSI}_i = \lambda_i - \mu_i, \qquad i = 1, \ldots, n.$$

Larger values of PSI indicate a site has a higher collision rate than would be expected from the network, and so would suggest it be a better candidate for treatment. Clearly the accuracy of estimate of PSI is heavily dependent on having an accurate and well-fitting SPF (in order to have an accurate μ), and since we cannot guarantee this for datasets with limited covariate information, it may be sensible in such cases to simply rank by λ outright.

We further note the importance of proactive hotspot prediction, as opposed to retrospective hotspot identification. It is common practice for hotspot locations to be designated and treated reactively [U.S. Department of Transportation, 2018], that is solely using past data, usually when some form of threshold criteria (usually in relation to collision/casualty totals) is exceeded. This is clearly not ideal, since it requires a given number of negative events to occur before any treatment is applied at a dangerous location. A better approach would be to implement a proactive approach whereby treatment is allocated based on future levels of risk at locations, without the need to wait for a threshold level of collisions etc to be exceeded.

5.2 Halle Dataset

We shall demonstrate our hotspot prediction model using a real dataset taken from the city of Halle, Saxony-Aanholt, Germany. The full dataset comprises information at 734 "nodes", where a node is a location at which the road setup changes (often this will be at an intersection but it could also be at other situations, e.g. where two roads merge rather than intersect), since nodes are specific locations, as opposed to routes or links, we can consider them the same as the sites discussed in chapter 2. The data is taken on a monthly basis, although for the purposes of this chapter we shall aggregate data to annual observations, taken over 9 years, from 2004-2012. Comprising this data are collision counts which are broken down by severity - although again in this chapter we shall aggregate counts across severities, as well as covariate information from which we can build an SPF. The included covariates are:

- Volume: The average number of vehicles at the node in a day over the year (taken on the log scale)
- MinorVolume: The average number of vehicles along the major arm of the node (taken on the log scale, equal to Volume if the node is not an intersection)
- MinorVolume: The average number of vehicles along the minor arm of the node (taken on the log scale, 0 if the node is not an intersection)
- Urban: 1 if the node is in an urban location (0 otherwise)
- Intersection: 1 if the node is at an intersection (0 otherwise)
- Signalised: 1 if the node is at a signalised intersection (0 otherwise)
- SpeedLimit: The speed limit at the node (30, 45, 50, 60, 70 or 80km/h)
- MajorRoad: 1 if the node is at a major road (0 otherwise)
- MajorIntersection: 1 if the node is at an intersection involving a major road (0 otherwise)
- FourLegs: 1 if the node is at an intersection with four arms (0 otherwise)

A summary of the collision and covariate data is given in Table 5.1.

We observe the majority of the nodes in the Halle dataset were located at urban intersections, which is to be expected given the location of the dataset, however we still have a reasonable mixture for each categorical variable. From table 5.1a we observe an apparent negative trend in total collision counts, most notably for the final 3 years of the dataset.

Year	Total	Mean	Variance
2004	2678	3.649	21.437
2005	2738	3.730	24.484
2006	2621	3.571	25.689
2007	2726	3.714	24.846
2008	2609	3.554	20.378
2009	2671	3.639	23.047
2010	2414	3.289	19.324
2011	2281	3.108	19.381
2012	2181	2.971	20.901
Total	22919	3.469	22.207

(a) Summary of collision counts from 2004 to 2012 across all 734 nodes in the Halle dataset

Covariate	Mean	St. Dev.
Volume	6.898	3.187
MajorVolume	6.648	3.241
MinorVolume	3.783	3.648

(b) Summaries for continuous covariates in the Halle dataset

Covariate	Prop.	Speed Limit	Prop.
Urban	0.911	30	0.369
Intersection	0.857	45	0.119
Signalised	0.270	50	0.215
MajorRoad	0.063	60	0.178
MajorInt	0.196	70	0.045
FourLegs	0.244	80	0.072

(c) Relative frequencies of categorical variables in the Halle dataset

Table 5.1: Summaries of the Halle dataset

5.3 Adapting Scheme Evaluation Methods

We retain the framework used to address RTM and trend described in section 2.2 by retaining a Bayesian structure to model our accident rate λ, only this time rather than be restricted to just two years of data, we allow for there to be any number of years of data from which we can form an analysis. Our model structure is given,

$$Y_{i,0}|\lambda_i(0) \sim Pois\left(\lambda_i(0)\right) \qquad i = 1, ..., n, \qquad t = 0 \tag{5.1}$$

$$Y_{i,t}|\lambda_i(t) \sim NegBin\left(p = \frac{1}{c_i(t)}, r = \frac{\lambda_i(t)}{c_i(t) - 1}\right) \qquad i = 1, ..., n, \qquad t < 0 \tag{5.2}$$

Here we see that in the current year (fixed to be time $t = 0$), we retain a Poisson distribution for our number of collisions, however as we move further into the past, $t < 0$, we switch to a Negative Binomial distribution which, whilst still having mean λ, has an increased variance λc, where c is a monotonically increasing function as t grows increasingly negative. We include this since clearly observations further into the past will be less informative to our current accident rate, and thus any predictions we form, than observations taken more recently. To model this we downweigh observations according to how far in the past they were made, by increasing the value of c, and thus increasing the variance of our Negative Binomial distribution [Fawcett et al., 2017].

$$c_i(t) = \exp(-t\tau_i), \qquad i = 1, \dots, n \qquad t < 0, \tag{5.3}$$

$$\tau_i \sim Ga(2, 20), \qquad i = 1, \dots, n. \tag{5.4}$$

Our choice of prior for τ_i was elicited from road safety experts at PTV Group through standard elicitation methods [Garthwaite et al., 2005] and chosen to be the distribution which they best felt described the rate at which past observations should be down-weighed in our model. In this elicitation procedure, we identified along with an expert, reasonable quantiles that our prior for τ should follow, based on the multiplicative effect caused by moving several years into the past. Restrictions were specified in order to avoid the variance increasing too rapidly, thereby effectively removing data points from informing our analysis, and ultimately it was decided a $Ga(2, 20)$ distribution best met the required criteria.

The purpose of this model is to form predictions regarding the number of collisions at a given site in future years, and we can form these predictions by appealing to the Bayesian posterior predictive distribution:

$$f(y_{i,1} = y|y_i) = \int_{\Lambda} f(y_{i,1} = y|\lambda_i(1))\pi(\theta_j|y_j)d\lambda_j.$$

We approximate this predictive distribution by estimating values of the underlying accident rate for the future time period, $\lambda_{i,\text{pred}}$ by extending our global and site specific trends (and assuming covariate levels are the same as in the final observation). We retain the same level of uncertainty regarding the condition in future years at our site, as we did for the past, and hence we increase the variance of observations in the next year by a factor of $c_i(1)$ and hence obtain,

$$y_{i,1}|\lambda_i(1) \sim NegBin\left(p = \frac{1}{c_i(1)}, r = \frac{\lambda_i(1)}{c_i(1)-1}\right), \tag{5.5}$$

and so obtain the predictive probability from this distribution.

5.4 Rate Parameter Structure

We model the accident rate λ as being comprised of three parts: the effect due to covariates - given by μ, the site effect - given by σ, and the site specific trend effect - given by α.

$$\lambda_i(t) = \exp(\mu_i(t) + \sigma_i + \alpha_i t), \qquad i = 1, ..., n, \qquad -\infty < t < \infty. \tag{5.6}$$

The covariate effect μ is obtained via the standard SPF structure as outlined in section 1.3, however now since we have multiple years of data, we model trend by including time t as a covariate, in order to account for global trends across the network. Here we assume linear trends in data, partly for simplicity – as assuming complicated trend structures would significantly increase the longitudinal data requirements of the model, and partly since the primary purpose of this model is short term prediction, we wish to avoid overfitting trend effects. Clearly these linear trends could be replaced by a more complicated trend function $g(t)$, should there be particular reason to suspect non-linear trends are present. Furthermore since all potential hotspot sites are currently untreated, we no longer have “reference” and “treated” pools of sites, as was the case in chapter 2 and so the SPF is formed from a single pool containing every available site,

$$\mu_i(t) = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_{n_p} x_{n_p} + \beta_t t),$$

where n_p is the number of covariates in the SPF. We do not assume any prior knowledge for the general form of our model, and so assign independent, vague prior distributions to our regression coefficients,

$$\beta_j \sim N(0, 100) \qquad j = 1, \ldots, n_p. \tag{5.7}$$

However because we fit a single model to the entire network, we cannot guarantee it will be representative of every individual site within the network, a problem which becomes more pronounced the larger the dataset. In order to account for any systemic (ie. fixed in time) differences between $\boldsymbol{\mu}$ obtained from the SPF, and the observed values at the site

$\mathbf{y}$, we include a site effect parameter, σ, which accounts for this discrepancy, as well as for any unobserved covariate effects. Since it entirely possible for this systemic deviation to be positive or negative, in the general form of the model we propose a vague Normal prior for σ,

$$\sigma_i \sim N(0, 100), \qquad i, \ldots, n. \tag{5.8}$$

Whilst $\boldsymbol{\sigma}$ accounts for any systemic deviations from the SPF, we still retain the very real possibility that not all sites will display the same trend as that estimated by β_t in the SPF formula. To account for this we propose an additional site-specific trend parameter, α, to account for site-specific deviations from this global trend. However we must guard against the possibility of trends forming in $\boldsymbol{y}$ due to chance, rather than a true trend being present, a problem which becomes more pertinent when there are few years of data available at each site. In order to decrease the likelihood of false trend detection, we impose a zero-inflated structure for α, thereby providing more weight to the case where there is no site-specific deviation from the global trend, while still allowing it to be possible, should significant evidence of it be present. Again in the general form of the model we assume prior ignorance regarding the likelihood and direction of site specific trends, and so we impose a vague zero-inflated Normal prior for α,

$$\alpha_i = \alpha_{N_i} \alpha_{Z_i}, \tag{5.9}$$

$$\alpha_{N_i} \sim N(0, 100), \qquad i = 1, \ldots, n \tag{5.10}$$

$$\alpha_{Z_i} \sim Bern(0.5). \tag{5.11}$$

We note that the Bernoulli probability could be adjusted depending on any expert prior information (i.e. the expected heterogeneity or lack thereof among trends across the network), thereby increasing or decreasing the level of zero-inflation. Here we choose 0.5 as a sensible default, so as to reduce the probability of false trends being detected whilst still allowing for a good sample size to be drawn for the local trend should it be present. Given the Normal distributions assigned to the site effect parameter, σ, and site-specific trend effect, α, the rate parameters, λ, is effectively lognormally distributed, conditional on the covariate effect parameter, μ, obtained from the SPF, i.e.

$$\log\left(\lambda_i(t) \mid \mu_i(t)\right) \sim N(\mu_i(t) + \sigma_i + \alpha_i t, 100 + 75t^2).$$

This means that in the final year of data, corresponding to $t = 0$, the model becomes analogous to a Poisson-Lognormal model, commonly used in literature ([Kitali and Sando, 2017b], Zhan et al. 2015, Wang & Kockleman 2013).

5.5 The 1 and 2 year case

We would expect, as with any statistical model, for the accuracy of our estimates, particularly those pertaining to trend, to increase with more years of data available. However we must also contend with the possibility that very few years of data will be available, with just 1 or 2 years of available data possible. Clearly in such cases the methods for estimating site-specific trend outlined in equation (5.9) are no longer valid, however we may retain the estimate of global trend since this uses information from across the network and so will have multiple data points per year. Hence in the case where there are only 2 years of data available, we remove the α term from our expression for λ given in equation (5.6), to give,

$$\lambda_i(t) = \exp\left(\mu_i(t) + \sigma_i\right)$$

and hence in this case we fit the global observed trend to each site.

In the more extreme case where only 1 year of data is available, most of our model features become invalid. Clearly we can no longer fit any form of trend, but additionally our site effect term σ becomes redundant, since this will simply become the difference between the SPF estimate, μ, and the observed collision count y, so as to make the collision rate, λ, equal to y, i.e.

$$\begin{aligned}
\sigma_i &= \log\left(y_i\right) - \mu_i, \\
\lambda_i &= \exp\left(\mu_i + \sigma_i\right), \\
&= \exp\left(\mu_i + \log\left(y_i\right) - \mu_i\right), \\
&= \exp\left(\log\left(y_i\right)\right) \\
&= y_i.
\end{aligned}$$

Due to the lack of suitability of the additional parameters we impose in our hotspot prediction model, in the case of only 1 year of available data we revert to the Fully Bayesian method discussed in section 2.2 to estimate the underlying rate parameter λ.

5.6 Model Application

5.6.1 MCMC Algorithm

We employ a Markov Chain Monte Carlo (MCMC) algorithm in order to fit the main hotspot prediction model. Given the non-conjugate structure of our model, we make use of a Metropolis-Hastings algorithm in order to carry out updates of the parameter vector, $\boldsymbol{\Psi}$. Given various components of our model (e.g. $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, c) are deterministic conditional on other parameters, the parameter vector we must carry out inference on is,

$$\boldsymbol{\Psi} = \left(\beta_0, \ldots, \beta_{n_p}, \beta_t, \sigma_1, \ldots, \sigma_n, \alpha_1, \ldots, \alpha_n, \tau_1, \ldots, \tau_n\right)$$

where here $n_p = 16$ and $n = 734$. The MCMC algorithm is implemented as follows:

1. Initialise the chain at $\boldsymbol{\Psi}^{(0)} = \left(\beta_0^0, \ldots, \beta_{n_p}^0, \beta_t^0, \sigma_1^0, \ldots, \sigma_n^0, \alpha_1^0, \ldots, \alpha_n^0, \tau_1^0, \ldots, \tau_n^0\right)$, where each parameter's initial value is the mean of the corresponding prior distribution. Set iteration counter i = 1.

2. For each element Ψ_j generate a proposed update value, Ψ_j^*. If $\Psi_j \notin (\tau_1, \ldots, \tau_n)$ the proposal is generated via a Normal random walk, i.e.

$$\Psi_j^* \sim N\left(\Psi_j^{m-1}, \epsilon_j\right).$$

   Alternatively if $\Psi_j \in (\tau_1, \ldots, \tau_n)$, the proposal is generated from a Gamma proposal distribution,

$$\Psi_j^* \sim Ga\left(\frac{\left(\Psi_j^{(m-1)}\right)^2}{\epsilon j}, \frac{\Psi_j^{(m-1)}}{\epsilon j}\right).$$

   The parametrisation of the Gamma proposal distribution was chosen such that it has mean $\Psi_j^{(m-1)}$ and variance ϵ_j as with the Normal distribution for the other parameters.

3. Set $\Psi_j^{(m)} = \Psi_j^*$ with probability p_{ij},

$$p_{ij} = min\left(1, \frac{f(\Psi_j^{(*)})}{f(\Psi_j^{(m-1)})} \frac{\pi(\Psi_j^{(*)})}{\pi(\Psi_j^{(m-1)})}\right)$$

   where $f(\Psi_j^{(*)})$ is the model likelihood evaluated using the proposed value, and $f(\Psi_j^{(m-1)})$ is the current model likelihood. Set $\Psi_j^{(m)} = \Psi_j^{(m-1)}$ otherwise.

4. Set m = m + 1. Go to step 2.

Because the model parameters (not including the regression coefficients $\beta_0, \ldots, \beta_{n_p}, \beta_t$) for a site are independent of those for all other sites, the updates for these parameters can be carried out in parallel in order to improve computational efficiency. A burn-in period of 1,000 iterations was carried out to ensure the chain reached convergence, followed by a full run 50,000 iterations, which were thinned by a factor of 5 to reduce autocorrelation, giving a resulting posterior sample size of 10,000.

Variable	Estimate	Std. Error	p-value
Intercept	-2.305	0.645	< 0.001
Urban	0.319	0.073	< 0.001
Intersection	1.129	0.046	< 0.001
Signalised	0.493	0.033	< 0.001
Sp. Lim. 30	1.638	0.639	0.010
Sp. Lim. 45	2.050	0.641	0.001
Sp. Lim. 50	1.680	0.641	0.008
Sp. Lim. 60	1.815	0.641	0.005
Sp. Lim. 70	1.404	0.644	0.029
Sp. Lim. 80	1.313	0.646	0.042
MajorIntersection	0.277	0.039	< 0.001
FourLegs	0.435	0.031	< 0.001
MajorVolume	-0.031	0.074	0.680
MinorVolume	0.043	0.007	< 0.001
Year	-0.029	0.005	< 0.001

Table 5.2: Output from the SPF fitted to the Halle dataset

5.6.2 Exploratory Data Analysis

We apply the hotspot prediction model to the Halle dataset outlined in section 5.2. We first build the SPF in order to obtain estimates of μ, for which we use the covariate data from all 734 nodes (not including Volume so as to avoid issues of multicollinearity), and fit a Negative Binomial generalised linear model using the `glm.nb` function within the `MASS` package in `R`. The output from this regression is given in table 5.2.

From table 5.2 we can see almost all (with the exception of MajorVolume) of the covariates are highly statistically significant in estimating the number of collisions at a given node. The lack of significance of MajorVolume could be due to its relatively high correlation with MinorVolume ($r = 0.59$), a possible by-product of how the traffic flows were simulated, and the node locations decided within the traffic model. We further observe that, as expected given the results shown in table 5.1a, there is a negative trend in collision counts, although the strength of this trend is diluted by the relative stagnation of counts in the first 6 years of data. As discussed in section 5.4, we do not expect all nodes to behave exactly as the SPF suggests, with regards to covariate and/or trend effects, and hence include the site and local trend parameters, σ and α, to account for any divergence from this estimate.

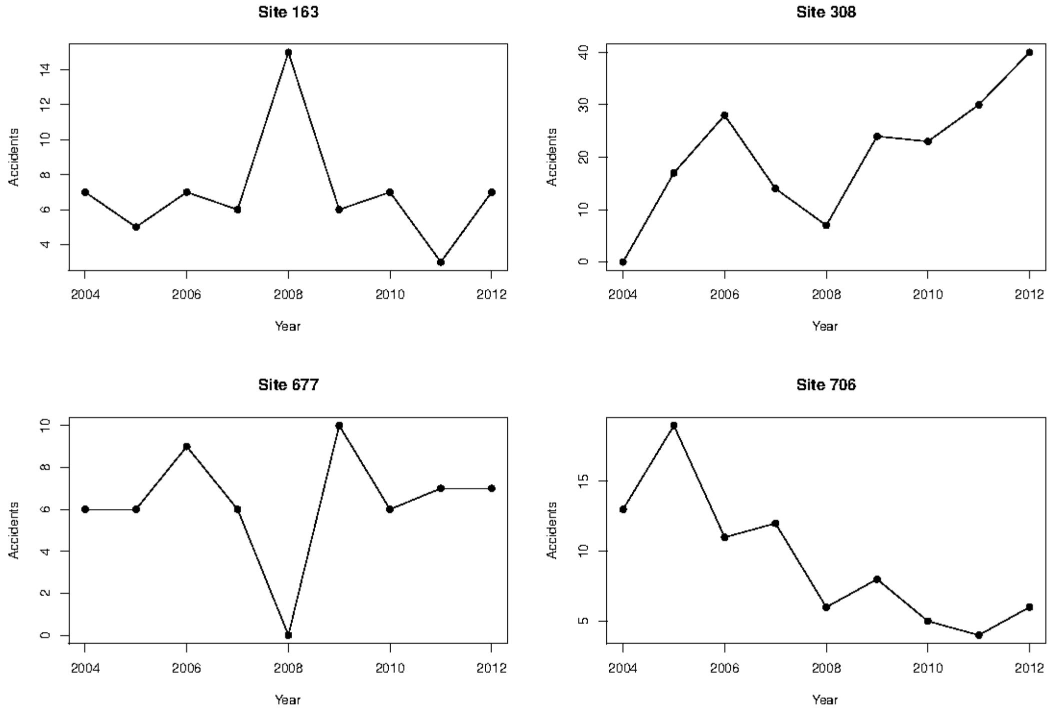


Figure 5.1: Plots of collision counts at 4 nodes from the Halle dataset from 2004-2012.

5.6.3 Model Results

We can now apply the full model to each of the 734 nodes across the Halle network in order to predict future collision counts and thus determine where on the network should be considered future hotspots and thus candidates for treatment. We choose to highlight four nodes in particular so as to best demonstrate the features of the model, which are given in figure 5.1.

From figure 5.1 we observe at nodes 163 and 677, for the year 2008 there is a clear blip which is distinct from the underlying collision rate at each node. It is clearly highly important that each is identified as a blip, as failing to do so in the case of node 163 would potentially lead it to being falsely identified as a dangerous location, and the opposite for node 677. In both cases this would lead to resources being misallocated with potentially dangerous nodes remaining untreated, and treatments being falsely allocated to locations where they will have little to no benefit, as seen in Chapter 2. Furthermore in figure 5.1, we observe clear trends in collision counts at nodes 308 and 706, with node 308 demonstrating a clear increasing trend, and node 706 demonstrating a clear decreasing trend. It is important to acknowledge these trends since the purpose of our model is to estimate the level of safety at locations in the future, we should estimate and extrapolate any trends present at each location. For instance at a location such as node 308, even if its current collision count does not deem it sufficiently dangerous as to warrant treatment, the fact the collision numbers are growing suggests it is likely to become sufficiently dangerous in the future, and so intervention should be made before this occurs. Conversely for sites demonstrating a negative trend such as node 706, even if its current collision counts suggest it should be a candidate for treatment, the fact the collision totals are reducing by themselves, without treatment, may suggest it is more prudent to withhold treatment, at least in the short term, to see if the collision rate returns to an acceptable level without need for intervention. Output from fitting the hotspot model to these nodes is given in figure 5.2

From figure 5.2 we observe that at nodes 308 and 677, the blip values in 2008 have no effect on the estimated collision rate, shown by the estimated collision rate fitting well to the rest of the points but leaving the blip value as an outlier. This is important since it demonstrates any RTM effect will be removed from the model by disregarding any blips which may be present. When comparing the red and blue lines in the plots for nodes 163 and 677, we observe the lines to be virtually parallel, meaning the model does not detect any site specific deviation from the global trend at these nodes (i.e. α will be 0). However for nodes 308 and 706, we observe the red and blue lines have very different gradients, showing the model has detected sufficient site specific deviation from the global trend to fit a significantly different trend at these nodes (in the case of node 308 we will have $\alpha > 0$, and for node 706 $\alpha < 0$). The fact there is no site specific deviation in trend for nodes 163 and 677, where the appears to be little evidence to warrant it from the data, but significantly different trends are fitted to nodes 308 and 706 is encouraging, as it suggests the zero-inflated distribution on α is having the desired effect, restricting where deviations from the global trend are included, but allowing for them where there is clear evidence of site-specific deviations from the global trend.

In figure 5.2 we observe the underlying collision rate λ is extrapolated to become the mean of the predictive distribution, as described in equation 5.5, alongside its 95% prediction interval. We can investigate this in more detail by viewing the full posterior predictive distributions, given in figure 5.3.

From figure 5.3 we observe the predictive probability of observing any number of collisions in the coming year at each of the 4 nodes we are analysing. While we report the posterior predictive mean in the plots in figure 5.2, since we wish practitioners to make decisions based on the underlying level of safety at a site as opposed to individual counts, if a practitioner was solely interested in the number of collisions in the coming year, the posterior mode may be analysed instead. We can also infer the level of confidence in the prediction, for instance the distribution for node 308 is much wider than for the other three nodes, implying we have much less certainty regarding future collision counts

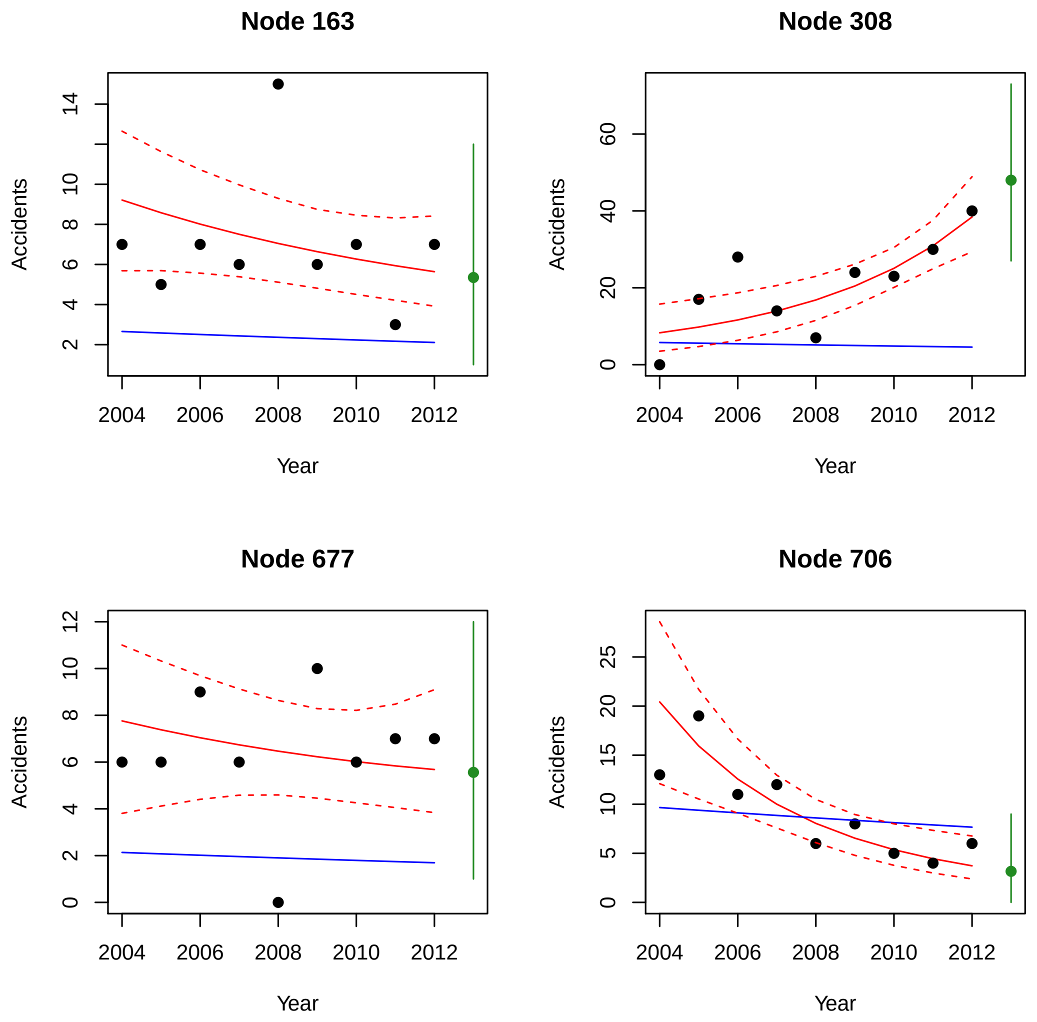


Figure 5.2: Plots displaying output from the hotspot prediction model at 4 nodes from the Halle dataset. Collision counts are given as black dots, the red line and red dotted lines correspond to the estimated value of λ and the 95% credible interval for λ respectively. The blue line corresponds to the fitted values from the SPF for each node, and the green dot with line corresponds to the mean of the predictive distribution, with 95% prediction interval.

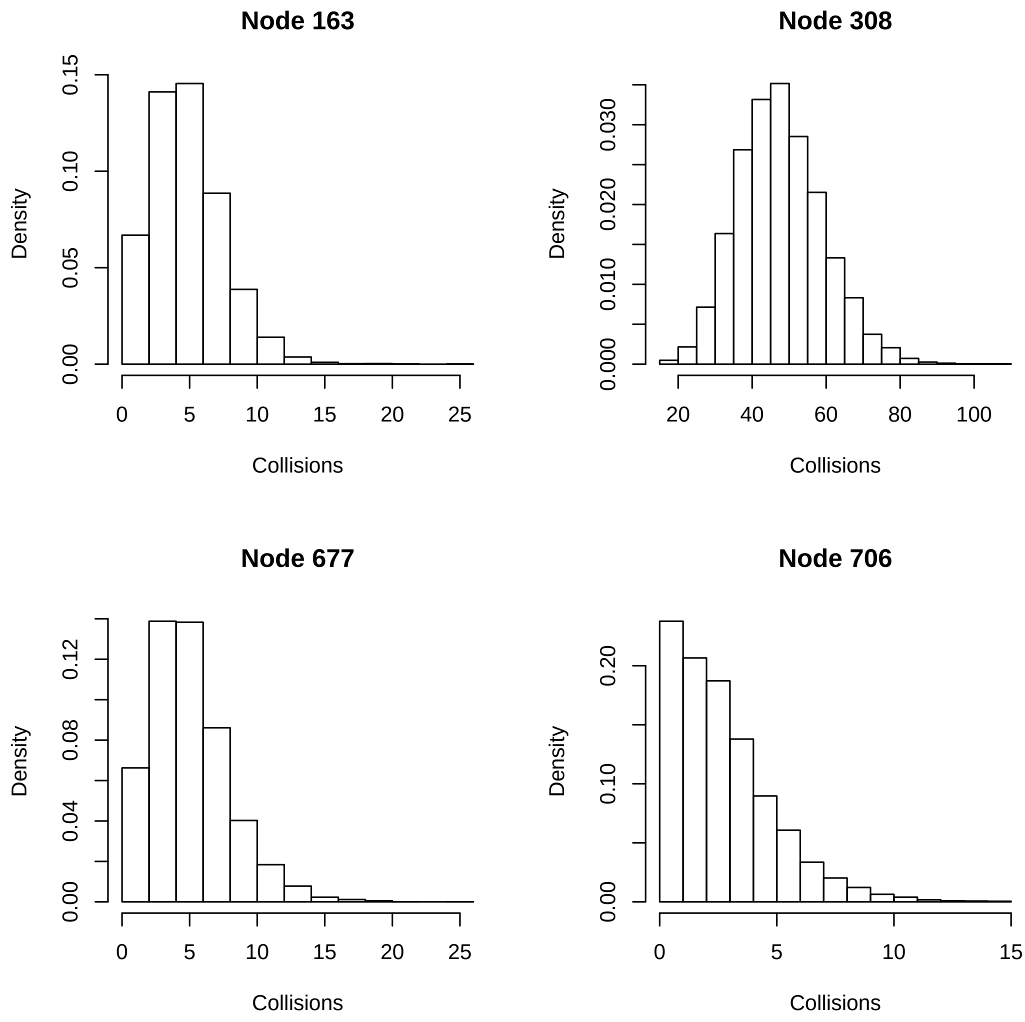


Figure 5.3: Histograms displaying the posterior predictive distributions for the number of collisions in 2013, for 4 nodes in the Halle dataset.

there, conversely node 706 has a rather narrow distribution, implying a much greater level of certainty regarding future counts there. These uncertainties must of course be considered relative to the size of the predictive estimate, with node 308 having a much larger predictive mean, and so it may be expected that is has a large posterior predictive variance relative to nodes with a smaller predictive mean.

5.7 Model Validation

Clearly it is important for any model, particularly for which the primary purpose is prediction, to be validated as fitting real data well. There are numerous statistical metrics with which we can ascertain the goodness of fit of a model in comparison to others, (see Section 5.7.1), however these do not necessarily provide us with information regarding the goodness of fit of a model in isolation. In order to be confident in the predictive capability of our model we should attempt to verify that our model marries well to observed reality in isolation, rather than simply it performing better than some alternatives. A simple, and widely used ([Deublein et al., 2015], [Anastasopoulos, 2016], [Hou et al., 2018]) approach toward this is to remove a section of a dataset to form a “training” dataset, from which we shall predict the remaining “non-training” data. A natural method for this with regards to hotspot prediction, would be to fit the model to the initial $n-1$ years of data, and predict the nth collision count. In the case of the Halle dataset, we fit our model to data from 2004-2011, and predict 2012, giving the results shown in figure 5.4.

In Figure 5.4 we observe a strong positive correlation of 0.858 between the collision counts observed in 2012 for the Halle dataset, and the means of the corresponding posterior predictive distributions for each node. This indicates a good matching between prediction and observation and gives evidence that we can be satisfied our model has fit well to the data, however this alone is not sufficient, since if our model consistently predicted 10 collisions more than were observed would yield a correlation coefficient of 1, but clearly these predictions would be unsatisfactory. Hence we cannot simply solely rely on the correlation coefficient as a measure of model validation, and so we also calculate the mean absolute error (MAE) calculated as,

$$\text{MAE} = \frac{1}{734}\sum_{i=1}^{734} |(y_{i,\text{OBS}} - y_{i,\text{PRE}})|.$$

The choice to use the MAE, as opposed to commonly used alternatives e.g. mean squared error (MSE), is based upon the direct layman interpretation of the MAE output as the “average prediction error produced by the model”, something objects such as the MSE cannot provide. However we should be wary of the MAE’s inability to standardise residuals, thereby meaning our validation metric could be overwhelmingly influenced by a

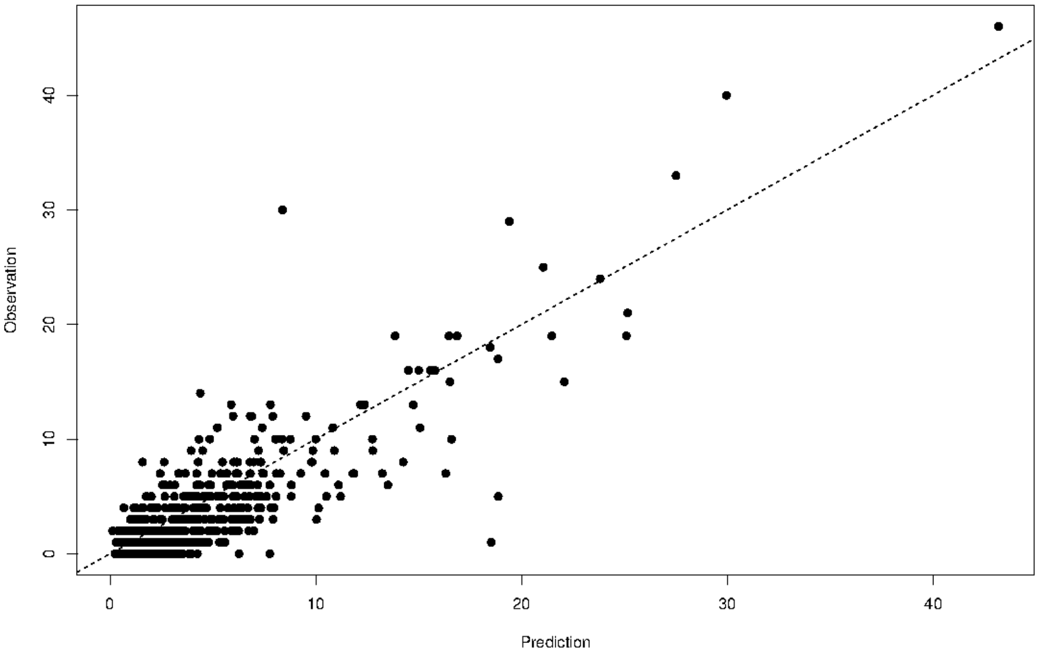


Figure 5.4: A plot comparing predicted and observed collision counts for the 734 nodes in the Halle dataset for the year 2012.

large number of small nodes with low collision counts heavily down weighting the overall MAE. Again as we can see from figure 5.4, the model validation procedure for the Halle dataset gave an MAE value of 1.626, meaning the prediction was typically between 1 and 2 collisions from the observed count at each node. Again this value indicates that the model predictions are generally close to the observed counts, further validating its predictive capability. It is however clear from figure 5.4 that this average is clearly not representative of the model's predictive accuracy at all nodes, where the predictive error is clearly proportional to, and thus increases with, the predicted/observed counts. This is not necessarily a cause for concern however, since as demonstrated in Chapter 2, when a large number of individual collision counts are observed, it is expected that many will be blips which are not representative of the underlying collision rate at their site. An example of this would be node 65, highlighted on figure 5.4 which has a predicted value of less than 10, but an observed count of 30. While normally this would be a cause for concern, suggesting a poor model fit at this node, closer inspection of the historic counts at node 65, given in figure 5.5 show that there has only been a single annual count across the entire dataset greater than 10, and indeed there appears to be a general negative trend displayed up to and included 2011, the final year of the training data. While we cannot yet know for certain, it would appear this observed value for 2012 is highly likely

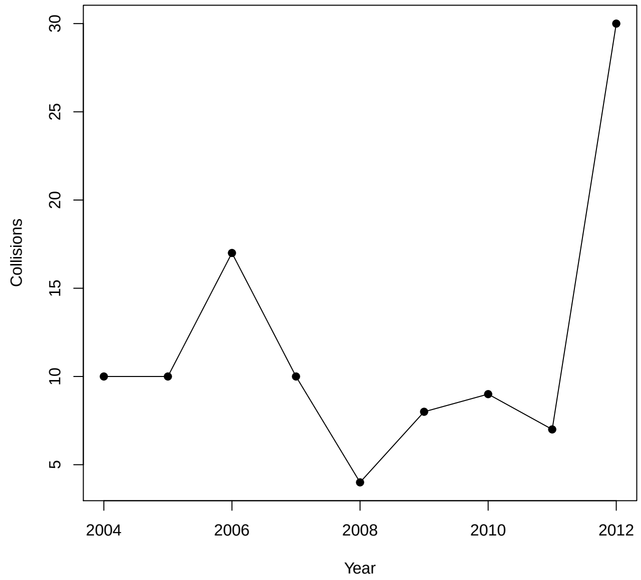


Figure 5.5: Collision counts at node 65 from the Halle dataset, showing a clear outlier in the year 2012

to be a blip, and reduce by itself without need for intervention, and so should not be an indicator of poor model performance. Should future counts remain in this high territory, then clearly treatment should be considered, however this would also be reflected by the model when the new data is included.

We can be satisfied from the above analysis that our model fits the Halle dataset well when we have a training dataset comprising 9 years of data. However clearly not all practitioners will have as much as 8 years of data available, and so it is important to investigate the robustness of the model to fewer years of data in the training dataset. We can investigate this for the Halle dataset using the same strategy as above, by using the data corresponding to years up to 2011, and attempting to predict the corresponding counts for 2012. In order to ensure our model is able to predict adequately, with reduced data available, we shall reduce the size of our training dataset by removing the earliest years. We investigate the predictive validity of the model using a training dataset comprising of 8 years of data as above (i.e. the years 2004-2011), 7 years (2005-2011), 6 years (2006-2011) and 5 years (2007-2011). We then compare the predicted value from these models with their observed counterparts, with the resulting plots given in figure 5.6.

From figure 5.6 we can see there is very little fluctuation in the accuracy of the model predictions, with all 4 plots appearing very similar, showing that the model predictions are highly robust to reduced number of years. Hence we can be satisfied our model fits

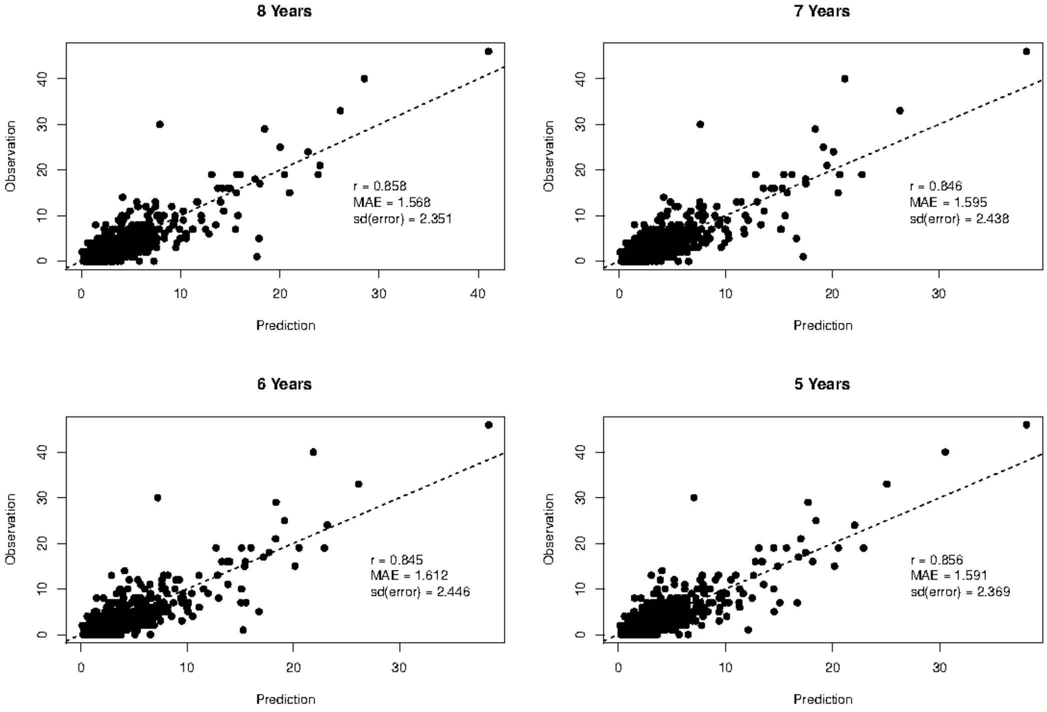


Figure 5.6: A plot comparing predicted and observed collision counts for the 734 nodes in the Halle dataset for the year 2012 using datasets of length 8, 7, 6 and 5 years respectively up to the year 2011.

the data well and generally forms accurate predictions, even with a reduced amount of available data. However there are many other approaches toward hotspot identification, and we should investigate the effectiveness of our model against others which attempt the same analysis.

5.7.1 Methods of Model Comparison

As discussed in Section 5.7 there are many metrics by which we can evaluate model performance. Whereas in Section 5.7 where we were primarily concerned in evaluating the quality of model fit to a specific dataset, here we simply wish to compare different model fits in order to determine the best performing model. There are a variety of methods from which we can choose, across a spectrum of statistical complexity. Perhaps the most commonly used statistical tests for this purpose are the information criteria, typically the Aikake Information Criterion [Akaike, 1998], Bayesian Information Criterion [Schwarz et al., 1978] or Deviance Information Criterion [Spiegelhalter et al., 2002] are applied, with the model minimising the resulting value being deemed to be the best fit to the data. However in the case of models for hotspot identification, specific statistical tests have been developed for the purpose of method comparison, namely: the site consistency test, the method consistency test, the total rank differences test, and the total score test [Montella, 2010]. A common feature of these approaches is that they assume hotspots are identified as top $\alpha\%$ of all sites, when sites are ordered by (modelled) collision rate. This is a relatively restrictive assumption, as it does not take into account the various other ways hotspots may be defined (for instance, as discussed in Section 5.6.3 hotspots can be simply categorised as any site exceeding certain criteria, in this case a given collision rate). In practice, rather than set out to treat a given percentage of potential hotspots, it is more likely a practitioner will have a budget for a particular number of treatments, and so will select this many hotspots (which can be trivially converted into a percentage). These methods also require each method of analysis to be applied in two distinct time periods and compare the hotspots identified by each method in both time periods. This presents practicality issues since, as has been referred to throughout this thesis, we cannot guarantee or assume an abundance of data available for analysis and failing to ensure a good amount of data for both time periods being analysed makes it much more difficult to achieve a good model fit (increasingly so for more complicated models with more parameters) and leave such analyses susceptible to factors such as the RTM effect outlined in Chapter 1. The four methods in question are outlined as follows:

- **Site Consistency Test (SCT)**: The test statistic for the SCT for method m is defined as:

$$\text{SCT}_m = \frac{1}{n}\sum_{i=1}^{n} y_{h_i,t+1},$$

where $y_{h_i,t+1}$ is the collision count at hotspot i in time period $t+1$, where the hotspot has been identified at time t, and n is the number of hotspots identified. This test is designed to analyse the ability of a method to consistently identify dangerous sites across multiple time periods, hence a method which identifies sites as hotspots which go on to have high collision counts in the future will be favoured by the SCT statistic. Larger values of SCT are preferred, with SCT being defined on the range $0 \leq \text{SCT} < \infty$.

- **Method Consistency Test (MCT)**: The test statistic for the MCT for method m is defined as:

$$\text{MCT}_m = \{h_1, h_2, \ldots, h_n\}_{m,t} \cap \{h_1, h_2, \ldots, h_n\}_{m,t+1},$$

where $(h_1, h_2, \ldots, h_n)_{i,t}$ is defined as the set of hotspots $\boldsymbol{h}$ identified by method m in time period t. The rationale behind this test is that if a site is truly a hotspot then it should remain a hotspot throughout time periods i and hence a method should consistently identify it as such. Methods which consistently identify the same, or similar groups of sites as hotspots should be preferred and so are favoured by the MCT statistic. Larger values of MCT are preferred, with MCT being defined on the range $0 \leq \text{MCT} \leq n$.

- **Total Rank Differences Test (TRDT)**: The test statistic for the TRDT for method m is defined by:

$$\text{TRDT}_m = \sum_{i=1}^{n} |R_{m,i,t} - R_{m,i,t+1}|$$

where $R_{m,i,t}$ is the ranked position of hotspot i at time period t in terms of collision count according to method m. So for example if hotspot i had the greatest collision count in time period t but only the third highest in time period $t+1$ it would have $|R_{m,i,t} - R_{m,i,t+1}| = |1 - 3| = 2$. This test is designed to investigate consistency in relative hotspot severity, rather than simply whether or not sites were classed as hotspots, since clearly methods which give the most consistent picture of the most dangerous hotspots on a network will be preferred as to ensure optimal allocation of treatment resource. Unlike the SCT and MCT, smaller values of TRDT are preferred, with TRDT being defined on the range $0 \leq TRDT \leq \frac{1}{2}n^2$.

- **Total Score Test (TST)**: The TST is a combination of the three above tests, with test statistic defined as:

  $$\mathrm{TST}_m = \left(\frac{\mathrm{SCT}_m}{\max_m(\mathrm{SCT})}\right) + \left(\frac{\mathrm{MCT}_m}{\max_m(\mathrm{MCT})}\right) + \left(1 - \frac{\mathrm{TRDT}_m - \min_m(TRDT)}{\max_m(\mathrm{TRDT})}\right).$$

  This test combines the three test discussed above into a single overall metric, which provides equal weighting to each test (although this can easily be modified into a weighted sum if preferred). As discussed above, large values of SCT and MCT with small values of TRDT are preferred, and so large values of TST indicate a method performs consistently overall, with TST being defined on the range $0 \leq TST \leq 3$ (some authors multiply by a factor e.g. $\frac{1}{3}$ or $\frac{100}{3}$ to rescale TST however this is purely a matter of personal preference and does not affect the analysis conclusion).

A comparison study was carried out in [Guo et al., 2019], comparing the hierarchical Bayesian hotspot model described here in Chapter 5 with an EB approach to hotspot identification and a non-parametric approach referred to as the crash rate method which simply uses observed collision rates at locations as a means to rank hotspots without appealing to any external sources of information. Comparisons between the 3 methods were carried out using the four tests (SCT, MCT, TRDT and TST) on a subset of the Halle dataset. Model consistency was assessed by splitting the years of observations into 4 consecutive sets of 2 years (04-05, 06-06, 08-09, 10-11), enabling 3 comparisons to be made per test. These tests were repeated for classifying the top 2.5%, 5% and 7.5% of sites as hotspots, meaning 36 individual consistency tests were carried out in total. Of those 36 tests, the Bayesian hierarchical model performed best for 35 out of the 36 tests, providing strong evidence that it provides an improved method for informing hotspot identification over EB and naive ranking approaches.

5.8 Hotspot Identification

The model developed in this chapter allows a road safety practitioner to make detailed inferences regarding the collision rate at various points on the network. The model however deliberately stops short of specifically labelling particular locations as being collision hotspots, largely due to the highly subjective nature of how a hotspot is defined, which can vary significantly between countries, local authorities and practitioners. Some practitioners may only consider a location to be a hotspot based on the number of collisions of a particular severity, or due to a particular set of causes, which the extended hotspot model is able to identify (see Chapter 6). Even within the output from the model described in this chapter however, there are competing metrics by which a practitioner may wish to define a hotspot, with each parameter in the model potentially indicative of a site requiring treatment. Perhaps the most obvious initial statistic to consider would be the predicted number of collisions in the coming year $\mathbb{E}\left(y_{i1}\right)$, where clearly locations expected to have high numbers of collisions will be of greater concern to practitioners than those expected to have fewer collisions. This does not tell the whole story however, since a location which is predicted to have a high number of collisions, but has a noticeably declining trend in collision rate $\alpha_i < 0$, may soon no longer be considered among the more dangerous locations on the network without need for intervention, conversely a location which currently is not predicted to have a particularly high collision count, but has an increasing trend in collisions $\alpha_i > 0$, may require proactive preventative measures to prevent the level of danger worsening. Finally there is the site effect, σ_i to consider, which describes a site's collision rate relative to that expected from the network via the SPF. Even if a location is predicted to have a high number of collisions, if is still displaying typical behaviour for a site on the network with its set of covariates (i.e. if the collision rate is close to the value predicted by the SPF, and so $\sigma \approx 0$) then road safety treatments may not be effective at reducing collisions here, but rather infrastructural changes may be needed to reduce the covariate effects $\boldsymbol{\beta}$ across the network. Hence special priority may be given to sites with the largest values of σ, i.e. those with the greatest potential for safety improvement, as their collision rate is most abnormal for the network and so may be more likely to be improved by a safety countermeasure.

Chapter 6

Extensions to the Hotspot Prediction Model

Notation

Below is a summary of the notation used in this chapter. Much of the notation carries over from Chapter 5 and has the same meaning here.

Notation	Meaning
i	Site indicator, $i = 1, \dots, n$
j	Covariate indicator, $j = 1, \dots, P$
t	Time period indicator, $t = n_y - 1, \dots, 0$
$x_{i,j}$	The value of covariate j at site i
$y_{i,t}$	Observed collision count in time period t at site i
λ_i	Underlying collision rate at site i
σ_i	Site effect effect at site i
α_i	Site specific trend effect at site i
α_{N_i}	Zero-inflation component of α_i
α_{Z_i}	The Normal distribution component of α_i
τ_i	Variance inflation parameter for site i
μ_i	Fitted estimate from the SPF for site i
$\beta_{i,j}$	SPF regression coefficient for covariate j (at site i)
ϕ_s	Seasonal effect of season s, $s = 1, \dots, S$
$\pi_{i,t,k}$	The probability of a given collision at site i in time period t being of severity k
k	Severity indicator $k = 1, \dots, K$
$\theta_{k,i}$	The latent threshold at site i to determine if a collision reaches severity $k + 1$
f	Collision factor/type indicator $f = 1, \dots, F$

6.1 Accounting for Seasonality and Spatial Trend

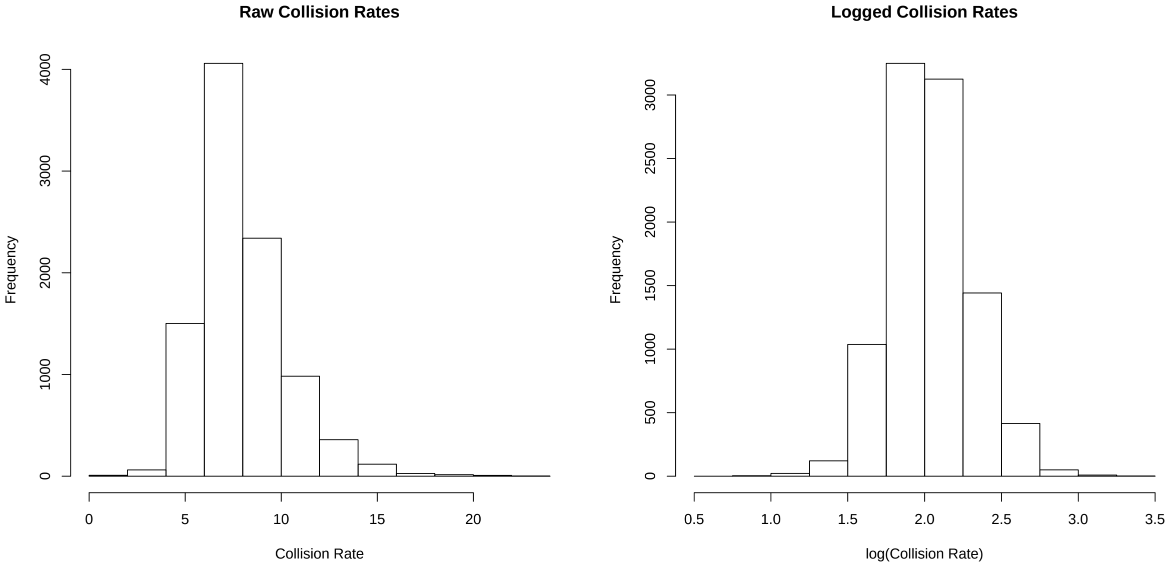
The hotspot prediction model described thus far in sections 5.3 and 5.4 is designed to analyse data on an annual (or larger) scale, and hence does not anticipate any seasonal patterns to form in the data. While this would be acceptable for practitioners making their decisions on an annual basis (as road safety practitioners do), it is not satisfactory for those who wish to make decisions on a finer scale. This is since the current model only fits a linear trend to the data and so does not account for cyclical patterns, which we would expect to be present in data taken on a seasonal level. This would likely be due to recurring effects on a road network for instance, significant seasonal weather patterns affecting the difficulty of driving, or possibly significant trends in the number of non-local drivers on the network due to tourism/events taking place (e.g. festivals), which in turn could lead to patterns in the number of road users driving while impaired etc. Clearly in order to account for this we require an additional component to be included in our estimate of the underlying collision rate, λ, in order to capture this seasonal variation.

An additional limitation of our current model, is that each element of the site effect parameter vector $\boldsymbol{\sigma}$ is evaluated independently of the others, i.e. there is no sharing of information pertaining to the site effect between sites. This is sub-optimal when we consider the possibility that factors contributing to sites being more or less dangerous than we expect are likely to be similar to those around them (changing socio-demographic areas, vulnerability to adverse weather conditions, popular areas for tourists who are new to the system etc), i.e. there is reason to suspect the site effects may be spatially correlated. Allowing for this spatial dependence between site effects should not only improve the fit of the model by allowing information to be shared between sites, but also could potentially provide more information to practitioners by developing a picture of relative risk on a more mesoscopic level, as opposed to the micro, site-specific, scale previously adopted.

As discussed at various points it is the aim of this thesis to develop methodologies which are as versatile as possible, and as such retain their utility to even more limited datasets. By incorporating seasonal and spatial structure into the model, and hence sharing of information between datapoints, we further enhance the model's ability to evaluate seasonal and spatial effects, even when data is limited.

6.1.1 The Florida Panhandle

A good example of the kind of location which would greatly benefit from, if not require, a spatial and seasonal component being included is the state of Florida, U.S.A., which in addition to covering a large geographic area, also undergoes wildly changing weather patterns between seasons. We can carry out an initial pre-analysis purely to investigate



(a) Raw collision rates

(b) Logged collision rates

Figure 6.1: Histograms of raw and logged collision rates for the Florida dataset.

the potential presence of seasonal and spatial correlations within the data, i.e. we take the model for the current year given in equation 6.1, but only consider seasonal and spatial effects (removing any dependence on covariates and trend). In this case the response variable is not collisions at locations, but rather collision rates on segments (collisions per km). Hence the Poisson/Negative Binomial distribution used for the count data in Section 5.3 would no longer be appropriate here, and so must be replaced by a continuous analogue. Plotting the collision rates in Figure 6.1a, reveals the typical slight positively skewed density associated with crash data, which is clearly bounded below by 0 since they are rates. Plotting the logged rates shown in Figure 6.1b shows a more symmetric distribution, approximately Gaussian in shape.

For convenience we shall therefore model the logged collision rates, allowing us to use a conjugate Normal-Normal structure,

$$y_{i,s} \sim N\left(\lambda_{i,s}, \tau_{i,s}^{-1}\right), \tag{6.1}$$

$$\lambda_{i,s} = \sigma_i + \phi_s, \tag{6.2}$$

$$\sigma_i \sim N\left(m_\sigma, t_\sigma^{-1}\right), \tag{6.3}$$

$$\phi_s \sim N\left(m_\phi, t_\phi^{-1}\right) \tag{6.4}$$

$$\tau_{i,s} \sim Ga(g_\tau, h_\tau), \tag{6.5}$$

hence the prior distribution for the mean of the collision rates can be given

$$\lambda_{i,s} \sim N\left(m_\sigma + m_\phi, (t_\sigma + t_\phi)^{-1}\right).$$

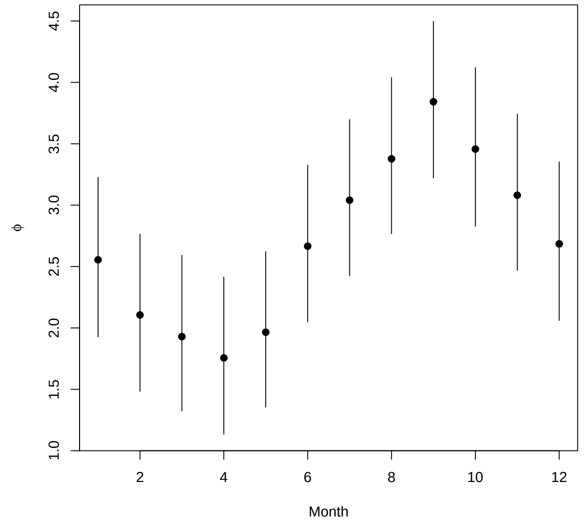
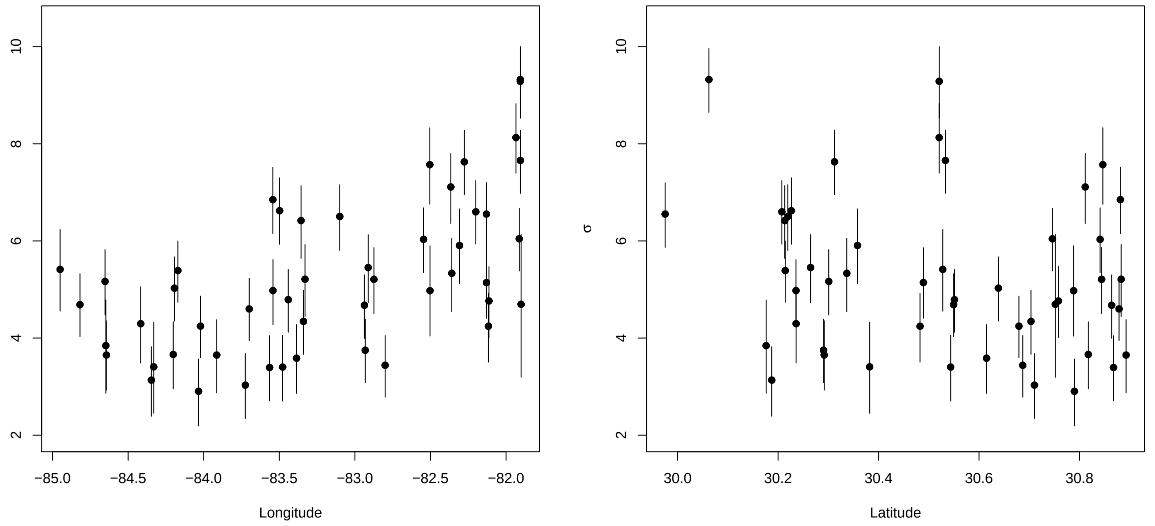


Figure 6.2: A plot of posterior means for ϕ_s against month s

In this modelling representation, $y_{i,s}$ is the log of the observed collision rate, $\lambda_{i,s}$ and $\tau_{i,s}$ are the mean and variance for the distribution of collision rates at site i in month s, and σ_i and ϕ_s are the spatial and seasonal effects at site i and season s respectively. Given the conjugate model structure we can employ a Gibbs sampler to sample from the posterior distributions $\pi\left(\sigma_i|\boldsymbol{\sigma}_{-i},\boldsymbol{\phi},\boldsymbol{y}\right)$, $\pi\left(\phi_s|\boldsymbol{\sigma},\boldsymbol{\phi}_{-s},\boldsymbol{y}\right)$ for $i=1,\ldots,n$ and $s=1,\ldots,S$.

We apply the model to the Florida dataset in which we have $n=52$ zones of data with monthly collision rates, meaning we have $S=12$ seasons. We initialise each parameter at its prior mean and run the MCMC algorithm for $M=1000000$ iterations after discarding the first 1000 iterations for burn-in and thinning by 100 to remove autocorrelation. Posterior means for $\boldsymbol{\phi}$ against month are given in Figure 6.2 and posterior means for $\boldsymbol{\sigma}$ against longitude and latitude of traffic analysis zone (TAZ) centroid are given in Figure 6.3.

We plot the posterior means for $\boldsymbol{\sigma}$ against longitude and latitude of the TAZ centroid, given in Figure 6.3. From Figure 6.2 we see a clear fluctuation in posterior mean for ϕ_s between months, with lower collision rates associated with months earlier in the year, and higher collision rates associated with the late summer months, possibly coinciding with the tropical storm season frequently affecting the south eastern coast of the USA. In Figure 6.3 we do not observe much of an effect of changing latitude on posterior mean for σ_i, however there appears to be a clear increase in collision rate with increasing longitude (moving from west to east, from central Florida towards the coast), again possibly due to climatic effects since coastal regions are normally those most affected by extreme weather effects. While we analyse the Florida dataset here to demonstrate the potential for seasonal and spatial effects in collision data, the dataset is very limited beyond this and so we shall



(a) Posterior means of σ against longitude (b) Posterior means of σ against latitude

Figure 6.3: Plots of posterior means of σ against longitude and latitude for the Florida dataset.

return to the Halle dataset to demonstrate expansions to the hotspot prediction model.

6.2 Updating the Hotspot Model

We wish to extend the model developed in Chapter 5 to account for subtleties in the data which are ignored by only modelling raw annual collision counts with no structure applied to the spatial effects of the data. In order to do this we shall consider counts which occur on a finer scale than annually, thereby presenting the possibility of seasonal effects affecting collision counts; spatial structure which may be present in the data by fitting structures to the site effect terms in our model; the ability to inform our prior beliefs with respect to the SPF coefficient vector $\boldsymbol{\beta}$ using external sources of exchangeable data so as to better inform our analysis.

6.2.1 Data Augmented Priors

When formulating our prior beliefs with respect to the regression coefficients $\boldsymbol{\beta}$ in Chapter 5, we stated that unless expert prior knowledge was available, we would adopt the position of prior ignorance, and so would fit independent, diffuse prior distribution to each element β_j of $\boldsymbol{\beta}$,

$$\beta_j \sim N\left(0, 10^2\right) \qquad j = 1, \ldots, P.$$

An improvement on this approach in absence of expert prior knowledge is to appeal to supplementary, external data which we can use to inform our beliefs with regards to covariate effects. This practice, known as forming data augmented priors (DAPs) [Ntzoufras, 2011], involves obtaining estimates of covariate effects (and covariances) using external data, and using that information to inform prior beliefs regarding the covariate effects of the data we wish to analyse. If we denote the fitted effect of covariate j to be β'_j then a DAP for β_j would be

$$\beta_j \sim N\left(\hat{\beta}'_j, k \times s.e.\left(\hat{\beta}'_j\right)^2\right), \tag{6.6}$$

where $\hat{\beta}'_j$ denotes the estimated value of β'_j, for example a maximum likelihood estimated obtained using `glm.nb`, $s.e.\left(\hat{\beta}'_j\right)$ denotes the standard error of this estimate, and k is a constant by which we multiply to account for uncertainty between the supplementary dataset, and the dataset on which the analysis will take place. It is possible to generalise this further if it is suspected there may be strong correlation between parameters, by adopting a multivariate prior distribution for the entire covariate effect vector (minus the intercept which by definition cannot be correlated with any parameter), here denoted $\boldsymbol{\beta}_{-0} = (\beta_1, \ldots, \beta_P)$. Hence a multivariate DAP would be,

$$\begin{aligned}\beta_0 &\sim N\left(\hat{\beta}_0, k \times s.e.\left(\hat{\beta}_0\right)\right)\\ \boldsymbol{\beta}_{-0} &\sim N_{P-1}\left(\hat{\boldsymbol{\beta}}_{-0}, \hat{\Sigma}_\beta\right)\end{aligned}$$

where $\hat{\Sigma}_\beta$ is the fitted variance-covariance matrix of the estimated $\hat{\beta}'_j$ coefficients, where the i, jth element would be given $\hat{\Sigma}_{\beta(i,j)} = Cov\left(\hat{\beta}'_i, \hat{\beta}'_j\right)$.

We formulate DAPs for the Halle data by randomly selecting 50 sites of our data to run a hotspot analysis on, retaining the 684 remaining sites as supplementary training data for the DAP. As demonstrated in Chapter 3 it is important when using supplementary data to inform an analysis, that the supplementary data is sufficiently exchangeable with the data on which the main analysis shall be run. In this case we carry out the permutation tests described in Section 3.4 on the covariates used to form the SPF, as well as the Mahalanobis distance on the combined covariate vector, for the supplementary and analysis datasets. In this case the true p-value requires calculating $\binom{734}{50} \approx 1.15 \times 10^{78}$ combinations, and so again we obtain an approximate p-value, $\hat{p}$, via a Monte Carlo permutation test (as in Chapter 3) with $N = 100000$ simulations, with results given in Table 6.1. From Table 6.1 we see that for all covariates and the Mahalanobis distance we have $p > 0.05$ suggesting we do not have evidence to reject exchangeability for any covariate and so can be satisfied the supplementary pool chosen will provide suitable exeternal information to inform our prior befliefs for $\boldsymbol{\beta}$ for the analysis pool. We note that although the pools are exchangeable they are not identical, and so retain the multiplicative factor k in our prior standard errors for

	x_1	x_2	x_3	x_{4A}	x_{4B}	x_{4C}	x_{4D}	x_{4E}	x_5	x_6	x_7	x_8	Mahal
$\hat{p}$	0.61	0.84	0.74	0.50	0.71	0.34	0.50	0.25	0.89	0.61	0.13	0.62	0.55

Table 6.1: Table giving approximate permutation test p-values for each covariate in the Halle dataset and the Mahalanobis distance for the supplementary and analysis datasets

Covariate	Estimate	Std. Error	p-value
β'_0	49.09	6.13	1.1×10^{-15}
β'_1	0.38	0.05	6.4×10^{-15}
β'_2	1.19	0.04	$< 2 \times 10^{-16}$
β'_3	0.46	0.02	$< 2 \times 10^{-16}$
β'_{4A}	0.35	0.03	$< 2 \times 10^{-16}$
β'_{4B}	-0.06	0.03	0.05
β'_{4C}	0.11	0.03	3.91×10^{-4}
β'_{4D}	-0.26	0.05	3×10^{-7}
β'_{4E}	-0.31	0.06	7.6×10^{-8}
β'_5	0.31	0.02	$< 2 \times 10^{-16}$
β'_6	0.42	0.02	$< 2 \times 10^{-16}$
β'_7	0.03	4.6×10^{-3}	2.1×10^{-12}
β'_8	0.05	3.1×10^{-3}	$< 2 \times 10^{-16}$
β'_9	-0.03	3.1×10^{-3}	$< 2 \times 10^{-16}$

Table 6.2: Table showing output from `glm.nb` in R showing maximum likelihood estimates, with associated standard errors and p-values from tests of significance, for the covariates in the supplementary Halle dataset.

the elements of $\boldsymbol{\beta}$. Given a lack of significant correlation between elements of $\hat{\beta}'_j$ we choose to employ independent DAPs for $\boldsymbol{\beta}$ in the main analysis. Output from fitting a Negative Binomial GLM to the supplementary data using `glm.nb` in R, is given in Table 6.2.

From Table 6.2 we observe that each variable has p-value of less than 0.05 suggesting statistical significance (with the exception of x_{4B} corresponding to the indicator for a 50km/h site). We additionally observe very small fitted standard errors, which is to be expected given the large amount of data used to train the model. To avoid overly specific priors we therefore use a multiplicative factor $k = 3$ when forming the prior standard deviations for β_j.

6.2.2 Modelling Seasonal Effects

As discussed in sections 6.1 and 6.1.1, we have reason to wish to incorporate a seasonal effect into the hotspot prediction model. This seasonal effect can be taken on a variety of scales, e.g. monthly/bi-monthly/quarterly, at the practitioner's discretion, with the model structure remaining broadly the same, simply containing a set of indicator parameters corresponding to each season within the year. Since we treated the seasonal effects as constant throughout the data this is analogous to fitting a seasonal random intercept term to our model. We define $\boldsymbol{\phi}$ to be the vector of seasonal effects, meaning if we define S to be the number of seasonal periods in our data, we have

$$\boldsymbol{\phi} = (\phi_1, \ldots, \phi_S)^{\mathrm{T}}.$$

We note these seasonal effects are not site specific, and instead reflect the global effect of each season across the network. This is done so as to allow information to be shared across the different sites and thus improve the precision of the estimates of $\boldsymbol{\phi}$, which becomes all the more important in datasets where the number of years is low. The seasonal effects can either be modelled independently, or with a dependence structure built in between differing seasons. A standard approach towards fitting independent season effects would be,

$$\phi_s \overset{\text{iid}}{\sim} N(0, 10), \qquad s = 1, \ldots, S,$$

where each ϕ_s corresponds with the effect of being in season s on collision rate. Here we shall assume prior ignorance, hence the vague Normal distributions assigned to each element of $\boldsymbol{\phi}$, however clearly any expert practitioner could adjust these according to their own prior beliefs. The decision to assign independent seasonal effects has varying merit, depending largely on the value of S. In the case where $S = 2$ or 3, the decision to keep seasonal effects independent seems sensible, since it would be expected for conditions to vary quite dramatically between consecutive periods of 6 or 4 months (respectively), and so it is unlikely including this information would better inform our beliefs regarding the current season. In the case for $S > 3$ however, we would expect to see in most cases, a correlation emerge between the level of safety in consecutive seasonal periods, and so including information regarding the surrounding seasons should provide more useful information regarding the effect in the current season. One such method to inform the effect of one seasonal effect using surrounding, consecutive seasonal effects, is conditional autoregressive modelling.

6.2.3 Conditional Autoregressive Models

Conditional autoregressive (CAR) models [Gelfand and Vounatsou, 2003] are conditional distributions which inform the prior beliefs regarding an element of a parameter vector, using neighbouring consecutive elements of the same vector. These can take various forms but are generally a useful modelling specification when it is believed there will be significant correlation among consecutive parameters in a parameter vector, in this case we believe this to be likely in the case of the parameter vectors corresponding to the seasonal effects, $\boldsymbol{\phi}$.

CAR models can be thought of as a special, discretised, case of the kernel density models we shall investigate in Section 6.2.4, as they do not account for varying distances between points, and so the weighting provided to each other parameter in the parameter vector follows a discrete structure. This discretised structure lends itself well to the modelling of seasonal variation, since we fix our seasons to be a constant distance apart, and so we shall use a CAR model to inform our beliefs regarding elements of our seasonal effect parameter, $\boldsymbol{\phi}$. Since the data we wish to analyse is given on a monthly scale, we shall use a lag 1 autoregressive CAR model with weight 0.5, so in effect we have,

$$\phi_s \sim \begin{cases} N\left(\frac{1}{2}\left(\phi_{s-1}+\phi_{s+1}\right), 10^2\right), & 2 \leq s \leq 11 \\ N\left(\frac{1}{2}\left(\phi_2+\phi_{12}\right), 10^2\right), & s = 1 \\ N\left(\frac{1}{2}\left(\phi_1+\phi_{11}\right), 10^2\right), & 12. \end{cases}$$

The choice of model for a seasonal parameter vector, $\boldsymbol{\phi}$ clearly depends upon the choice of S, since for instance if we were to take $S = 2$, with seasons covering the summer and winter months, it would not make sense to try and inform the summer parameter using information from the two surrounding winter seasons, and so a CAR model may not be appropriate. Likewise should our data be on a finer scale, e.g. weekly, it may be preferable to use a higher lag autoregressive structure, i.e. include more than simply the most immediate surrounding observations.

In addition to the seasonal effect vector $\boldsymbol{\phi}$ we can utilise a CAR structure to inform our beliefs regarding the parameter vector for spatial effects $\boldsymbol{\sigma}$. This seems intuitive since there is a strong possibility the heterogeneity accounted for by the σ parameter could be due to unobserved effects which vary geographically, e.g. exposure to climate, changing socio-economic levels, sites being in a poorly maintained local area etc. Such an effect becomes more pronounced when we analyse areas on a more mesoscopic level, i.e. grouping locations into areas (for instance the TAZs commonly used in the United States) rather than on a site level. Analysing data on a zonal level makes it far easier to meaningfully employ the key concept of CAR modelling, that our beliefs regarding a location should be influenced by neighbouring locations. Clearly it would be difficult to meaningfully specify what constitutes a neighbouring location should we carry out an analysis at a site specific level, however at a zonal level it comes much more straightforward, with any 2 zones which share a border being classed as “neighbours”. From this we can therefore express a general CAR model for spatial effects [Xie et al., 2014], [Papadimitriou et al., 2013], [Quddus, 2008a],

$$\sigma \sim N\left(\frac{\Sigma_{i\neq j}\omega_{ij}\sigma_j}{\Sigma_{i\neq j}\omega_{ij}}, 10^2\right)$$

where ω_{ij} represents the proximity indicator, for zones i and j,

$$\omega_{ij} = \begin{cases} 1, & \text{if zones } i \text{ and } j \text{ are neighbours} \\ 0, & \text{otherwise.} \end{cases}$$

This approach is a simple extension of that utilised for $\boldsymbol{\phi}$ previously, where now our prior mean is the mean of the site effects of all neighbouring zones, as opposed to in the seasonality case where the number of neighbours was fixed to be 2.

The advantages of using a CAR approach when modelling spatial effects are that it allows for information sharing between geographically similar sites, which as described above, we have reason to believe could be correlated. The form of the prior distribution is simple and intuitive, and allows for flexibility in the number of neighbouring zones, meaning it can be applied to virtually any zonal structure. The binary weighting structure involved can be considered overly rigid though, with any zones other than the immediate neighbours deemed to provide zero information regarding the spatial effect at our zone of interest. Furthermore, the equal weighting involved, assumes all neighbouring zones provide the same amount of information, which may be considered unrealistic, particularly considering zones can often vary significantly in size.

While dealing with zonal data makes improvements upon this CAR approach difficult, should we consider site data we can overcome these disadvantages quite easily, by replacing our CAR prior, with one formed by a kernel density smoother.

6.2.4 Kernel Density Smoothers

Kernel density smoothers (KDEs) are in many ways a generalisation of the CAR models discussed in Section 6.2.3, since in the KDE framework, each site is provided with its own (normally) non-zero weighting. Hence we have, for a vector $\mathbf{x}$,

$$x_i \sim N\left(\sum_{j\neq i} w_{ij}x_j, \epsilon_x^2\right)$$

where w_j is the weight associated with element j, obtained by a given pre-defined weight function, and ϵ_x^2 is the prior uncertainty associated with the vector $\mathbf{x}$.

It makes little sense to apply a KDE approach to estimate seasonal effects since these are discretised intervals already and so are better suited to a CAR approach. However in the case of spatial effects, with the concept being that we believe sites situated closely together geographically are likely to be correlated, it would make sense to employ a KDE approach with the vector of weights, $\boldsymbol{w}$ a function of the distance from the site in question,

$$\begin{aligned} \phi_i &\sim N\left(\sum_{j \neq i} w_{ij}\phi_j, \epsilon_\phi^2\right), \\ w_{ij} &= f\left(d_{ij}\right), \end{aligned}$$

where d_{ij} is the distance between sites i and j. Here we shall employ the Gaussian weight function to form our kernel density estimator, and hence obtain,

$$\begin{aligned} \phi_i &\sim N\left(\sum_{j \neq i} w_{ij}\phi_j, \epsilon_\phi^2\right), \\ w_{ij} &= \exp\left(\frac{d_{ij}^2}{b}\right), \end{aligned}$$

where b is the bandwidth distance. As documented in literature, see for example [Yu et al., 2014b], the choice of weight function is less important relative to the value of the bandwidth. Adopting a Bayesian approach towards this, including the bandwidth as a variable in the MCMC analysis, allows the data to determine the most appropriate value of the bandwidth. Furthermore, adopting a Bayesian approach allows for proper inclusion of uncertainty regarding the bandwidth, which would be lost in some non-Bayesian approaches which use, for example cross-validation techniques, to find a single value of the bandwidth which is treated as fixed. Again this would lead to over optimistic estimates of uncertainty with regards to the spatial weights, which propogates through as to over optimistic estimates of uncertainty regarding the spatial effects $\boldsymbol{\sigma}$ and hence of the collision rate $\boldsymbol{\lambda}$.

6.2.5 Application of the Extended Model

Combining the elements discussed thus far in Section 6.2 we hence obtain our extended hotspot model,

$$
\begin{aligned}
Y_{i,s,t}|\lambda_{i,s}(t) &\sim \begin{cases} Pois\left(\lambda_{i,s}(t)\right), & t=0 \\ NegBin\left(\text{Mean} = \lambda_{i,s}(t), \text{Variance} = \lambda_{i,s}(t)c_i(t)\right), & t<0, \end{cases} \\
\lambda_{i,s}(t) &= \exp\left(\mu_i(t) + \sigma_i + \phi_s + \alpha_i t\right), \qquad i=1,\ldots,n \\
\mu_i(t) &= \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_P x_{i,P} + \beta_t t, \\
\sigma_i &\sim N\left(\sum_{j\neq i} w_{ij}\sigma_j, \epsilon_\sigma^2\right) \\
\phi_s &\sim \begin{cases} N\left(\frac{1}{2}\left(\phi_{s-1}+\phi_{s+1}\right), 10^2\right), & 2 \le s \le S-1 \\ N\left(\frac{1}{2}\left(\phi_2+\phi_S\right), 10^2\right), & s=1 \\ N\left(\frac{1}{2}\left(\phi_1+\phi_{S-1}\right), 10^2\right), & S. \end{cases}
\end{aligned}
$$

We adopt data augmented priors for the regression coefficients $\boldsymbol{\beta}$,

$$\beta_j \sim N\left(\hat{\beta}'_j, k \times s.e.\left(\hat{\beta}'_j\right)^2\right)$$

where we take $k=3$ to account for uncertainty between the supplementary and analysis datasets, as described in Section 6.2.1. We impose a Gaussian weight structure to form our KDE in the expression for the site effect σ_i,

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{b}\right)$$

where d_{ij} is the geographic distance between sites. We choose d_{ij} to be the Euclidean distance between sites (as opposed to link distance) since the effects we expect to capture are geographic/climatic, and so should be captured by geographic distance measures. Of course other distance measures are possible, depending on the type of spatial effect anticipated in the data. Here b is the bandwidth parameter for which we assume prior ignorance, and so assign a vague Normal prior distribution to the log of the bandwidth (so as to overcome the non-negativity contraint)

$$\log(b) \sim N(0, 10^2).$$

We assume prior ignorance for each of the seasonal and site effects and so assign each parameter a diffuse Normal prior distribution,

$$
\begin{aligned}
\phi_s &\sim N\left(0, 10^2\right), \qquad 1,\ldots,12 \\
\sigma_i &\sim N\left(0, 10^2\right), \qquad 1,\ldots,50.
\end{aligned}
$$

6.2.6 MCMC Algorithm

We employ a Markov Chain Monte Carlo (MCMC) algorithm in order to fit the extended hotspot prediction model, which is largely similar to the algorithm used to fit the model in Chapter 5. Given the non-conjugate structure of our model, we make use of a Metropolis-Hastings algorithm in order to carry out updates of the parameter vector, $\boldsymbol{\Psi}$. Given various components of our model (e.g. $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, c) are deterministic conditional on other parameters, the parameter vector we must carry out inference on is,

$$\boldsymbol{\Psi} = (\beta_0, \ldots, \beta_P, \beta_t, \sigma_1, \ldots, \sigma_n, \phi_1, \ldots, \phi_S, \alpha_1, \ldots, \alpha_n, \tau_1, \ldots, \tau_n, b)$$

where here $P = 16$, $S = 12$ and $n = 50$. The MCMC algorithm is implemented as follows:

1. Initialise the chain at $\boldsymbol{\Psi}^{(0)} = (\beta_0^0, \ldots, \beta_P^0, \beta_t^0, \sigma_1^0, \ldots, \sigma_n^0, \alpha_1^0, \ldots, \alpha_n^0, \tau_1^0, \ldots, \tau_n^0)$, where each parameter's initial value is the mean of the corresponding prior distribution. Set iteration counter m = 1.

2. For each element Ψ_j generate a proposed update value, Ψ_j^*. If $\Psi_j \in (\beta_0, \ldots, \beta_t, b)$ the proposal is generated via a Normal random walk, i.e.

$$\Psi_j^* \sim N\left(\Psi_j^{m-1}, \epsilon_j\right).$$

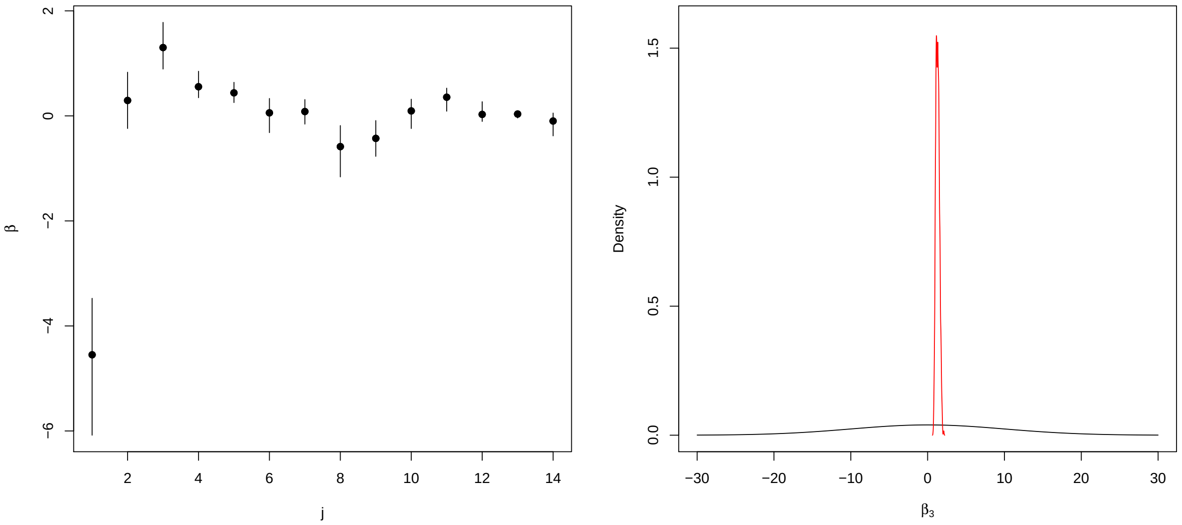
If $\Psi_j \in (\phi_1, \ldots, \phi_S)$ the proposal distribution remains Normal, however the mean is obtained via the CAR structure discussed in Section 6.2.3. Similarly if $\Psi_j \in (\sigma_1, \ldots, \sigma_n)$ the proposal distribution is Normal with mean given by the result of the KDE described in Section 6.2.4. If $\Psi_j \in (\alpha_1, \ldots, \alpha_n)$ the proposal distribution is a zero-inflated Normal distribution as in Chapter 5. Finally if $\Psi_j \in (\tau_1, \ldots, \tau_n)$, the proposal is generated from a Gamma proposal distribution,

$$\Psi_j^* \sim Ga\left(\frac{\left(\Psi_j^{(m-1)}\right)^2}{\epsilon j}, \frac{\Psi_j^{(m-1)}}{\epsilon j}\right).$$

The parametrisation of the Gamma proposal distribution was chosen such that it has mean $\Psi_j^{(m-1)}$ and variance ϵ_j as with the Normal distribution for the other parameters.

3. Set $\Psi_j^{(m)} = \Psi_j^*$ with probability p_{ij},

$$p_{ij} = min\left(1, \frac{f(\Psi_j^*)}{f\left(\Psi_j^{(m-1)}\right)} \frac{\pi(\Psi_j^{(*)})}{\pi(\Psi_j^{(m-1)})}\right)$$



(a) Posterior means and 95% CIs for $\boldsymbol{\beta}$ (b) DAP (black) and posterior densities for β_3

Figure 6.4: Posterior output for the covariate effect parameter vector $\boldsymbol{\beta}$

where $f(\Psi_j^{(*)})$ is the model likelihood evaluated using the proposed value, and $f\left(\Psi_j^{(m-1)}\right)$ is the current model likelihood. Set $\Psi_j^{(m)} = \Psi_j^{(m-1)}$ otherwise. In the case $\Psi_j = b$, since the bandwidth b has no explicit representation in the likelihood, for each proposal $b^{(*)}$, we generate pseudo-proposal values $\boldsymbol{\sigma}^*$, where each element σ_i^* is a deterministic weighted sum of the current vector $\boldsymbol{\sigma}^{(m)}$ with weights calculated using the proposed bandwidth, i.e.

$$\sigma_i = \sum_{j \neq i} w_{ij}^* \sigma_j,$$
$$w_{ij}^* = f\left(b^*\right).$$

We note here that in this context $\boldsymbol{\sigma}^*$ are just pseudo-proposals, and so are discarded (that is to say $\boldsymbol{\sigma}^{(m)}$ does not change) regardless of whether b^* is accepted as a sample from $\pi\left(b|\boldsymbol{y}\right)$ or not.

4. Set m = m + 1. Go to step 2.

We initialise each parameter at its prior mean, tune the innovation parameters to reach an acceptance rate of approximately 20-30% before running the model for 10000 iterations and discarding the first 1000 as burn-in.

For the covariate effects $\boldsymbol{\beta}$ we obtain posterior means and 95% credible intervals given in Figure 6.4a.

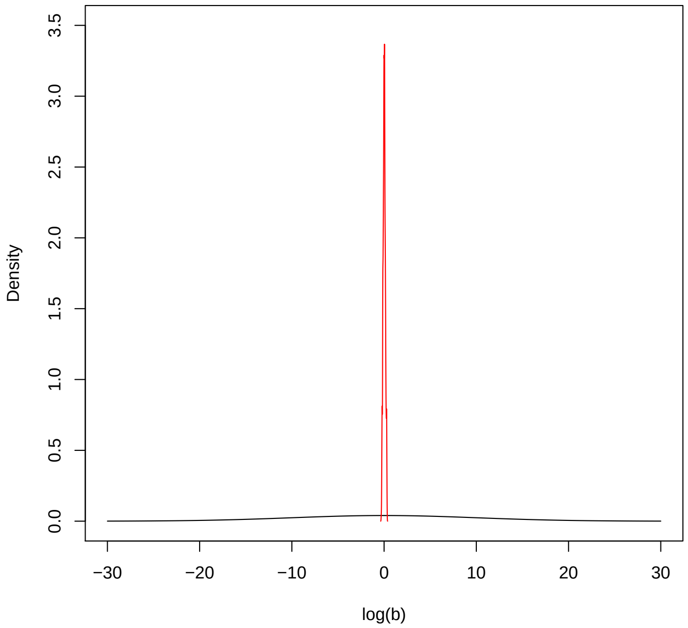


Figure 6.5: The prior (black) and posterior (red) densities for the logarithm of the bandwidth parameter b

From Figure 6.4 we observe a relatively high level of posterior precision with regards to the parameter estimates with very small 95% credible intervals for many of the elements of $\boldsymbol{\beta}$, which may be in part due to the data augmented prior distributions we impose which were designed to inform our analysis and thus reduce posterior uncertainty. We observe many of the densities remain close to zero suggesting the covariates do not have a large effect on the collision rate at a site, although we note β_2 is significantly greater than 0, suggesting sites at intersections have a higher collision rate than sites not at intersections.

We allow the data to select the most appropriate values of spatial smoothing bandwidth parameter b by including it as a parameter in the model. In Figure 6.5 we see prior and posterior densities for $\log(b)$, chosen so as to overcome non-negativity constraints, and observe a huge increase in certainty around the true value of the parameter. We obtain a posterior mean for $\log(b)$ of 0.018 with a 95% credible interval of (-0.221, 0.274), meaning the posterior mean and 95% credible interval for the bandwidth b are 1.03 and (0.801, 1.32) respectively.

We observe a posterior mean for the global trend parameter to be -0.1, however we also incorporate the potential for site specific deviations from the global trend via the α_i parameter, although we add a zero-inflation component to this parameter to reduce the risk of erroneous site level trends being detected by the model. This zero-inflation

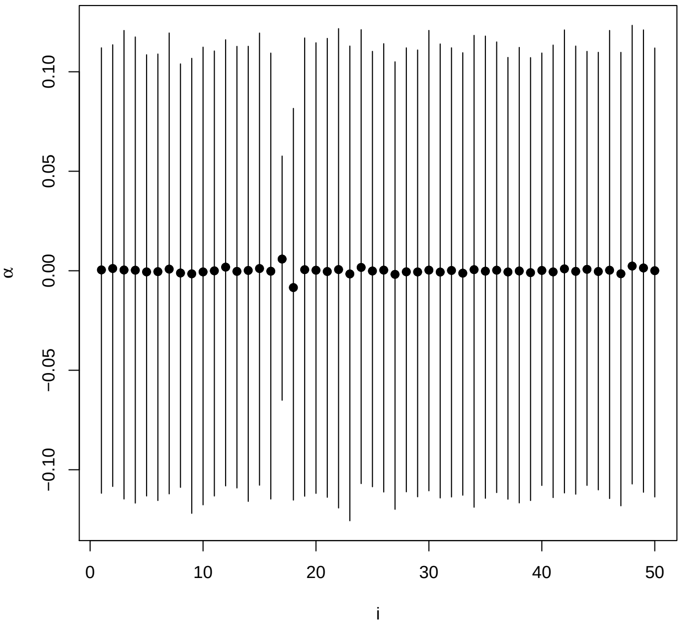


Figure 6.6: Posterior means and 95% credible intervals for α_i

is highly evident in Figure 6.6, where we observe posterior means for α_i at all sites to be very close to zero, suggesting there was no site specific trend present anywhere in our data, and we can reasonably conclude the trend effect is constant across the sites we have analysed.

6.3 Interpolating Spatial Effects

A commonly used advantage to deploying spatial models when analysing data is the ability to interpolate between datapoints to form an estimate of the spatial effect at locations for which data have not been collected. While the concept of interpolatiing spatial effects becomes less advantageous, and arguably less meaningful, for small scale road safety datasets (e.g. small towns and cities), for larger scale macroanalyses (e.g. counties/states and countries) the ability to generate a general picture of local effects across a large region (where fitting the model from Section 6.1 becomes computationally impractical) becomes much more useful. In cases where this is of interest we propose fitting independent spatial effects without any form of local smoother,

$$\sigma_i \sim N\left(0, 10^2\right)$$

allowing us to fit spatial models to these raw spatial effects. At each iteration j of the MCMC algorithm we obtain a sample of the spatial effects $\boldsymbol{\sigma}^{(j)} = \left(\sigma_1^j, \ldots, \sigma_n^j\right)$. Using this sample we can carry out kriging [Cressie, 1992], a process by which we can estimate the spatial effect across a given grid of points beyond just those where which we have observed data. Repeating this procedure at each iteration of the MCMC algorithm allows a distribution of interpolated spatial effects to be constructed, and hence a mean and standard error for the spatial effect at each point on the grid can be obtained.

Beyond knowing simply which regions appear to have an elevated collision compared to what would be expected across the network, given by the parameter σ, practitioners may also be interested in investigated trend deviations, given by α. Hence the spatial modelling approach outlined above can be replicated for successive samples of α, i.e. $\boldsymbol{\alpha}^{(j)} = \left(\alpha_1^j, \ldots, \alpha_n^j\right)$, with a spatial model being applied to $\boldsymbol{\alpha}^{(j)}$ allowing us to interpolate local effects across a large region.

We apply the hotspot model as defined in Section 6.2 to the 50 Halle sites on which we fitted the full model, except here we assume independence between seasonal and spatial effects (initially), giving rise to the model

$$Y_{i,s,t}|\lambda_{i,s}(t) \sim \begin{cases} Pois\left(\lambda_{i,s}(t)\right), & t = 0 \\ NegBin\left(\text{Mean} = \lambda_{i,s}(t), \text{Variance} = \lambda_{i,s}(t)c_i(t)\right), & t < 0, \end{cases}$$

$$\begin{aligned} \lambda_{i,s}(t) &= \exp\left(\mu_i(t) + \sigma_i + \phi_s + \alpha_i t\right) \qquad i = 1, \ldots, 50 \\ \mu_i(t) &= \exp\left(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_P x_{i,P}\right), \\ \beta_j &\sim N\left(0, 10^2\right) \qquad j = 1 \ldots, P \\ \sigma_i &\sim N\left(0, 10^2\right) \\ \phi_s &\sim N\left(0, 10^2\right), \qquad s = 1, \ldots, 12. \end{aligned}$$

We run this model using `rjags` and obtain 1000 samples of the posterior distribution for $\boldsymbol{\sigma}$. On each sample $\boldsymbol{\sigma}^{(m)} = \left(\sigma_1^{(m)}, \ldots, \sigma_{59}^{(m)}\right)$ we carry out kriging to obtain 1000 interpolated estimates of the site effect σ at points for which we do not have data. The locations of the 50 locations for which we do have data are given in Figure 6.7.

In order to carry out kriging we make use of the `sp` and `gstat` packages in `R` [Bivand et al., 2013]. Examples of the heat maps that can be created as a result of carrying out kriging on samples of $\boldsymbol{\sigma}$ are given in Figure 6.8.

In addition, if for example we wished to obtain information regarding the site effect at a new point at which we don't have data, say at co-ordinates (11.85, 51.4) we can obtain estimates of the site effect at this point for each sample of $\boldsymbol{\sigma}^{(m)}$ and combine these to form a sampling distribution for the estimated site effect at this point. It is important to note here that the sampling distribution alone will not give us adequate representation of

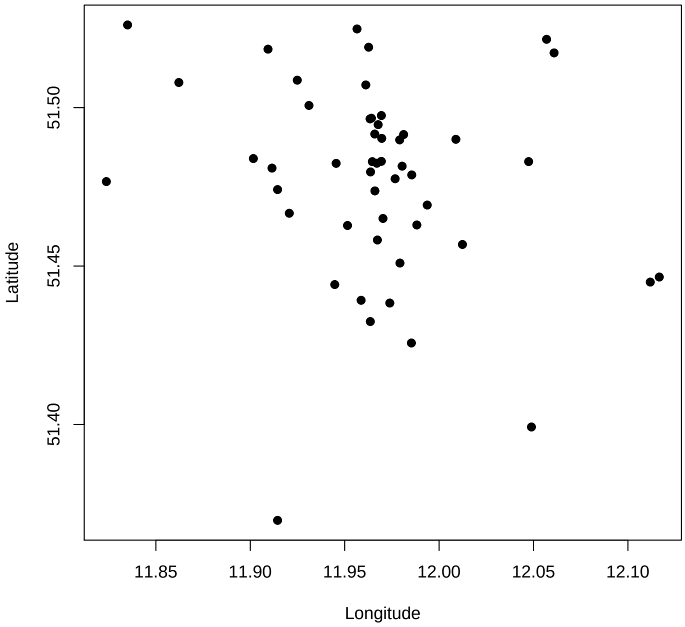
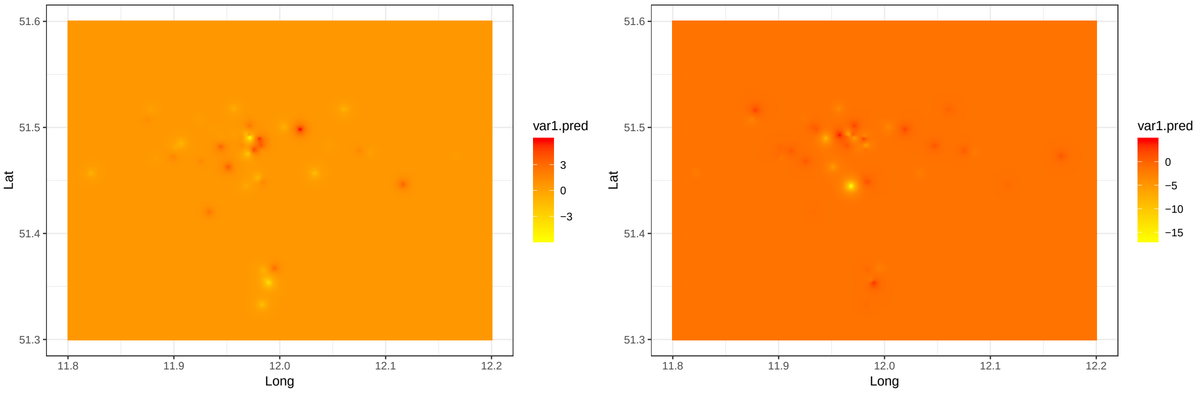


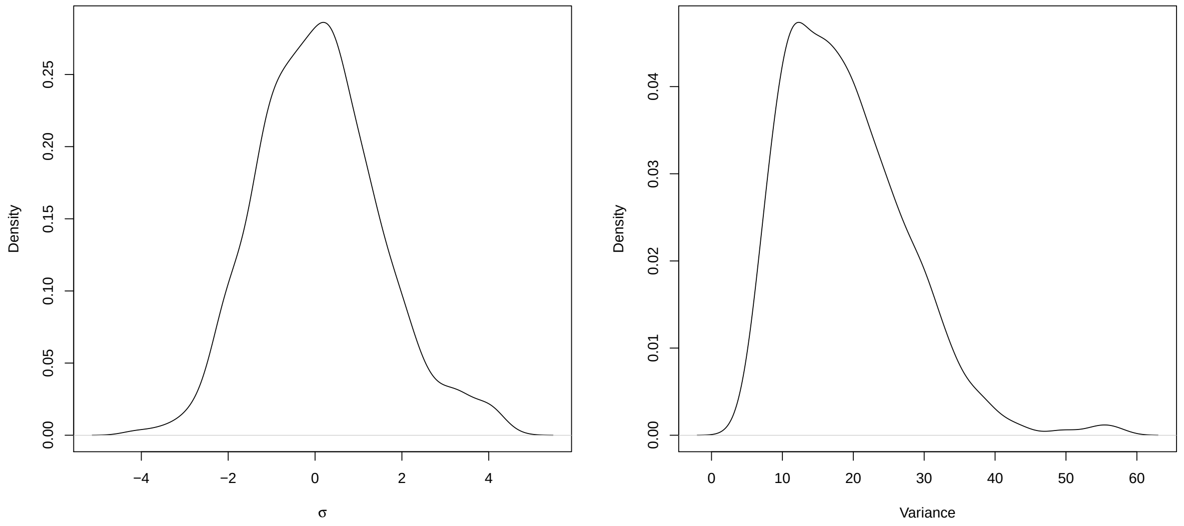
Figure 6.7: The geographic co-ordinates of the 50 Halle sites onto the site effects $\boldsymbol{\sigma}$ of which a spatial model can be fitted



(a) $\boldsymbol{\sigma}^{(25)}$

(b) $\boldsymbol{\sigma}^{(93)}$

Figure 6.8: Heat maps obtained as a result of kriging with an Exponential model using site effect samples $\boldsymbol{\sigma}^{(25)}$ and $\boldsymbol{\sigma}^{(93)}$



(a) Sampling distribution for σ

(b) Sampling distribution for Var(σ)

Figure 6.9: Sampling distributions of σ and Var(σ) for the point at co-ordinates (11.85, 51.4)

the uncertainty surrounding the estimate since this will decrease with increasing numbers of estimates M. Hence to properly represent the uncertainty of our estimate, we also obtain a sampling distribution of the variance of the interpolated estimate to account for uncertainty in the spatial model, particularly where we are attempting to obtain estimates at a point far from our observed data. Examples of this using a null model to estimate σ at the point (11.85, 51.4) are given in Figure 6.9.

6.4 Accounting for Collision Severity

The hotspot prediction model described in Chapter 5, and thus far in this chapter has sought to analyse and predict collision counts at locations in order to determine which locations are sufficiently dangerous to warrant treatment. A flaw in this approach is that it is assumed all collisions are equal, regardless of their severity, with respect to the level of danger at a location and thus its need for treatment. This is clearly flawed, since if for a given time period a location had 9 collisions all causing fatalities, this would clearly be more dangerous than a location which has 10 collisions in which no one was hurt. Hence we must extend the mantra outlined in Chapter 1 whereby we claimed the number of collisions to be a proxy for the level of safety at a location, to include the severity, and thus threat to health/life of these collisions. There are many ways by which collision severity can be disaggregated, in the interests of ensuring significant numbers of each (frequency decreases rapidly as severity increases), here we choose to model 3 severities (although the methodology works the same for any number of severity classes):

1. KSI (Killed or Seriously Injured): A collision directly leading to one (or more) deaths or severe injuries
2. Slight: A collision directly leading to one (or more) non-severe injuries
3. PDO (Property Damage Only): A collision in which no significant injuries were sustained

Every collision must fall into one of these categories (collisions leading to injuries of varying severity are categorised by the most severe), and so the total number of collisions can be disaggregated into the number of collisions of each severity,

$$y_{i,t} = y_{i,t,\mathrm{K}} + y_{i,t,\mathrm{S}} + y_{i,t,\mathrm{P}}, \tag{6.7}$$

where $y_{i,t,\mathrm{K}}$, $y_{i,t,\mathrm{S}}$ and $y_{i,t,\mathrm{P}}$ denote the number of KSI, slight and PDO collisions respectively. From this we can therefore model the vector of severity totals, $\boldsymbol{y}_{i,t} = (y_{i,t,\mathrm{K}}, y_{i,t,\mathrm{S}}, y_{i,t,\mathrm{P}})$ as an observation from a Multinomial distribution,

$$\boldsymbol{y}_{i,t} \sim Mult\left(n = y_{i,t}, p = (\pi_{i,\mathrm{K}}(t), \pi_{i,\mathrm{S}}(t), \pi_{i,\mathrm{P}}(t))\right) \tag{6.8}$$

where $\pi_{i,\mathrm{K}}(t)$, $\pi_{i,\mathrm{S}}(t)$, $\pi_{i,\mathrm{P}}(t)$ are the probabilities a given collision at site i in time period t will have severity KSI, slight or PDO respectively. Since the vector $\boldsymbol{\pi}_i(t) = (\pi_{i,\mathrm{K}}(t), \pi_{i,\mathrm{S}}(t), \pi_{i,\mathrm{P}}(t))$ is made up of probabilities, it seems logical to carry out a multinomial logistic regression to model each element of $\boldsymbol{\pi}_i(t)$, however to do so would be to ignore the clear ordinal structure to the danger, that is that a KSI collision is clearly worse than a slight injury collision is worse than a PDO collision. Hence we must include this ordered structure to the data in our analysis, and so carry out an ordinal logistic regression on the probability vector $\boldsymbol{\pi}_i(t)$. If we define z to be the severity indicator for a given collision, such that,

$$z = \begin{cases} 1, & \text{the collision has severity PDO,} \\ 2, & \text{the collision has severity ``slight''}, \\ 3, & \text{the collision has severity KSI.} \end{cases}$$

We can model this using a continuous, latent variable z^*, the value of which determines the value of z,

$$z = \begin{cases} 1, & \theta_0 < z^* < \theta_1, \\ 2, & \theta_1 < z^* < \theta_2, \\ 3, & \theta_2 < z^* < \theta_3, \end{cases}$$

where θ_1 and θ_2 can be thought of as severity thresholds, which will be determined from the data, and θ_0 and θ_3 are set to be $-\infty$ and ∞ respectively. Hence we obtain

$$Pr(z = k|\boldsymbol{x}) = Pr\left(\theta_{k-1} < z^* < \theta_k|\boldsymbol{x}\right), \qquad k = 1, \dots, 3 \tag{6.9}$$

It is common here ([Haleem and Abdel-Aty, 2010], [Abdel-Aty, 2003], [Al-Bdairi and Hernandez, 201 to assume a standard linear regression form for z^*,

$$z^* = \boldsymbol{\beta x} + \epsilon,$$

where $\boldsymbol{x}$ is a set of covariates believed to affect the prevalence of each class of severity, and $\boldsymbol{\beta}$ is the corresponding coefficient vector. We extend this here to incorporate features incorporated into the rate parameter λ discussed in Section 5.4, namely global and local trends, a site effect and seasonal effect, hence we obtain

$$z^*_{i,s}(t) = \beta_0 + \boldsymbol{\beta x}_i + \beta_t t + \sigma_i + \phi_s + \alpha_i t. \tag{6.10}$$

As opposed to the case in Section 5.4, here we do not impose any modelling structure to the seasonal effects $\boldsymbol{\phi}$ nor the spatial effects $\boldsymbol{\sigma}$, since extreme values in these vectors will be of interest, and hence we have independent priors for each effect

$$\begin{aligned} \sigma_i &\sim N(0, 10) \\ \phi_s &\sim N(0, 10) \end{aligned}$$

with the local trend parameter α retaining the same zero-inflated Normal structure,

$$\begin{aligned} \alpha_i &= \alpha_N \alpha_Z \\ \alpha_N &\sim N(0, 10) \\ \alpha_Z &\sim Bern(0.5). \end{aligned}$$

Incorporating these additional parameters provides the potential for much more informative inference compared with just including covariate effects. This model now has the ability to:

- Detect seasonal variability in severity proportion. The assumption of constant severity proportions across seasons will be far too restrictive in cases where there is high climatic variability, e.g. in Florida during the hurricane season, not only do we expect a higher rate of collisions, but also that these collisions will be more severe relative to outside of the tropical storm season.

- Detect site specific deviations from the expected severity proportions as determined by the covariates. This not only improves the accuracy of the model, it also allows for easy inference as to which locations appear to have more or less severe collisions than would be expected from a standard covariate model (analogous to the site effect discussed in Section 5.4). This would provide important information as it pertains to the allocation of treatment, since even if a location is not predicted to have a particularly high collision count, if these collisions are still predicted to have a high degree of severity this may still be indicative of the site requiring treatment.

- Detect and extrapolate severity trends at a global and local level. As discussed in Chapter 5 detecting trends in collision totals allows for proactive decision making with respect to hotspot identification, while avoiding unnecessary treatment at locations which appear to be improving without intervention. This same logic applies in the area of severity proportion identification, where it is important any apparent trends in severity are extrapolated to the future, in order to allow for proactive hotspot identification. Hence we incorporate a global trend component reflecting severity changes across the network, while allowing for the same site specific deviations (with the case favoured for no deviation) at a site level.

We can therefore use Equation 6.9 to derive

$$\begin{aligned} Pr\left(z = k|\boldsymbol{x}\right) &= Pr\left(\theta_{k-1} < z^* < \theta_k|\boldsymbol{x}\right) \\ &= Pr\left(z^* < \theta_k|\boldsymbol{x}\right) - Pr\left(z^* < \theta_{k-1}|\boldsymbol{x}\right) \\ &= Pr\left(z^* + \epsilon < \theta_k\right) - Pr\left(z^* + \epsilon < \theta_{k-1}\right) \\ &= Pr\left(\epsilon < \theta_k - z^*\right) - Pr\left(\epsilon < \theta_{k-1} - z^*\right) \\ &= \Phi\left(\theta_k - z^*\right) - \Phi\left(\theta_{k-1} - z^*\right), \end{aligned} \qquad (6.11)$$

where $\Phi\left(\cdot\right)$ is the standard Normal CDF.

From this we can therefore obtain,

$$\begin{aligned} Pr(z = 1|\boldsymbol{x}) &= \Phi(\theta_1 - z^*), \\ Pr(z = 2|\boldsymbol{x}) &= \Phi(\theta_2 - z^*) - \Phi(\theta_1 - z^*), \\ Pr(z = 3|\boldsymbol{x}) &= 1 - \Phi(\theta_2 - z^*), \end{aligned} \qquad (6.12)$$

and a visual representation of this structure is given in Figure 6.10. It is clear from Equation 6.11 that in order for $Pr(z = k|\boldsymbol{x})$ to be a well defined probability (i.e. non-negative) we must ensure the sequence of $\theta_{k,i}$ is increasing, i.e. $-\infty < \theta_{1,i} < \theta_{2,i} < \infty$. It is this restriction that forms the key modelling difference between ordinal regression and standard regression. Since we have no such restrictions on the elements of the regression coefficient vector $\boldsymbol{\beta}_i = (\beta_{0,i}, \beta_{1,i}, \ldots, \beta_{P,i})$ we choose to assign independent vague prior Normal distributions to each,

$$\beta_{j,i} \sim N\left(0, 10^2\right), \qquad j = 1, \ldots, P.$$

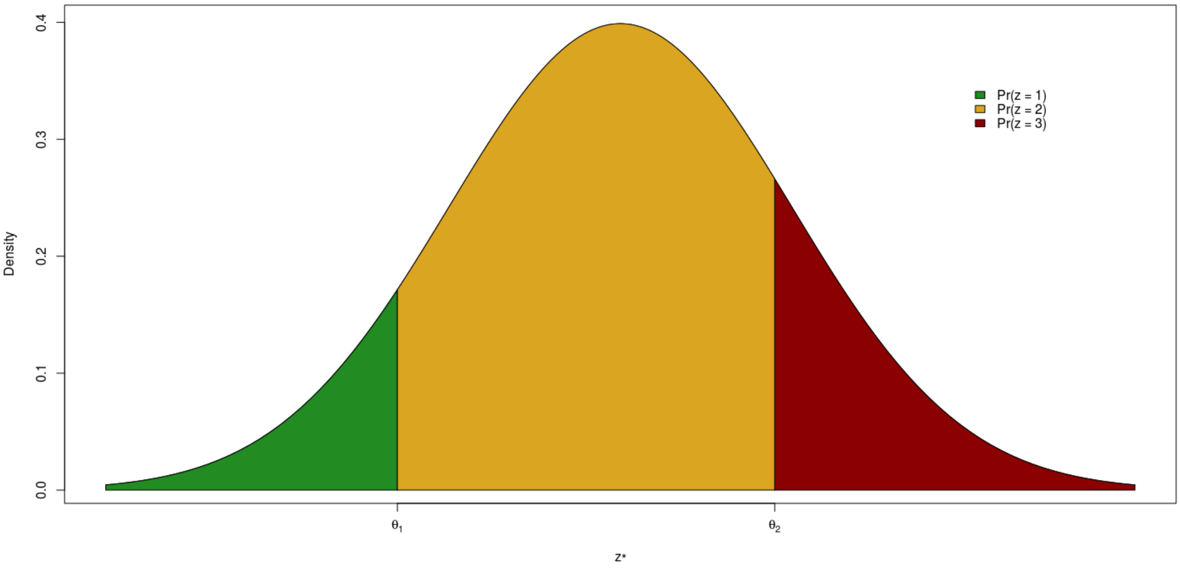


Figure 6.10: A visual representation of the ordinal regression modelling structure, showing how increased values of the latent variable z^* cause the value of z to change from 1 ($z^* < \theta_1$ - the green region) to 2 ($\theta_1 < z^* < \theta_2$ - yellow) to 3 ($z^* > \theta_2$ - red)

.

We note here that the covariates $\boldsymbol{x}$ used to model z^* need not necessarily be the same as used in the SPF when estimating μ, and should only consist of covariates believed to affect the likelihood of a given collision being of a more/less severe nature. We then interpret the values of the coefficient vector $\boldsymbol{\beta}$ to be that more positive the value of $\beta_{j,i}$, the more increased levels of covariate j are associated with greater proportions of severe collisions, and vice-versa. The final parameters we estimate are the threshold parameters $\theta_{1,i}$ and $\theta_{2,i}$, which as described earlier are constrained by their ordering, in that we must have $\theta_{1,i} < \theta_{2,i}$. In order to model this we allow $\theta_{1,i}$ to be unconstrained (technically its only restriction being $-\infty < \theta_{1,i}$ and then model the conditional distribution $\theta_{2,i}|\theta_{1,i}$. Hence for θ_1 we again fit a vague Normal prior distribution,

$$\theta_{1,i} \sim N\left(0, 10^2\right).$$

In the non-ordinal case we would assign the same vague Normal prior to $\theta_{2,i}$, however we are restricted by the ordering $\theta_{1,i} < \theta_{2,i}$, and so we assign $\theta_{2,i}$ to have a vague truncated Normal distribution, bounded below at $\theta_{1,i}$ and unbounded above. Hence $\theta_{2,i}$ has conditional PDF given by,

$$f\left(\theta_{2,i}|\theta_{1,i}\right) = \begin{cases} \phi\left(0, 10^2\right) I(\theta_{1,i}, \infty) & \theta_{2,i} > \theta_{1,i} \\ 0, & \theta_{2,i} < \theta_{1,i}. \end{cases}$$

where $\phi(\cdot)$ is the Normal PDF function, and $I(\theta_{1,i}, \infty)$ is the Normal truncation function bounded below at $\theta_{1,i}$, here given by

$$I(\theta_{1,i}, \infty) = 10 \left(1 - \Phi\left(\frac{\theta_{1,i}}{10}\right)\right)^{-1},$$

where $\Phi(\cdot)$ is the standard Normal CDF function.

We can combine the posterior distributions of our model parameters to form posterior distributions for the latent severity variable z^* at site i, in season s in the following time period $t = 1$,

$$\pi\left(z^*_{i,s}(1)|\boldsymbol{y}\right) = \pi\left(\mu_i(1)|\boldsymbol{y}\right) + \pi\left(\sigma_i|\boldsymbol{y}\right) + \pi\left(\phi_s|\boldsymbol{y}\right) + \left(\pi\left(\alpha_i|\boldsymbol{y}\right) \times 1\right). \tag{6.13}$$

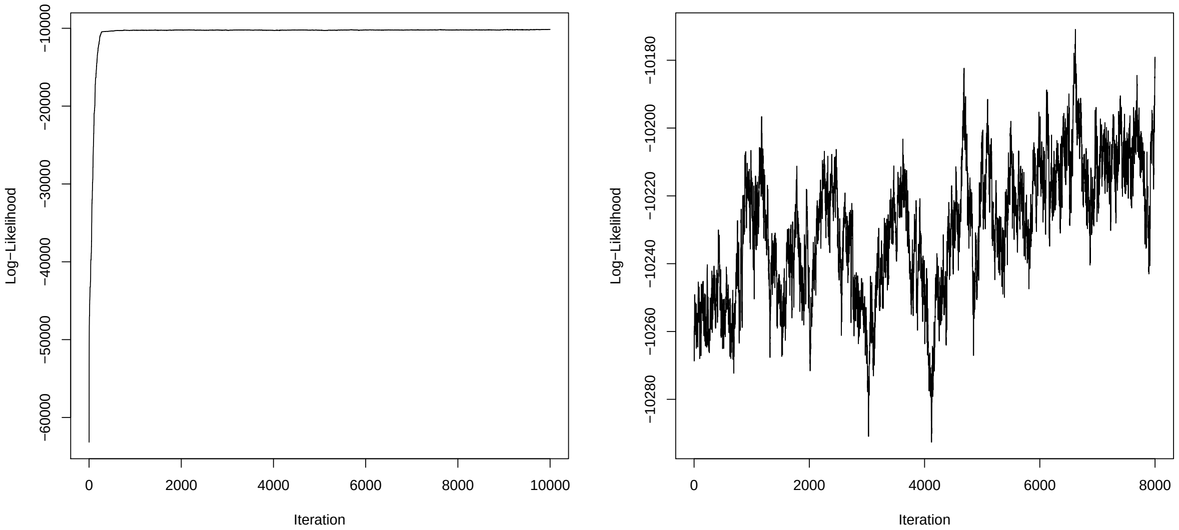
As in Equation 6.12, we define the proportion of collisions of severity k to be

$$\pi_{i,s,k} = \Pr\left(\theta_{k-1} < z^*_{i,s} < \theta_k\right)$$

then we can obtain a sample m from the posterior distribution $\pi_{i,s,k}$ to be the proportion of the posterior samples of $z^*_{i,s,t}$ which lie between $\theta^{(m)}_{k-1}$ and $\theta^{(m)}_k$, and combine these over all M posterior samples to obtain the posterior distribution $\pi\left(\pi_{i,s,k}|\boldsymbol{y}\right)$. Evaluating this posterior distribution using samples from $\pi\left(z^*_{i,s}(1)|\boldsymbol{y}\right)$ as defined in Equation (6.13) therefore provides predictive samples of the proportion of collisions of severity k at site i in season s in the following year. Running a standard hotspot prediction model as defined in Chapter 5 or Section 6.2 obtains a posterior predictive distribution for the collision count at site i in season s of the following year, $\pi\left(y_{i,s}(1)|\boldsymbol{y}\right)$. Multiplying these two distributions therefore allows us to obtain a posterior predictive distribution for the number of collisions of severity k at site i in season s of the following year, $\pi\left(y_{i,s,1,k}|\boldsymbol{y}, \boldsymbol{\pi}\right)$.

6.4.1 Model Application

We apply this model to the zonal Halle data described in Section 1.6, where we have $i = 1, \ldots, 59$ zones of data over $t + 1 = 9$ years. We do not have covariate data for this dataset and so our SPF will simply be the global trend parameter $\mu(t) = \beta_t$. Our data is disaggregated over $S = 12$ monthly periods, and $K = 3$ severities. Hence the full parameter vector for this model is $(\beta_t, \phi_1, \ldots, \phi_{12}, \sigma_1, \ldots, \sigma_{59}, \alpha_1, \ldots, \alpha_{59}, \theta_1, \theta_2)$. For each parameter we assume prior ignorance and so assign independent $N\left(0, 10^2\right)$ prior distributions to each, with the exception of the site specific trend parameter, for which we retain the zero-inflated Normal structure as in Section 6.2. We initialise each parameter in our model at its prior mean of 0 with the exception of θ_2 which we initialise at 0.1 to respect the $\theta_2 > \theta_1$ condition. We use random walk proposals to update each element of the parameter vector, with the exception of θ_2 for which we use a Normal distribution



(a) Log-likelihood for the full model

(b) Log-likelihood after burn-in removal

Figure 6.11: Log-likelihoods for the severity model against iteration before and after burn-in removal

lower truncated at θ_1. We update each parameter independently with the exception of θ_1 and θ_2 for which we carry out a block update so as to maintain the parameter ordering. We tune the innovations on the random walk proposals to achieve in acceptance rate of 20-30% for each parameter and run our model for $M = 10000$ iterations, with plots of the model log-likelihood given in Figure 6.11. From Figure 6.11a we can observe a relatively low log-likelihood value at initialisation as expected, which then increases as the parameter vector approaches the posterior distribution, where the log-likelihood levels off. Informally from this plot therefore we can estimate a burn-in period of 2000 iterations to be removed, with the remaining 8000 log-likelihood scores given in Figure 6.11b, where the log-likelihood remains in the same area, suggesting model convergence. We note the mixing in this model is not ideal but tolerable, and more precise posterior summaries can be obtained by extending the run of the model and taking a high degree of thinning to reduce posterior autocorrelation.

Analysing the vector of seasonal effects $\boldsymbol{\phi}$ gives results shown in Figure 6.12 where we observe all values of ϕ to be significantly below 0, suggesting collisions of a lower severity are more common in all months s. Here we observe a lower value of ϕ, suggesting a lower likelihood of a severe collision in the winter months $s = 12, 1, 2, 3$ corresponding to December to March, perhaps due to increased awareness of the dangers on roads and so fewer high speed collisions etc. Including covariate effects such as average speed in the

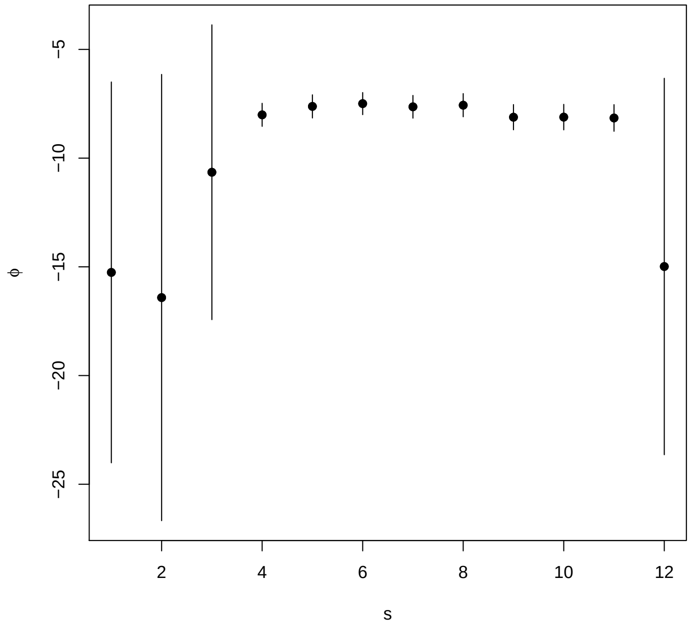


Figure 6.12: Plot of posterior means for ϕ with 95% credible intervals for the severity model

model may change the seasonal estimates significantly. We also observe a much higher posterior standard deviation for the winter months as well, suggesting the data was much less consistent in these months, as opposed to the summer months where the posterior standard deviations are relatively much smaller. Accounting for seasonal dependence such as using a CAR stucture as in Section 6.2.3 could reduce these posterior standard deviations by sharing information between seasons.

In terms of trend, we obtain a posterior mean for β_t to be 0.653 and 95% credible interval (0.038, 0.865), suggesting generally collisions are getting slightly more severe across the entirety of the 59 zones being analysed. A comparison of prior and posterior densities for β_t is given in Figure 6.13 showing the substantial increase in precision over the estimate of the global trend effect in light of the data.

We incorporate the potential for site-specific deviations from this global trend with the (zero-inflated) site specific trend deviation parameter α_i. Posterior means for $\boldsymbol{\alpha}$ with 95% credible intervals are given in Figure 6.14.

From Figure 6.14a as we would expect we observe the majority of sites with $\alpha_i \approx 0$ as a result of the zero-inflation component designed to discourage erroneous deviations from the global trend except when there is clear evidence. The majority of clear deviations from the global trend appear to be negative, suggesting positive trend in the remaining

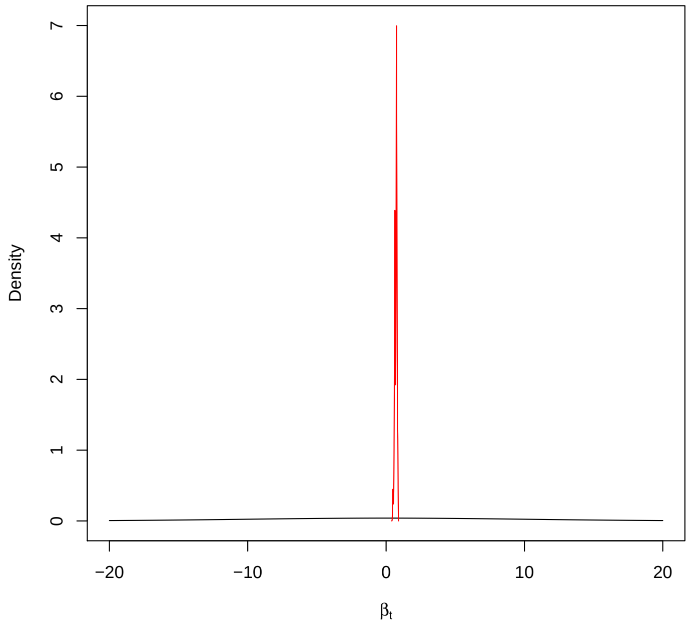
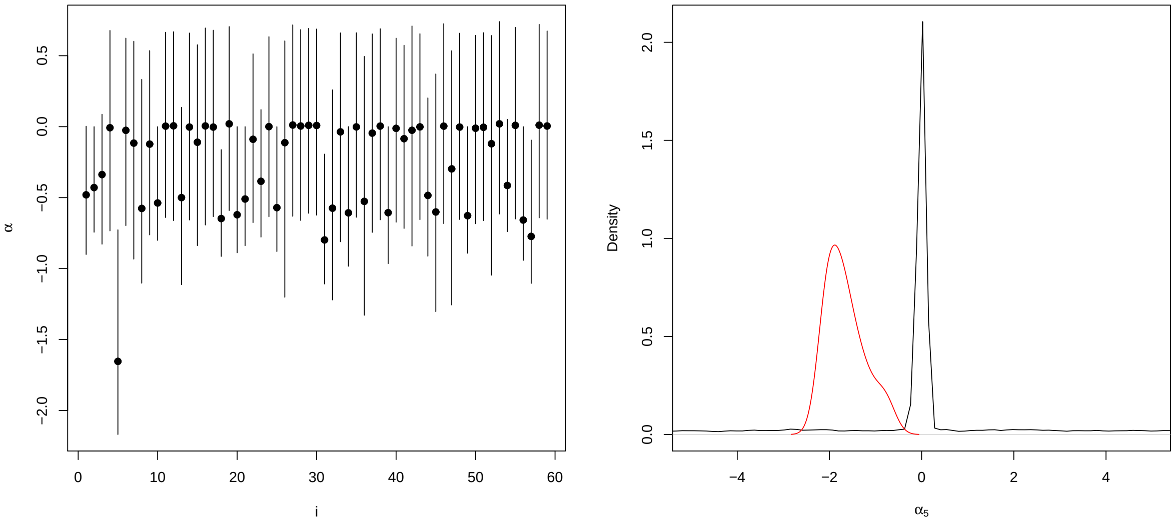


Figure 6.13: Plot of prior (black) and posterior (red) densities for β_t



(a) Posterior means and 95% CIs for α_i (b) Prior (black) and posterior densities for α_5

Figure 6.14: Posterior output for $\boldsymbol{\alpha}$ in the severity model

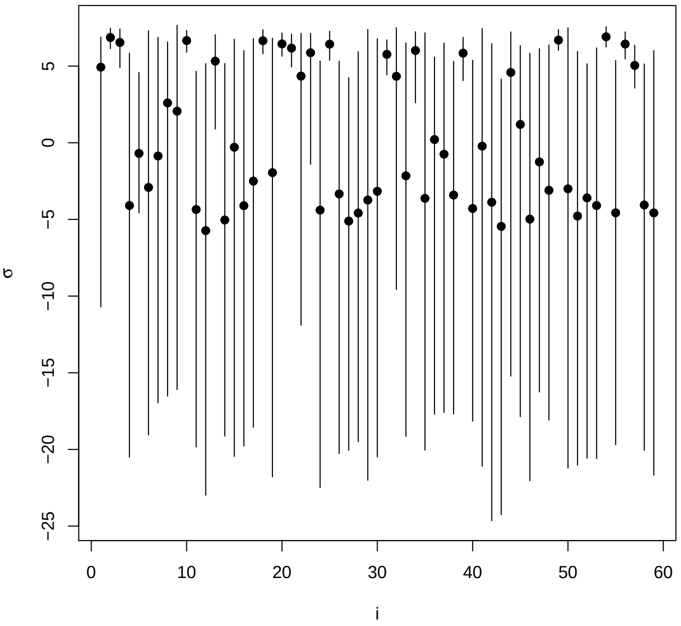


Figure 6.15: Posterior means and 95% credible intervals for $\boldsymbol{\sigma}$ against site i

sites was only slight and so heavily dampened by the zero-inflation component of α_i. The site which deviates from this global trend most heavily is site 5, which has posterior mean of -1.654 and 95% credible interval of (-2.17,-0.727), with prior and posterior density given in Figure 6.14b, suggesting the proportion of severe collisions at this site is decreasing at a much quicker rate compared to the rest of the network.

We plot posterior means and 95% credible intervals for the site effect $\boldsymbol{\sigma}$ in Figure 6.15. From Figure 6.15 we observe a clear mixture of sites with $\sigma_i > 0$, indicating a higher risk of severe collisions occuring, and those with $\sigma_i < 0$ indicating a lower risk of severe collisions. We furthermore notice a significant discrepancy in posterior standard deviation, with sites with more positive values of σ_i generally having a very low standard deviation, and hence we have a much higher level of certainty in the estimate, compared with sites with strongly negative values of σ_i for which the posterior standard deviation is much higher.

For the severity threshold parameters we obtain posterior means of 1.28 and 2.21 for θ_1 and θ_2 with respective 95% credible intervals of (1.255, 1.321) and (2.181, 2.259), suggesting a very high level of posterior precision for these parameters. The posterior mean of θ_1 being greater than 0 suggests the majority of collision are of lowest severity, with a relatively high posterior mean of θ_2 (relative to a standard Normal random variable) suggests a very small number of fatal collisions in our dataset. We can visualise this by

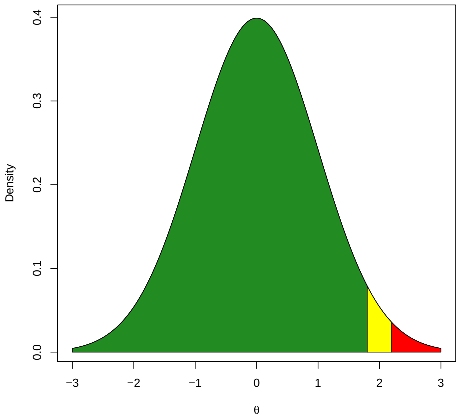


Figure 6.16: Visual representation of the fitted ordinal regression model for severity, with the green region corresponding to collisions of severity 1 (PDO); yellow corresponding to severity 2 (slight injury) and red corresponding to severity 3 (fatal or serious injury)

reproducing Figure 6.10 for our posterior means of θ_1 and θ_2, given in Figure 6.16.

Finally we wish to extrapolate the latent severity variable z^* so that we can form predictions regarding the number of collisions at each severity in future time periods as discussed in Section 6.4. In order to do this we form a posterior distribution for $z^*_{i,s}(1)$ by combining posterior distributions for model parameters,

$$\pi\left(z^*_{i,s}(1)|\boldsymbol{y}\right) = \pi\left(\beta_t|\boldsymbol{y}\right) + \pi\left(\sigma_i|\boldsymbol{y}\right) + \pi\left(\phi_s|\boldsymbol{y}\right) + \pi\left(\alpha_i|\boldsymbol{y}\right).$$

Hence we can form a posterior distribution for the proportion of collisions of each severity in the future time point, $\pi_{i,s,k}(1)$, by estimating

$$\pi_{i,s,k}(1) = \Pr\left(\theta_{k-1} < z^*_{i,s}(1) < \theta_k\right)$$

from posterior samples. We can therefore obtain posterior samples from $\pi_{i,s,k}(1)$,

$$\pi^{(m)}_{i,s,k}(1) = \Pr\left(\theta^{(m)}_{k-1} < z^*_{i,s}(1)|\boldsymbol{y} < \theta^{(m)}_k\right)$$

and hence combining over all posterior samples of $\boldsymbol{\theta}$ gives a full posterior distribution for $\pi(1)$. In order to obtain predictive distributions for the number of collisions of each severity at a future time point, we must obtain a posterior predictive distribution for the overall number of collisions. Here we apply a simplified version of the hotspot prediction model described in Section 6.2, where we assume independence between parameters, to the zonal Halle dataset described in Section 1.6. Hence we have,

$$Y_{i,s,t}|\lambda_{i,s}(t) \sim \begin{cases} Pois\left(\lambda_{i,s}(t)\right), & t = 0 \\ NegBin\left(\text{Mean} = \lambda_{i,s}(t), \text{Variance} = \lambda_{i,s}(t)c_i(t)\right), & t < 0, \end{cases}$$

$$\begin{aligned} \lambda_{i,s}(t) &= \exp\left(\beta_t t + \sigma_i + \phi_s + \alpha_i t\right) \qquad i = 1, \ldots, 59 \\ \sigma_i &\sim N\left(0, 10^2\right) \\ \phi_s &\sim N\left(0, 10^2\right), \qquad s = 1, \ldots, 12. \end{aligned}$$

From this we can therefore obtain a posterior predictive distribution for the number of collisions in the next time point $\pi\left(y_{i,s}(1)|\boldsymbol{y}\right)$, in the same way as in Chapter 5. We can then obtain a posterior predictive number of collisions of severity k by multiplying the posterior predictive distribution for the overall number of collisions by the posterior estimate of the proportion of collisions of severity k in month s of time period $t = 1$,

$$\pi\left(y_{i,s,k}(1)|\boldsymbol{y}\right) = \pi_{i,s,k}(1)\pi\left(y_{i,s}(1)|\boldsymbol{y}\right).$$

6.5 Accounting for Causation

In Section 6.4 we upgraded the hotspot prediction model by allowing it to provide more precise information regarding the level of safety at a location by disaggregating the overall collision totals by severity. Whilst this provides additional information regarding where safety treatments should be deployed, it does little to inform the practitioner as to the causes of these collisions, and thus does not help advise as to which countermeasures should be deployed. Fortunately, we can make use of a similar framework to that used to account for severity in order to provide information regarding potential causation factors behind the overall collision totals. We can make use of information commonly provided in police reports regarding the circumstances of each collision to produce additional indicator causation variables to categorise the nature of each collision, e.g. due to speeding, due to drink driving, purely accidental. From this we can therefore produce total numbers of collisions due to each causation factor at a given site in a given year, exactly in the same way as for severity. Hence, as in the case for severity, we can perform inferences on the proportion of the total number of collisions corresponding to each causation factor, in order to predict the number of collisions due to each factor in a future time period. We can do this in the same way as for severity, except in this case there is no natural ordering to the different causation factors (e.g. there is no natural ordering to whether a collision was due to alcohol, speeding or neither), and so rather than carry out an ordinal multinomial logistic regression, we simply carry out a standard multinomial logistic regression.

Formally we can state that for a site i in time period t with collision count $y_{i,t}$, we have collision totals due to F causation factors, denoted $\boldsymbol{y}_{i,t} = (y_{i,t,1}, \ldots, y_{i,t,F})$. Unlike the case for severity we cannot assume each collision will be due to exactly one of the causation factors being investigated, there is the possibility for instance that a driver could be speeding and over the alcohol limit, or that a collision might not be due to any of the factors being investigated. We note here that we could force the causation factors to be Multinomially distributed by only recording the primary cause of the collision (thereby removing the possibility for a single collision to record multiple causation factors), and including "None" as an option, however this would mean discarding information regarding secondary factors (as well as introducing the potential for subjective biases when deciding the primary factor) and so we choose not to enforce this. Hence we cannot make an analogous statement to Equation 6.7, and hence cannot state that $\boldsymbol{y}_{i,t}$ is an observation from a Multinomial distribution. Instead we make the assumption that the prevalence of each causation factor is independent of the others, and hence each element of $\boldsymbol{y}_{i,t}$ can be modelled independently as an observation from a Binomial distribution,

$$y_{i,t,k} \sim Bin(y_{i,t}, \pi_{i,k}(t)), \qquad k = 1, \ldots, F.$$

We estimate each element of the causation factor probability vector $\boldsymbol{\pi}_{i,t} = (\pi_{i,t,1}, \ldots, \pi_{i,t,F})$ using a logistic regression model, again we are free to choose between any logistic regression structure, with the two most common being the logit and probit models, as with severity here we elect to use a probit model structure,

$$\pi_i(t) = \Phi\left(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_P x_{i,P}\right).$$

Unlike in the case for severity, we no longer have an ordinal structure to the data, and so we do not need to place any ordering constraints on the parameters in the logistic regression and so we can simply fit vague Normal prior distributions to each,

$$\beta_{i,j} \sim N(0, 10^2) \qquad j = 1, \ldots, P.$$

6.5.1 Model Application

We apply our causation model to the zonal Halle dataset, where again we have $n = 59$ sites over $t + 1 = 9$ years disaggregated over $S = 12$ monthly periods. In this case we do not disaggregate by severity, instead we have counts disaggregated by collision type, in this case we disaggregate by a binary indicator corresponding as to whether the collision occurred in darkness or not. We retain the rate parameter structure as in Section 6.4 and hence we have the following model structure,

$$\begin{aligned} y_{i,s,t,f} &\sim Bin\left(y_{i,s,t}, \pi_{i,s}(t)\right), \\ \pi_{i,s}(t) &= \beta_t + \sigma_i + \phi_s + \alpha_i t, \end{aligned}$$

where $y_{i,s,t,f}$ is the number of collisions of type f occurring at site i, in season s of year t. $\pi_{i,s}(t)$ is the proportion of collisions of type f occurring at site i in year t and so as in Section 6.4 we can extrapolate this into future time periods, and combine with predicted overall numbers of collisions in order to form predictions of the number of collisions of type f in future time periods. We assume independent diffuse Normal prior distributions for each element of $\boldsymbol{\sigma}$ and $\boldsymbol{\phi}$ along with β_t,

$$\begin{aligned} \sigma_i &\sim N\left(0, 10^2\right), \qquad && i = 1, \ldots, n \\ \phi_s &\sim N\left(0, 10^2\right), \qquad && 1, \ldots, S \\ \beta_t &\sim N\left(0, 10^2\right). \end{aligned}$$

We retain independent zero-inflated Normal prior distributions for the elements of $\boldsymbol{\alpha}$,

$$\begin{aligned} \alpha_i &= \alpha_N \alpha_Z \\ \alpha_N &\sim N\left(0, 10^2\right) \\ \alpha_Z &\sim N\left(0, 10^2\right). \end{aligned}$$

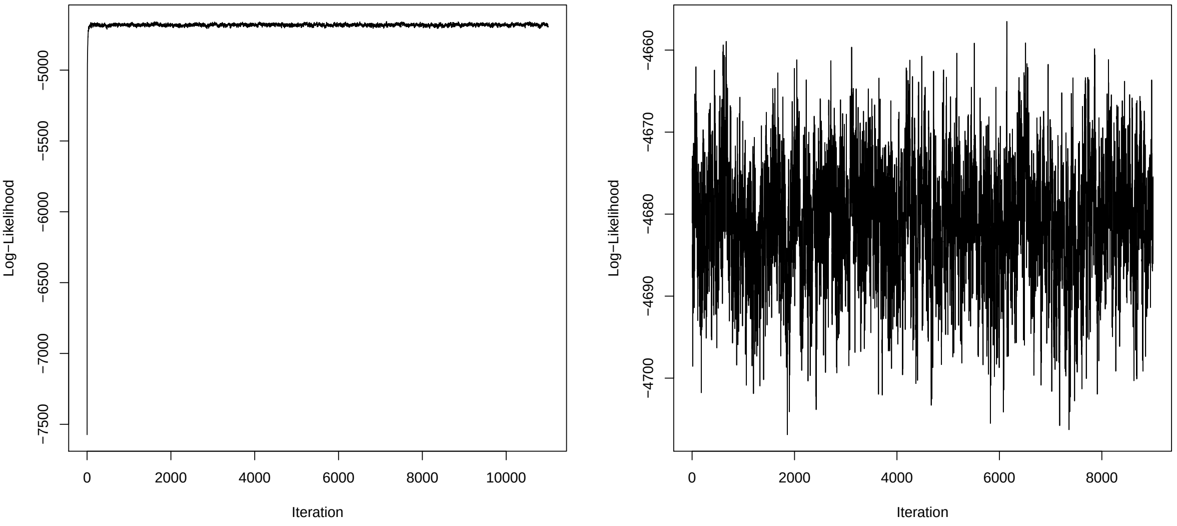
We adopt a Metropolis random walk to update each parameter, with innovation parameter tuned to give an acceptance rate between 20% - 30%. We initialise the chain at each parameter prior mean, and run our model for $M = 11000$ iterations, with traceplots of the log-likelihood given in Figure 6.17.

From Figure 6.17a we see that as with the severity model we have a strong increase in log-likelihood during the burn-in period, before the log-likelihood converges to a steady state once the chain has reached its posterior distribution. Removing the first 1000 observations as burn-in gives the remaining 10000 log-likelihoods shown in the traceplot in Figure 6.17b which show good convergence.

We obtain posterior means for the seasonal effects $\boldsymbol{\phi}$ with results shown in Figure 6.18.

From Figure 6.18 there does not appear to be any clear seasonal effect present in the data, with months $s = 4$ and 7 (April and July) corresponding to the highest posterior means of s, with respective posterior means of 0.334 and 0.401 and 95% credible intervals of (0, 0.641) and (0.098, 0.649). While these credible intervals suggest the ϕ_4 and ϕ_7 are greater than 0, none of the ϕ parameters appear to be sufficiently far from 0 to have a considerable effect on the collision rate, suggesting that no month makes the proportion of collisions due to darkness significantly more or less likely.

Posterior output for the site effect vector $\boldsymbol{\sigma}$ is given in Figure 6.19a, where in Figure 6.19a we obseve reasonably uniform site effects across most sites with $\sigma_i \approx -1$, suggesting collisions in darkness occur less frequently than those not in darkness at all sites. The clear exception to this rule is site 46, for which prior and posterior densities are compared in Figure 6.19b, where there is a hugely negative posterior mean of -3.62 and



(a) Log-likelihood for the full model

(b) Log-likelihood after burn-in removal

Figure 6.17: Log-likelihoods for the causation model against iteration before and after burn-in removal

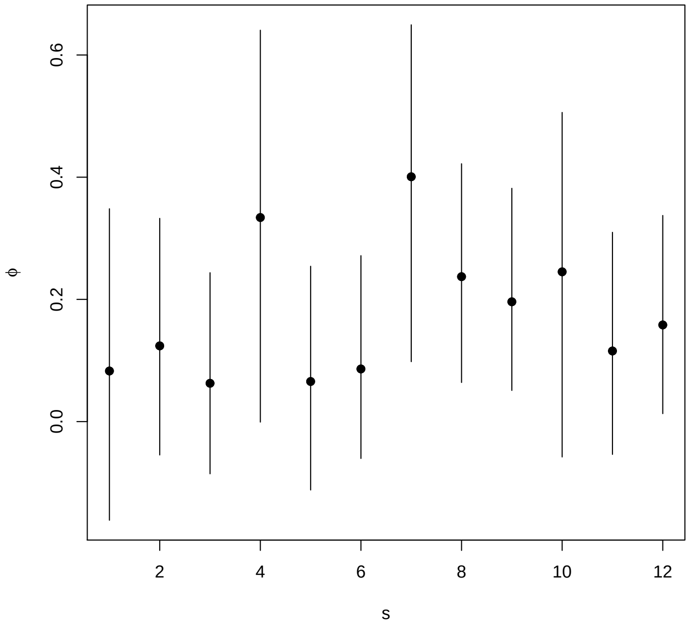
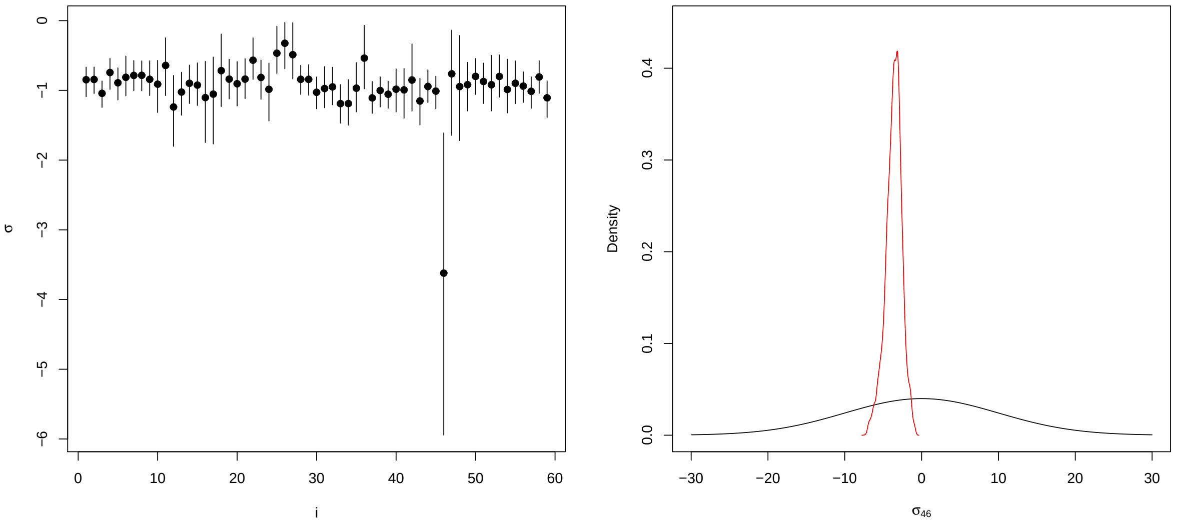


Figure 6.18: Plots of the posterior means and 95% credible intervals for ϕ_s against month s



(a) Posterior means and 95% credible intervals for σ_i

(b) Prior (black) and posterior (red) densities for σ_{46}

Figure 6.19: Posterior output for $\boldsymbol{\sigma}$ from the causation model

95% credible interval of (-5.946, -1.608), suggesting a much greatly reduced proportion of collisions taking place at site 46 once other factors have been accounted for, although there is relatively much lower certainty surrounding this estimate. It is not immediately clear why site 46 has such different results relative to other sites, for both σ_i as shown in Figure 6.19b, and for α_i shown in Figure 6.21b. It is possible, given the imperfect mixing of the chain that running for longer iterations and thinning by a greater degree may shrink the posterior distributions, in particular the posterior uncertainties, for σ_{46} and α_{46}, however this remains a subject for further investigation.

The posterior estimate for the global trend parameter β_t is 0.011, with 95% credible interval (0.001, 0.020), suggesting a very slight increase in the proportion of collisions due to darkness across the 59 sites, although this is unlikely to have any real impact.

We obtain posterior estimates for the site-specific trend parameter vector $\boldsymbol{\alpha}$, with posterior means and 95% credible intervals given in Figure 6.21a. From Figure 6.21a we observe again as expected, the majority of sites have $\alpha_i \approx 0$, again likely due to the zero-inflation component we incorporate into α_i. As with the case for σ_i we observe a clear outlier at site 46, with the posterior mean and 95% credible interval for α_{46} being -0.444 and (-0.809, -0.116) suggesting a highly significant deviation from the global trend at site 46, with proportions of collisions in darkness decreasing here at a much greater rate than any of other sites on the network. Some sites have a positive α_i, for instance site 11 has

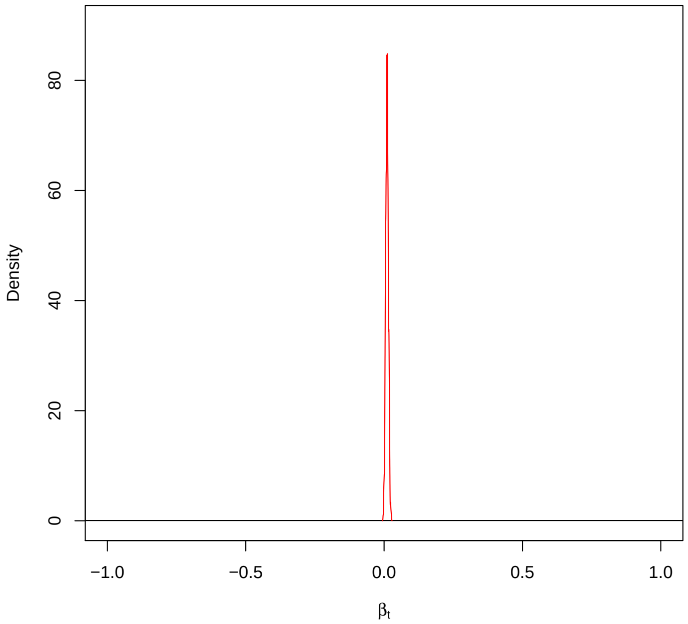


Figure 6.20: Prior (black) and posterior (red) densities for β_t for the causation dataset

a posterior mean of 0.100, with 95% credible interval given by (0, 0.208), suggesting the proportion of collisions in darkness here are increasing at a greater rate than any other site on the network.

As discussed in Section 6.5, a common goal for this analysis is to obtain a predictive distribution for the number of collisions due to each causation factor in a future time period. In order to do this we obtain the posterior distribution for $\pi_{i,s}(1)$ by summing the posterior distributions of the model parameters,

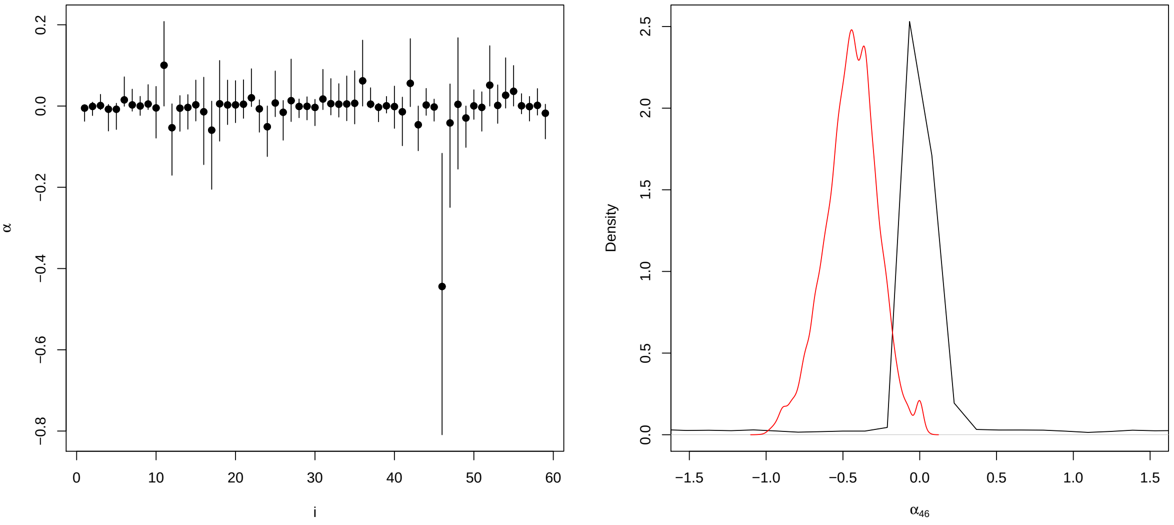
$$\pi\left(\pi_{i,s}(1)|\boldsymbol{y}\right) = \Phi\left(\pi\left(\beta_t|\boldsymbol{y}\right) + \pi\left(\sigma_i|\boldsymbol{y}\right) + \pi\left(\phi_s|\boldsymbol{y}\right) + \pi\left(\alpha_i|\boldsymbol{y}\right)\right).$$

Hence we are able to obtain a posterior predictive distribution for $y_{i,s,f}(1)$, the number of collisions of type f, in this case collisions in darkness, at site i in season s of the following time period,

$$\pi\left(y_{i,s,f}(1)|\boldsymbol{y}\right) = \pi\left(\pi_{i,s}(1)|\boldsymbol{y}\right)\pi\left(y_{i,s}(1)|\boldsymbol{y}\right).$$

Hence if we consider site 43 of the Halle zonal data, we obtain predictive output shown in Table 6.3.

From Table 6.3 we observe a possible seasonal effect, with higher numbers of collisions due to darkness expected in the late summer/autumn months of the following year, although this appears to be largely driven by the overall collision count fluctuations as



(a) Posterior means and 95% credible intervals for α_i

(b) Prior (black) and posterior (red) densities for α_{46}

Figure 6.21: Posterior output for $\boldsymbol{\alpha}$ from the causation model

s	1	2	3	4	5	6
$\pi\left(\pi(1)\|\boldsymbol{y}\right)$	0.14	0.15	0.13	0.20	0.14	0.14
	(0.07,0.23)	(0.08,0.23)	(0.07,0.21)	(0.10,0.32)	(0.07, 0.22)	(0.08, 0.22)
$\pi\left(y(1)\|\boldsymbol{y}\right)$	35.4	34.7	39.1	41.8	38.7	38.9
	(23, 49)	(23, 48)	(26, 53)	(28, 57)	(26, 53)	(26, 53)
$\pi\left(y_f(1)\|\boldsymbol{y},\pi\right)$	4.93	5.13	5.25	8.45	5.23	5.42
	(2.15, 9.18)	(2.41, 8.92)	(2.49, 9.21)	(3.81, 15.05)	(2.46, 9.55)	(2.59, 9.51)
s	7	8	9	10	11	12
$\pi\left(\pi(1)\|\boldsymbol{y}\right)$	0.22	0.18	0.17	0.18	0.15	0.16
	(0.12,0.34)	(0.10,0.28)	(0.09,0.25)	(0.08,0.30)	(0.07,0.23)	(0.09,0.24)
$\pi\left(y(1)\|\boldsymbol{y}\right)$	33.2	32.5	41.8	41.1	44.2	40.2
	(21, 46)	(21, 46)	(28, 56)	(28, 56)	(30, 59)	(27, 55)
$\pi\left(y_f(1)\|\boldsymbol{y},\pi\right)$	7.30	5.70	6.88	7.36	6.45	6.25
	(3.50, 12.8)	(2.66, 10.2)	(3.37, 11.8)	(3.12, 13.6)	(2.96, 11.2)	(3.10, 10.8)

Table 6.3: Table showing means and 95% credible intervals by month s for: the predicted proportion of collisions due to darkness ($\pi\left(\pi(1)|\boldsymbol{y}\right)$); the predicted number of total collisions ($\pi\left(y(1)|\boldsymbol{y}\right)$); and the predicted number of collisions due to darkness ($\pi\left(y_f(1)|\boldsymbol{y}\right)$) at site $i = 43$

opposed to the proportions of collisions due to darkness, which are predicted to remain largely constant between months.

Chapter 7

Conclusions

7.1 Summary

Throughout this thesis we have explored the challenges facing road safety practitioners, and analysed and developed statistical techniques to allow them to make best use of their data in order to address them.

In Chapter 1 we discussed the importance of maintaining and improving road safety, and costs incurred, both human and financial, due to collisions occurring on roads, both in the UK and around the world. We outlined the problems facing road safety practitioners who use data in order to inform their decisions, in particular the issue of regression to the mean, which can mislead conclusions drawn from data if not handled correctly. We discussed the history of regression to the mean, and provided clear real-world examples of the problems that it can cause. We then outlined the concept of safety performance functions, and how these can be used to overcome the issues posed by RTM, by harnessing external information from other sites in the network to help identify any unusual extreme data points which may have occurred at a particular site. We outlined the concept of Bayesian inference and described its utility when carrying out statistical analysis. We discussed Bayesian inferential methods such as Markov Chain Monte Carlo, which are necessary in order to execute the majority of statistical models, and heavily features in the implementation of the models developed in this thesis.

In Chapter 2 we focused on the issue of road safety scheme evaluation, and the statistical methods traditionally employed to carry out a scheme evaluation analysis. We discussed the potential role of RTM and trend in misleading a scheme evaluation analysis, and consequently why a naive before and after comparison would lead to biased conclusions regarding the estimate of treatment effect. We then discussed the merits of employing a Bayesian framework incorporating a safety performance function which allowed us to account for and estimate the RTM and trend effects present in the data, remove these, and thereby obtain a "cleaned" estimate of the treatment effect. We compared the Empirical Bayes and Full Bayes methods for carrying out a scheme evaluation analysis on the Northumbria dataset, noting how Empirical Bayes leads to a falsely optimistic estimate of posterior uncertainty of the treatment effect, and so Full Bayes is a preferable approach. We then demonstrated the usage of various prior distributions for the collision rate which are possible when using the Full Bayes approach, and compared the resulting models in terms of their goodness-of-fit to the data using the Deviance Information Criterion. We then discussed the relative merits of using the posterior distribution as a means of evaluating treatment effect, as is currently used in Empirical and Full Bayes analyses. We compared this with using the posterior predictive distribution instead, as this correctly accounts for uncertainty over future observations, thereby allowing us to properly estimate the uncertainty in our treatment effects, while accounting for trend in a more coherent way.

Chapter 3 was devoted to demonstrating and quantifying major methodological issues which cause problems with many forms of scheme evaluation analyses used as standard in the literature (including many of those discussed in Chapter 2). The main issue we sought to address affected all methods which make use of comparison data to inform the analysis - namely that of requiring exchangeable comparison pools of sites in order to minimise bias in estimates of RTM (and thus treatment) effects. We numerically demonstrated the risks incurred when non-exchangeable comparison pools are used, with a clear increase in RTM (and hence treatment effect) bias arising with decreasing exchangeability between treated and comparison pool. Methods for checking the exchangeability between a given comparison and treated pool were discussed, and the results of these tests were shown to correlate strongly with RTM bias in our simulated data – suggesting they are a good indicator of comparison data suitability. The method of propensity score matching was advocated as a proactive technique in determining site specific exchangeable comparison pools in Chapter 4. This method was then adapted into the propensity score weighted regression technique, which overcomes the issue of comparison pool exchangeability by weighting the SPF contribution of each individual comparison site by its similarity (as determined by the difference in propensity scores) to the treated site in question. This method was then applied to the simulated dataset, where it was shown to give more weight to sites found to be exchangeable with the treated pool (and hence induce a lower bias) and little to no weight to sites from non-exchangeable pools. This demonstrated PSWR's ability to autonomously extract the most suitable data from a large, heterogenous dataset, with no unnecessary data wastage nor requirement for external parameter selection. This method was then used to help identify site specific estimates of temporal trend in the data, by obtaining SPFs using covariate data collected in the before and after periods at each treated site, and including an “after” indicator variable in the second SPF. This again provides an entirely data driven approach toward trend estimation, which avoids assumptions of constant trend across the treated pool, avoids requirements of longitudinal data at each site, and makes no assumption of expert prior information.

In Chapter 5 we discussed an alternative issue facing road safety practitioners, namely road safety scheme allocation whereby dangerous hotspots on the network are identified as suitable locations for treatment. A model for proactive hotspot prediction was developed by extending the scheme evaluation methodology discussed in Chapter 2 in order to allow multiple time periods of data in which to fit the model, with additional parameters to account for site specific deviations from the global model included, as well as a mechanism to provide more weight to recent observations when forming predictions. This model was validated against real traffic datasets and was shown to predict accurately compared with the true observed collision counts, even with a reduced number of years of data, highlighting the model robustness and availability to datasets of all sizes.

The model was then developed further in Chapter 6 where the potential for including a spatiotemporal structure was motivated using a dataset from Florida, USA. These effects were included into the hotspot prediction model, with a CAR model being used to model seasonal effects in order to allow for sharing of information between consecutive seasons, and a kernel density smoother being used to model spatial effects, in order to allow for sharing of information between geographically similar locations. Furthermore the response variable was disaggregated in order to enable modelling by severity in order to allow for a more precise reflection of the level of danger at each location using an ordinal probit model, before a logistic model was described which allowed for proactive prediction of collision totals by type, hence allowing the practitioner to be better informed as to where treatment should be allocated.

In Abstract A we discuss the various practical applications for the research, and the real world impacts it can make, and has already made. Particular focus is paid to the RAPTOR suite of software applications, developed using the `Shiny` package within the statistical software `R`, which are designed to make the complex statistical methods discussed in this thesis available to all practitioners, regardless of their level of statistical knowledge.

The scheme evaluation app described in Section A.1.2 is demonstrated as a means to employ a Fully Bayesian scheme evaluation analysis as described in Section 2.2, and provide practitioners with estimates of the RTM, trend and treatment effects present in their data. The data is inputted in the form of two spreadsheets, corresponding to the comparison sites, from which the SPF is built, and the treated sites which are to be evaluated. There are options to investigate the suitability (i.e. exchangeability) of the comparison and treated sites, in order to avoid the dangers of biases in the RTM and treatment effect estimates, as described in Section 3.2. The data input files are in the form of a spreadsheet, with rows corresponding to each individual site, and columns containing collision counts and covariate data. There are additional options for experts to impart any prior knowledge they may have by modifying the prior distributions of the model parameters. This results are provided both graphically, in the form of a bar chart, and numerically, in the form of a table, thereby making it easier for the effectiveness of the scheme to be discerned. In addition, information regarding the SPF fitted as part of the analysis is provided, in the form of a summary table displaying the MLEs and accompanying p-values for each of the covariates provided, allowing the practitioner to see what effect (if any) the covariates had on the number of collisions on the network.

The hotspot prediction app described in Section A.1.3 implements the Bayesian hierarchical hotspot prediction model as described in Chapter 5. The data inputs are largely the same as for the scheme evaluation app, however in this case since no comparison data are needed there is only one dataset which is needed to be uploaded. Additionally since in most cases there should be multiple observations at each site (one for each time period), each row should correspond to a site in a given time period, and hence the uploaded file should have nn_{years} rows and n_p columns. Again the user has the option of imparting any prior knowledge they may have into the analysis, most notably with regards to their beliefs regarding the diversity of their network as it pertains to the likelihood of there being site-specific deviations from the global trend, and the rate at which the network has evolved and so past observations become less informative (by adjusting the prior distributions for b_n and τ respectively). As with the scheme evaluation app there is the functionality to alter the length of the MCMC chain to adapt to the users needs, although the defaults are expected to be used in most cases. Finally the results tab displays the output of the analysis, comprising of:

- A table displaying summary statistics for the posterior predictive distribution for the number of collisions in the following time period
- Plots showing the posterior means of λ and μ alongside the collision counts through time, along with the posterior predictive mean and 95% prediction interval for the future number of collisions. These plots can be directly exported for use in reports etc.
- A histogram of the posterior predictive density for the number of collisions in the following year
- A table summarising the SPF, including maximum likelihood estimates and p-values for each covariate

- A ranked and colour coded list of sites by their posterior predictive likelihood of exceeding a user specified threshold

The multinomial proportions application provides a method by which practitioners can investigate contributory factors toward collision totals. This was designed to address the potential gap in utility of the scheme evaluation and hotspot prediction apps which analyse collision totals, but not why these totals occurred. It is therefore hoped the multinomial proportions app would provide insight for practitioners as to which countermeasures would be most successful and hotspots once they have been identified. The method for the user to provide their data is simple, the data should be of the form of a contingency table containing a breakdown of collisions across the sites being investigated into the different conditions the user suspects there may be a difference between (i.e. if the user believes there are significantly more collisions at night than in the day, the contingency table would be broken down into different periods of day e.g. morning/afternoon/night). The output from this app is based around the p-value obtained from the hypothesis test that there is no significant difference between the conditions being investigated, along with an interpretation of the results for the user.

There are a number of users representing different organisations across the globe (New York Department of Transport, Highways England, Transport for London) whom have requested access to the RAPTOR applications both for the purposes of hotspot prediction and scheme evaluation. The usage of these methodologies is helped further by the addition of the hotspot prediction methodology to PTV Group's Visum Safety software, expanding the application's reach to PTV Group's global clientbase.

7.2 Further Research

Whilst the research and models established thus far in this thesis have been demonstrated to be effective in their own right there remain several possible avenues of research which could improve them yet further. Here we discuss several concepts which, due to either time constraints or lack of suitable data, have not yet been implemented, but would be potential extensions to the research outlined thus far.

7.2.1 Safety Performance Functions

The quantity of literature devoted to the derivation of safety performance functions in their own right, without a particular application e.g. scheme evaluation or hotspot prediction, is vast (see Chapter 1). As discussed in Section 1.3, we wish to avoid models which are dataset specific, that is - have been developed specifically for a certain dataset/location, and which have excessive data requirements in order to be fitted (e.g. time series models, which require a large number of sequential observations in order to achieve a good fit). The generic log-linear SPF described in Section 1.3 fulfils this criteria, since it can be applied when any number of consecutive observations are available (provided the number of data points is greater than the number of parameters, i.e. $nn_y > n_p$). Various alternatives were discussed in Section 1.3, where the possibilities of using GWR to form the SPF, in order to allow for geographically varying regression coefficients, as well as the possibility of using a DAP prior distribution for the regression coefficients, allowing a subset of the available data to inform our beliefs regarding the marginal and covariance structure of the regression coefficient parameter vector $\boldsymbol{\beta}$ as demonstrated in Section 6.2.

A key aspect of SPF development which has not been discussed in depth here, is that of covariate selection. As mentioned in Section 1.3, the standard approach used in the analyses in chapters 1, 2, and 5 is to fit a frequentist regression using the `glm.nb` command of the `MASS` package, and perform a backwards elimination to remove any non-significant covariates (as described in [Miller, 1984]). While this approach is time-tested and ensures all covariates used in formulating the SPF have significant p-values, this dependence on a frequentist technique may be seen as ill fitting to a Bayesian framework. There are a variety of Bayesian techniques which can be employed in place of backwards elimination, and thus make our modelling approach more self-consistent. Unfortunately many of these, such as a simple comparison of fit using various information criteria e.g. AIC/BIC/DIC, become unfeasible since they require the comparison of 2^{n_p} models, which clearly becomes computationally difficult for any large n_p. An alternative, proposed by [George and McCulloch, 1993], makes use of a Gibbs sampling procedure, as outlined in Section 6.1.1 to find an appropriate subset of all available covariates, to be selected to form the SPF. The method works by defining a latent vector of binary variables $\boldsymbol{\gamma}$, and assigning a mixture distribution to each element of the regression coefficient vector $\boldsymbol{\beta}$,

$$\begin{aligned} \beta_j|\gamma_j &\sim (1-\gamma_j)\, N\left(0, \tau_j^2\right) + \gamma_j N\left(0, c_j^2\tau_j^2\right), \qquad j = 1, \dots, n_p \\ \gamma_j &\sim Bern(p_j). \end{aligned}$$

Fixing $c_j > 1$ to be large, and $\tau_j > 0$ to be small, essentially results in the case where $\gamma_j = 0$ giving very little probability mass for β_j away from 0, with the converse being true for $\gamma_j = 1$. We can therefore interpret p_j as the prior probability for covariate j being included in the model, which allows for the inclusion of any expert prior knowledge which may be available, an option not available through standard backwards elimination. Many choices are available for the structure of $\pi(\boldsymbol{\gamma})$, assuming independence between covariate probabilities (which is not a requirement, although would often be reasonable) results in a binomial likelihood,

$$\pi(\boldsymbol{\gamma}) = \prod_{j=1}^{n_p} p_j^{\gamma_j} \left(1 - p_j\right)^{(1-\gamma_j)}$$

where if prior ignorance, i.e. $p_j = 0.5 \,\forall\, j$, is assumed, we simply have

$$\pi(\boldsymbol{\gamma}) = 2^{-p}.$$

7.2.2 Obtaining Traffic Sites

A potential issue with the techniques we have developed for both scheme evaluation and hotspot prediction, is that both require data from fixed locations to which accidents can be attributed, as seen in chapters 2 and 5 as site or nodes. This can cause a potential issue for road safety practitioners since clearly multiple collisions are highly unlikely to occur at the same exact location, and so collisions must be grouped, or “clustered” in order for the concept of a site to be meaningful. There are a variety of methods to carry out this clustering, and as such not all site clusterings will be the same, which can therefore lead to the overall conclusions of an analysis being dependent on the clustering used. Clearly it would be preferable to avoid inconsistency in results as a result of site clustering, which can be achieved by adopting a clustering algorithm into the overall modelling framework, enabling the analysis of raw road safety data without the need for pre-processing. This has the further advantage of enabling the methods to be accessed by practitioners who do not have access to mapping/clustering techniques, and so as of yet would be unable to make use of the scheme evaluation and hotspot prediction models. There are existing commercial software tools (e.g. *Visum* [PTV Group, a] and *AccsMap* [Buchanan Computing,]), as well as extensive research (e.g. [Deka and Quddus, 2014] and [Imprialou et al., 2014]) in the area of accurate collision mapping which clearly also plays a huge role in ensuring accurate clustering, as does data quality (see Section 7.2.4).

7.2.3 Scheme Evaluation

The research into scheme evaluation studies concluded by recommending a propensity score weighted regression approach, so as to avoid the issues caused by non-exchangeable comparison pools, the dangers caused by which were discussed and quantified in Section 3.3. While we are satisfied this approach gives an improvement on the standard Fully Bayesian framework used in scheme evaluation studies with regards to accounting for comparison pool exchangeability, there are further improvements which could be made. Possibly the most natural next steps of research would be to consider the hotspot prediction model, particularly the extensions made in Chapter 6 as a means towards extending the scheme evaluation equivalent. Given, as explained in Chapter 5, the hotspot prediction model is an extension of the methodologies used in Chapter 2 to carry out scheme evaluation analyses, it stands to reason that additional developments to the hotspot prediction model may also be suitable for the scheme evaluation counterpart. One particular area which we may seek to develop further, is the SPF component of the scheme evaluation analysis. We have discussed already in Section 7.2.1 how we may improve SPFs by considering techniques for variable selection which of course can be applied to the formation of the SPF used in scheme evaluation studies, however we can further consider adapting the SPF structure itself. In Section 6.2, modifications to the SPF structure were proposed, with the aim of removing unrealistic assumptions made in the standard log-linear structure, thereby forming a more realistic model, and therefore a model which provides a better fit. Since there is no reason to suspect different variables would be used for scheme evaluation or hotspot prediction it stands to reason the same adaptation can be employed in both analyses. Hence we could investigate the possibility of correlation between covariates by using a DAP, as described in Section 6.2.1, with a subset of the comparison dataset (should it be large enough) being used to inform the multivariate Normal prior distribution on the regression coefficient vector $\boldsymbol{\beta}$.

A further development to consider, should the scheme evaluation study take place over a sufficiently large area, would be a spatial aspect to the data. While clearly, as demonstrated in Section 5.5, it is not possible to include a site effect parameter in the model when only a single data point is available (as we assume for the scheme evaluation case, including multiple years leads to the model described in sections 5.3 & 5.4), it may be possible should the data be on a zonal scale. As discussed in Chapter 6, it is common for traffic data to be considered on a zonal scale, often referred to as traffic analysis zones (TAZs), in cases such as the USA, where large areas must be analysed at once. Whilst in Chapter 2 we mainly discussed schemes on a local scale, implemented by a local authority, it is entirely possible to investigate schemes on a national level, implemented by governments, and so it becomes highly possible for locations to be grouped into zones in order to aid this analysis. It therefore appears logical to consider a zonal effect within the model, either as an explicit parameter, as with σ in Chapter 5, or simply as an indicator variable within the SPF to as to better inform μ.

7.2.4 Hotspot Prediction

We have already discussed many avenues by which the generic hotspot prediction model described in Chapter 5 can be developed, with these given in Chapter 6. These cover the vast majority of features found in current hotspot prediction methods, and so improvements to the predictive capability cannot easily be obtained without redesigning the entire model structure and/or losing model generalisability, which is a core focus of the research. Despite this there are additional elements we may consider which can further improve the model's utility.

Data Issues

Perhaps one of the largest barriers to practitioners being able to make use of statistical models such as the hotspot prediction model outlined in chapters 5 & 6 is the requirement to have sufficient data available to build the model. Whilst great effort has been made to make the data requirements for the model as non-restrictive as possible, for some authorities it is still highly difficult to have good quality data, spanning a number of sites across the network for a number of years. This problem is exacerbated further by the model requirement for all data points to be available, that is we cannot allow for missing values. This places a great deal of further pressure on authorities where data is scarce, since often for various reasons records/measurements may be missing an observation for a particular year. This becomes problematic for the current hotspot prediction model since this can lead to sites, or in extreme cases entire covariates, being removed from the analysis, exacerbating issues caused by small datasets. An obvious improvement to this problem would be to allow our model to handle missing data, thereby removing the need to not include sites/covariates for which there are missing values. One such solution to this would be to incorporate data augmentation methods in order to approximate missing data, from data which is available. The adoption of the Bayesian framework makes this process very simple, whereby models can be assumed to describe covariate distributions, and hence missing data values can be imputed, with the uncertainty attached to the modelled missing value naturally handled by the Bayesian paradigm. Such analyses are made simpler by the fact that missing road safety data can be sometime considered missing completely at random (MCAR), i.e. missingness is completely independent of all data, or missing at random (MAR) i.e. missingness is independent of the true value of the missing data, since whether observations are recorded can depend on issues such as funding availability, independent of the covariate data itself. This may not always be the case though, since for example the decision to monitor speeds at a location may only be taken when it is believed speeding is a problem, thereby meaning the missing speeds could be assumed to be lower than the observed speeds, meaning the data in this case would be missing not at random (MNAR). The classification of missing data would therefore have to be taken on a case by case basis, with the modelling solution, and indeed whether modelling the missing data is worthwhile, dependant on this classification.

A related issue to this which is also of crucial importance to accurate modelling is that of data quality. The majority of collision data is obtained directly from reports from the attending police officer, and so there remains the possibility that information regarding the incident could be incorrectly reported. This is particularly problematic where there are areas of subjectivity, e.g. the primary causes of the collision, the severity of the collision etc, but even more fundamental information can be incorrectly reported (e.g. the location of the collision, and fixed covariate data such as the road class and speed limit at the location). The issue of data quality is known among practitioners, and has also been investigated in literature (for example [Imprialou and Quddus, 2017]). Clearly all of these present a problem for accurate collision modelling, and so it is important to account for this in order to achieve the aim of receiving collision data in raw form, and so some form of data cleaning will need to take place in order to ensure raw data collected is accurate.

Non-linear Trend

A key component of the hotspot prediction modelling framework, discussed in depth in Section 5.4 is the estimation of temporal trends in collision counts (and hence the underlying collision rate) in order to allow for accurate predictions regarding future levels of safety at each site. To this end there are two trend components, global and site-specific, included within the modelling framework. The variance inflation structure, described in Section 5.3 ensures recent trends are acknowledged more heavily than any apparent trend patterns found further into the past, again to ensure future predictions are as accurate as possible. However practitioners may not solely be interested in future predictions, they may also simply be interested in analysing how the collision rate has evolved over time at sites on the network. To this end, the current assumption of a single constant trend at each site throughout the data may not be realistic, with the possibility of events occurring during the span of the data which cause the trend to change, meaning a composite function may be more suitable. There is a significant quantity of research in road safety devoted to the analysis of change points in trends, usually within the context of interrupted time series models (see, for example [Ihueze and Onwurah, 2018]). For sufficient data such an approach would also provide a strong mechanism for scheme evaluation (since we would know implementation of the treatment to be the change point), however these require large numbers of observations in order to be fitted well, a restriction we do not expect to be feasible for a large quantity of practitioners and hence one we wish to avoid in this research. Despite this we can still investigate the possibility of interrupted or simply non-linear trends in road safety data, so as to achieve a better fit of our model throughout the data, although we would not expect this to impact predictions severely (since, as discussed, the variance inflation component provides stronger weight to more recent/current trends).

7.2.5 Impact Work

As discussed in Appendix A the main driving force behind this research has been the opportunity and potential for these methods to be used by practitioners. In order to facilitate this, the *RAPTOR* suite of software applications has been developed to allow practitioners to implement the methods discussed in this thesis, without any technical or computational requirements being made (see [Matthews et al., 2018]). While the software applications are still relatively new, and so we expect minor issues to arise and be resolved in time, there are major features which could be added which would strongly enhance the software capabilities.

We have purposefully maintained a Bayesian framework for the models for both scheme evaluation and hotspot prediction, and whilst in the general models given in chapters 2 and 5 we have assumed prior ignorance when specifying prior distributions, this does not have to be the case in practice. We would not expect every practitioner to have meaningful prior beliefs regarding every parameter incorporated in each model, however it is reasonable to believe some practitioners would have significant expert knowledge regarding conditions on their networks, particularly as it pertains to components such as temporal trends. Whilst there is the facility for experts to adjust, and hence inform, the prior distributions for the parameters in each model, this process currently would require some statistical knowledge (the ability to select distributions and make meaningful adjustments to the prior mean and variance) which may not be possible for all experts. It would therefore be a significant improvement to incorporate an elicitation mechanism, such as the MATCH [Morris et al., 2014] or SHELF [O'Hagan and Oakley, 2019] elicitation toolkits, which provide intuitive mechanisms for experts to impart their prior beliefs in a meaningful way.

Given the spatial nature of the analyses being carried out, particularly in the case of the hotspot prediction model, it would be of benefit to road safety practitioners analysing data to be able to view the analysis output on a map of the network rather than simply as a list of sites. Not only would this lend itself better to visual communication of information, it can also help diagnose causes of effects on the network, e.g. if sites for which treatment were all situated close together, and likewise for sites where the treatment was less effective, it may provide some clue as to why this is the case. The ability to display output for the hotspot prediction model is available as part of *Visum Safety* software, however it would be beneficial, and possible, to incorporate mapping capability into the RAPTOR software applications also.

7.2.6 Real Time Modelling

The majority of research discussed in this thesis has revolved around the concept of road safety treatments being applied at a fixed decision making point in time, after road safety data has been collected and analysed. Modern technology however allow for the possibility for data to be analysed, and decisions made in real time, as opposed to at pre-defined intervals. The advantages to this are obvious, extending the "proactive rather than reactive" mantra which motivated the hotspot prediction model described in chapters 5 and 6 to allow for proactive response to danger as soon as it becomes apparent. Data such as fixed sensors at potential hotspot locations, coupled with real time weather and mobile phone data for example, can give up to the minute estimates of covariate values included in an SPF and any collisions which have occurred, allowing for real time analysis and decision making to remove danger detected on the network. Research has already begun in this area (see for example [Hossain and Muromachi, 2013],[Sun and Sun, 2015],[Wang et al., 2015]) and this would provide an important addition to our research, particularly as pertains to the software applications for practitioner usage, discussed in Chapter A.

Appendix A

Impact Work

Clearly the motivations behind this research are heavily applied in nature, and indeed the work originated from a request to carry out a scheme evaluation analysis on 56 speed cameras carried out by the Northumbria Safety Reseach Initiative. Due to this my research has always hhas

A.1 Translational Research

Given the complex nature of the statistical modelling involed in this research, particularly in the hotspot prediction models discussed in Chapters 5 and 6, these models are usually only accessible to individuals with a significant amount of training in statistics. Since most road safety practitioners are usually transport/engineering experts, and usually have little statistical background, this can lead to problems in the research methods being implemented in practice. This is evidenced by the fact methods such as Empirical Bayes are still being used in practice (REFERENCE) despite multiple researchers evidencing it is an inferior method for carrying out a road safety analysis, as demonstrated in chapter 2. Clearly therefore efforts need to be made to address this divergence between methods used by practitioners, and thus shown by research to perform better. Attempting to make academic research become accessible available to practitioners is known as translational research (REFERENCE) and has been a key component of this work, as we have consistently attempted to engage practitioners with the methods developed in a bid to convince them to implement these methods themselves. There are multiple approaches to carry out translational research, the approach we felt best for this particular situation was to develop software applications to allow practitioners to access the various road safety analysis methods. Fortunately there is an add-on package in `R` which provides a GUI to an existing R script, known as `Shiny`. The usage of a GUI to access the model code therefore provides the possibility of an individual executing the analyses described thus far, and obtaining the relevant output, without needing to access the underlying code, nor be fully versed in the statistical methodologies underpinning the approaches. This therefore has potential to overcome which arise when practitioners lack any prior statistical expertise, and thus are unable to implement the methodologies, by removing any need for them to work with the statistical models themselves, rather allow the GUI to carry out the interaction for them.

A.1.1 Shiny Applications

Shiny applications, built using the package `shiny` in `R` are comprised of two functions: `server` and `ui` (these functions can equally either be in a single file or separate `server` and `ui` files). Both of these functions take two arguments: `input`, a list containing the data/options provided by the user; and `output`, a list containing the outputs provided by the application. The server function is responsible for carrying out the actual computation, in this case fitting the road safety models to the data, based on the input list provided by the user. These inputs can take a variety of forms, most commonly an uploaded dataset(s) and/or options selected from the various "widgets" (methods for inputting a value - e.g. drop down menu, slider bar, typed input). These inputs then specify the analysis...

Once the analysis is complete, the output list is then populated with the desired results. Outputs can again take a variety of forms, most commonly numeric outputs either individually or as a table, along with any plots of the resulting data. The means by which the inputs are provided and the outputs returned are controlled by the `ui` function, which as the name suggests, controls the interface by which the user accesses the analysis. It is the ui which allows the user to interact with the methodology without needing to understand the mechanics of the underlying model, as the ui formulates the computation to fit the input, and provides the output directly. In the context of the road safety analyses, the user input would primarily be the road safety datasets uploaded as a datafile (`.csv`, `.txt` etc) as well as the option for any expert prior knowledge should it be available. The output would primarily be tabulated numerical results, in the case of scheme evaluation - posterior estimates of the RTM, trend and treatment effects, and for hotspot prediction - the posterior predictive distribution for the number of collisions in the following time period. Further details on these applications can be found in Sections A.1.2 and A.1.3. Although here we make use of the R package to easily assimilate existing model code into shiny applications, they can equally be written in other languages such as JavaScript, which allow for people with web development skills to build specialist Shiny apps, while the standard R interface allows R users without web development skills/knowledge to also use Shiny.

A potential barrier in the usage of Shiny by practitioners would be the computational requirement still present, in order for practitioners to access any Shiny applications built, they would have to have R installed, along with any packages required by the analysis. Additionally there is the added drawback that the analyses can potentially be computationally expensive in terms of computation time, and so if computing resources are limited, the prospect of surrendering them for the sake of these in depth analyses may become unappealing to practitioners. In order to circumvent these issues, the decision was made to host the applications on a password protected external server, with practitioners being given login credentials in order to access it. This therefore immediately solves the aforementioned issues, since all necessary software and packages are downloaded to the server, there is no requirement for the user to download anything to access the apps (where issues such as administrator permissions/firewalls can cause problems). From this we can also monitor how many users have signed up for access to the RAPTOR suite of applications. At the time of writing there are 42 industry practitioners representing 19 organisations from the U.K. and worldwide that have requested access to the RAPTOR applications with a view to using them to analyse their road accident data (this does not include individuals from academic institutions that have been granted access). The RAPTOR suite has been promoted via various meetings with practitioners held around the U.K. in addition to attending conferences such as the Transport Practitioners Meeting (TPM) and the Annual Meeting of the Transportation Research Board (TRB) which are frequently well attended by road safety practitioners.

A.1.2 Scheme Evaluation

The scheme evaluation application was developed to allow users to carry out a fully Bayesian (FB) statistical analysis to determine the effectiveness of a given safety countermeasure deployed on their network. As discussed in Chapter 2, the most prevalent methods for scheme evaluation employed in practice are the Four Time Period and Empirical Bayes methods. We then discussed the methodological failings of these approaches, namely their oversimplistic and restrictive modelling assumptions, compared with methods such as FB which overcome these. However FB, by its very definition relative to EB, requires much more computational and statistical skill to deploy, one of the primary reasons for its lack of prevalence in practice relative to other more simplistic methods. The scheme evaluation app was developed to overcome these issues and thus enable practitioners to make use of research endorsed methods to carry out scheme evaluation. The analysis comprises several stages, and these are represented by different tabs within the application.

Data Upload

The first tab of the scheme evaluation app is the data upload tabwhich provides the user with a means to upload the data they wish to be analysed. The application requires two datasets to be uploaded, data from the untreated comparison sites (here called the "reference" dataset), and data from the sites which have been treated which are to be analysed. The datafiles can be in a variety of formats, with the default being `.csv` files, since Excel spreadsheets have been found to be the most common used by practitioners, however the upload file format can be edited by altering the "Separator" and "Quote" options in the data upload boxes. The "data preview" tables allow the user to ensure they have selected the correct format and that the data looks as they would expect. The datasets themselves should have a specific structure:

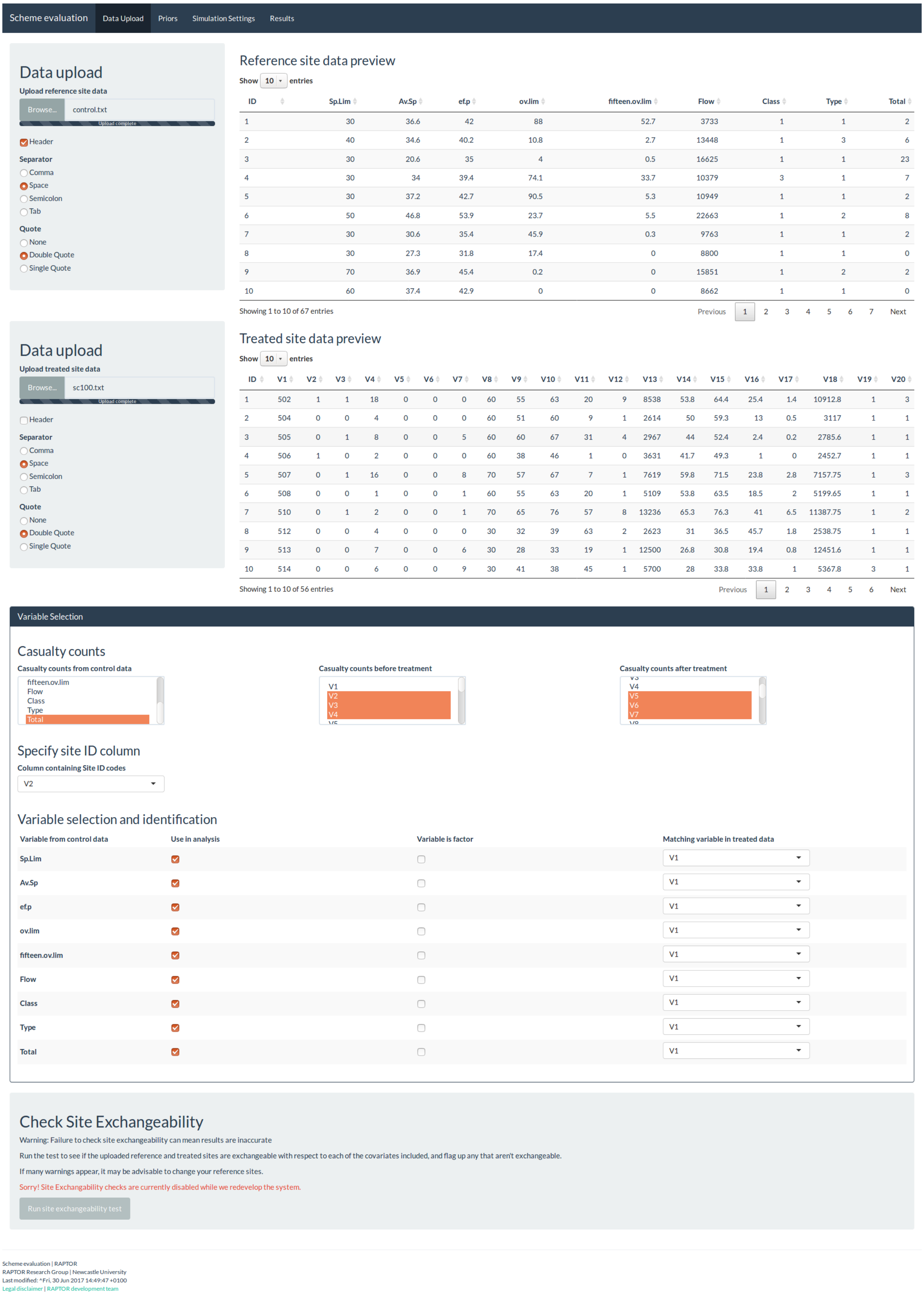
1. Each row in both datasets should correspond to a single site only, and each site should appear only once in the dataset. There should be an entry in each row identifying the site (site ID), which are used to identify the results of the analysis for each individual site.

2. There must be a column giving the number of collisions/casualties in the before period for both the treated and comparison datasets, and an additional column giving the totals in the after period for the treated dataset.

3. The remaining columns should describe the covariate information used to fit the SPF, with each column corresponding to an individual covariate, and each covariate should appear only once. The covariates themselves can be anything deemed by the practitioner to have a potential impact on the collision total, and these data can be of any type (discrete/categorical/continuous etc). The covariates used must be the same for both the comparison and treated datasets however, in order to be able to fit the SPF to the comparison sites and then apply the fitted model to the treated sites. The "Variable Selection" box allows the user to specify which columns contain which the collision counts and site ID indicators, as well as matching the covariates between comparison and treated datasets (although the application itself will attempt to do this based on the dataset column headings).

4. There must be no missing values in terms of collision or covariate count for any of the sites in either dataset, and so any sites for which all of the required information is not available are removed from the analysis.

Once the datasets have been uploaded the user can (and is encouraged to) then check the exchangeability of the comparison and treated datasets, so as to avoid any potential biases in the estimates of RTM and treatment effect, as discussed in Chapter 3. The checks



Data upload

Upload reference site data

Browse... control.txt

Upload complete

☑ Header

Separator

○ Comma
● Space
○ Semicolon
○ Tab

Quote

○ None
● Double Quote
○ Single Quote

Reference site data preview

Show 10 entries

ID	Sp.Lim	Av.Sp	ef.p	ov.lim	fifteen.ov.lim	Flow	Class	Type	Total
1	30	36.6	42	88	52.7	3733	1	1	2
2	40	34.6	40.2	10.8	2.7	13448	1	3	6
3	30	20.6	35	4	0.5	16625	1	1	23
4	30	34	39.4	74.1	33.7	10379	3	1	7
5	30	37.2	42.7	90.5	5.3	10949	1	1	2
6	50	46.8	53.9	23.7	5.5	22663	1	2	8
7	30	30.6	35.4	45.9	0.3	9763	1	1	2
8	30	27.3	31.8	17.4	0	8800	1	1	0
9	70	36.9	45.4	0.2	0	15851	1	2	2
10	60	37.4	42.9	0	0	8662	1	1	0

Showing 1 to 10 of 67 entries

Previous 1 2 3 4 5 6 7 Next

Data upload

Upload treated site data

Browse... sc100.txt

Upload complete

☐ Header

Separator

○ Comma
● Space
○ Semicolon
○ Tab

Quote

○ None
● Double Quote
○ Single Quote

Treated site data preview

Show 10 entries

ID	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
1	502	1	1	18	0	0	0	60	55	63	20	9	8538	53.8	64.4	25.4	1.4	10912.8	1	3
2	504	0	0	4	0	0	0	60	51	60	9	1	2614	50	59.3	13	0.5	3117	1	1
3	505	0	1	8	0	0	5	60	60	67	31	4	2967	44	52.4	2.4	0.2	2785.6	1	1
4	506	1	0	2	0	0	0	60	38	46	1	0	3631	41.7	49.3	1	0	2452.7	1	1
5	507	0	1	16	0	0	8	70	57	67	7	1	7619	59.8	71.5	23.8	2.8	7157.75	1	3
6	508	0	0	1	0	0	1	60	55	63	20	1	5109	53.8	63.5	18.5	2	5199.65	1	1
7	510	0	1	2	0	0	1	70	65	76	57	8	13236	65.3	76.3	41	6.5	11387.75	1	2
8	512	0	0	4	0	0	0	30	32	39	63	2	2623	31	36.5	45.7	1.8	2538.75	1	1
9	513	0	0	7	0	0	6	30	28	33	19	1	12500	26.8	30.8	19.4	0.8	12451.6	1	1
10	514	0	0	6	0	0	9	30	41	38	45	1	5700	28	33.8	33.8	1	5367.8	3	1

Showing 1 to 10 of 56 entries

Previous 1 2 3 4 5 6 Next

Variable Selection

Casualty counts

Casualty counts from control data

fifteen.ov.lim
Flow
Class
Type
Total

Casualty counts before treatment

V1
V2
V3
V4
V5

Casualty counts after treatment

V3
V4
V5
V6
V7
V8

Specify site ID column

Column containing Site ID codes

V2

Variable selection and identification

Variable from control data	Use in analysis	Variable is factor	Matching variable in treated data
Sp.Lim	☑	☐	V1
Av.Sp	☑	☐	V1
ef.p	☑	☐	V1
ov.lim	☑	☐	V1
fifteen.ov.lim	☑	☐	V1
Flow	☑	☐	V1
Class	☑	☐	V1
Type	☑	☐	V1
Total	☑	☐	V1

Check Site Exchangeability

Warning: Failure to check site exchangeability can mean results are inaccurate

Run the test to see if the uploaded reference and treated sites are exchangeable with respect to each of the covariates included, and flag up any that aren't exchangeable.

If many warnings appear, it may be advisable to change your reference sites.

Sorry! Site Exchangability checks are currently disabled while we redevelop the system.

Run site exchangeability test

Scheme evaluation | RAPTOR
RAPTOR Research Group | Newcastle University
Last modified: ^Fri, 30 Jun 2017 14:49:47 +0100
Legal disclaimer | RAPTOR development team

Figure A.1: The "Data Upload" tab of the scheme evaluation RAPTOR app.

work by carrying out the post-hoc tests described in Section 3.4, namely permutation tests on each of the uploaded covariates, along with a permutation test on the Mahalanobis distance of the two datasets in their entirety. Each permutation test which generates a significant result (a p-value of less than 0.05) generates a warning message and the user is encouraged to investigate any problems with the comparison dataset should these amount to a significant proportion of the covariates included (with particular emphasis on the test relating to the Mahalanobis distance).

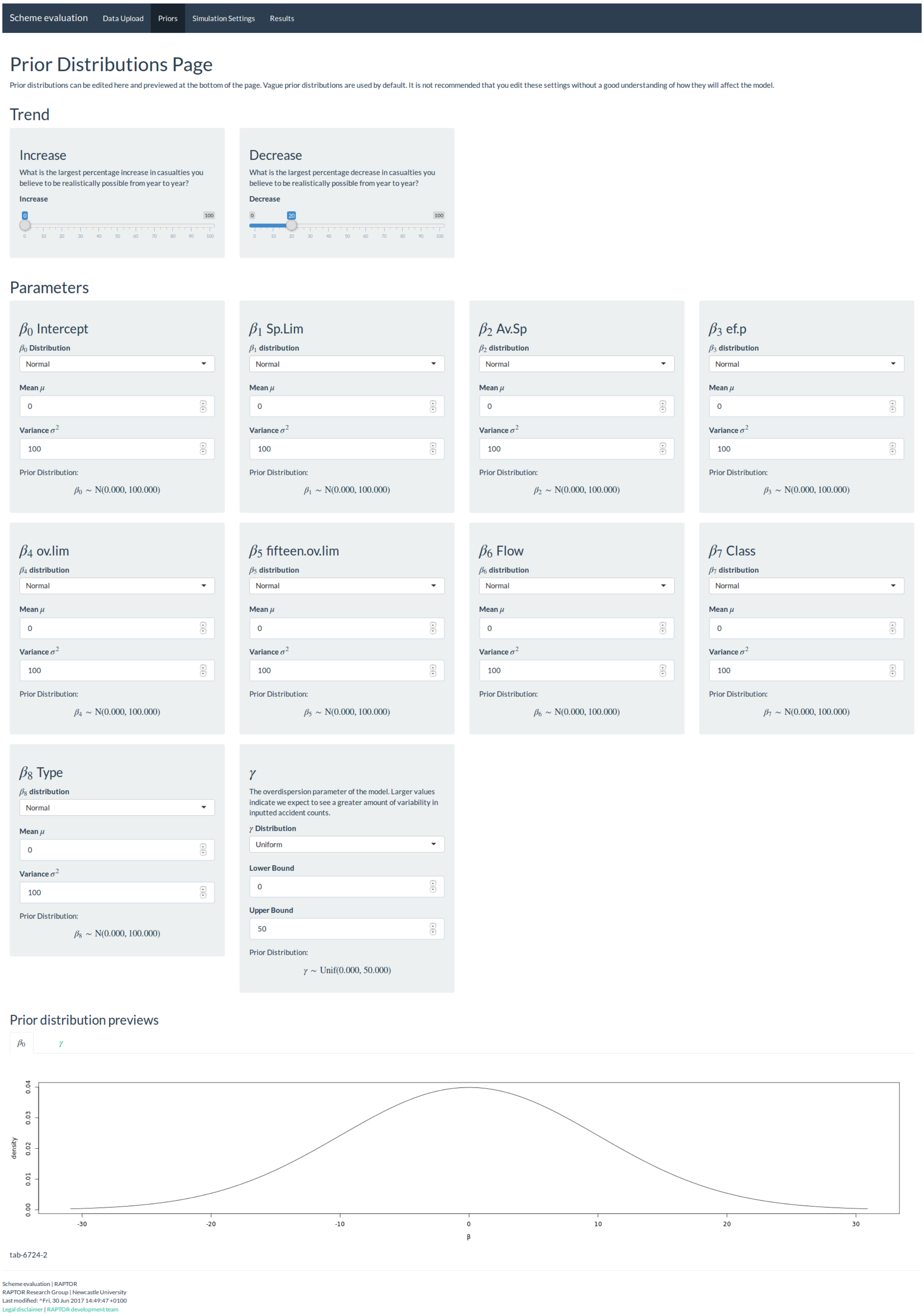
Priors

Whilst the main purpose of developing these applications was to remove any need for statistical knowledge from the user, it is important, particularly given the Bayesian framework from which the models are built, not to discount their potential for prior knowledge. In order to allow for this, the Priors tab allows the user to adjust the default vague priors attached to each regression coefficient,

$$\beta_j \sim N(0, 10^2), \qquad j = 1, \ldots, n_p \tag{A.1}$$

by changing the prior mean and variance in accordance with their prior beliefs. This is due to the likelihood that practitioners, particularly those more experienced, will have reasonably strong beliefs regarding how some covariates (particularly major covariates such as AADT, speed limit etc) will affect collisions on their network, and so it would be foolish not to allow this information to be included in the analysis. As has been the theme throughout this research however, this is by no means a requirement, and the standard vague prior distributions given in equation A.1 would be fitted should no adjustments be made, and have been shown to give satisfactory results (see Chapter 2). Text is included, describing the effect of adjusting each hyperparameter in order to help inform the user how best to include their beliefs, and plots of each resulting distribution are provided at the bottom of the page as a visual aid in order to help satisfy the user the prior distributions are as they would expect. It is expected that whilst some practitioners will have prior beliefs they wish to include, many may wish to simply run the standard analysis and so would ignore this tab, and particularly for large datasets it is not expected that this will come with any large cost.

What is significantly more significant for this tab is it provides the opportunity for the user to provide their beliefs regarding the effect of temporal trend across the treated sites. We allow the user to specify a range of possible values for the trend to take, by specifying an upper and lower bound of the likely percentage change in collision totals due to trend between the before and after periods, which translate to the upper and lower bounds of the Uniform distribution for the trend parameter ξ as specified in Sec-



Prior Distributions Page

Prior distributions can be edited here and previewed at the bottom of the page. Vague prior distributions are used by default. It is not recommended that you edit these settings without a good understanding of how they will affect the model.

Trend

Increase

What is the largest percentage increase in casualties you believe to be realistically possible from year to year?

Increase: 0

Decrease

What is the largest percentage decrease in casualties you believe to be realistically possible from year to year?

Decrease: 20

Parameters

β_0 Intercept

β_0 Distribution: Normal

Mean μ: 0

Variance σ^2: 100

Prior Distribution:

$$\beta_0 \sim \mathrm{N}(0.000, 100.000)$$

β_1 Sp.Lim

β_1 distribution: Normal

Mean μ: 0

Variance σ^2: 100

Prior Distribution:

$$\beta_1 \sim \mathrm{N}(0.000, 100.000)$$

β_2 Av.Sp

β_2 distribution: Normal

Mean μ: 0

Variance σ^2: 100

Prior Distribution:

$$\beta_2 \sim \mathrm{N}(0.000, 100.000)$$

β_3 ef.p

β_3 distribution: Normal

Mean μ: 0

Variance σ^2: 100

Prior Distribution:

$$\beta_3 \sim \mathrm{N}(0.000, 100.000)$$

β_4 ov.lim

β_4 distribution: Normal

Mean μ: 0

Variance σ^2: 100

Prior Distribution:

$$\beta_4 \sim \mathrm{N}(0.000, 100.000)$$

β_5 fifteen.ov.lim

β_5 distribution: Normal

Mean μ: 0

Variance σ^2: 100

Prior Distribution:

$$\beta_5 \sim \mathrm{N}(0.000, 100.000)$$

β_6 Flow

β_6 distribution: Normal

Mean μ: 0

Variance σ^2: 100

Prior Distribution:

$$\beta_6 \sim \mathrm{N}(0.000, 100.000)$$

β_7 Class

β_7 distribution: Normal

Mean μ: 0

Variance σ^2: 100

Prior Distribution:

$$\beta_7 \sim \mathrm{N}(0.000, 100.000)$$

β_8 Type

β_8 distribution: Normal

Mean μ: 0

Variance σ^2: 100

Prior Distribution:

$$\beta_8 \sim \mathrm{N}(0.000, 100.000)$$

γ

The overdispersion parameter of the model. Larger values indicate we expect to see a greater amount of variability in inputted accident counts.

γ Distribution: Uniform

Lower Bound: 0

Upper Bound: 50

Prior Distribution:

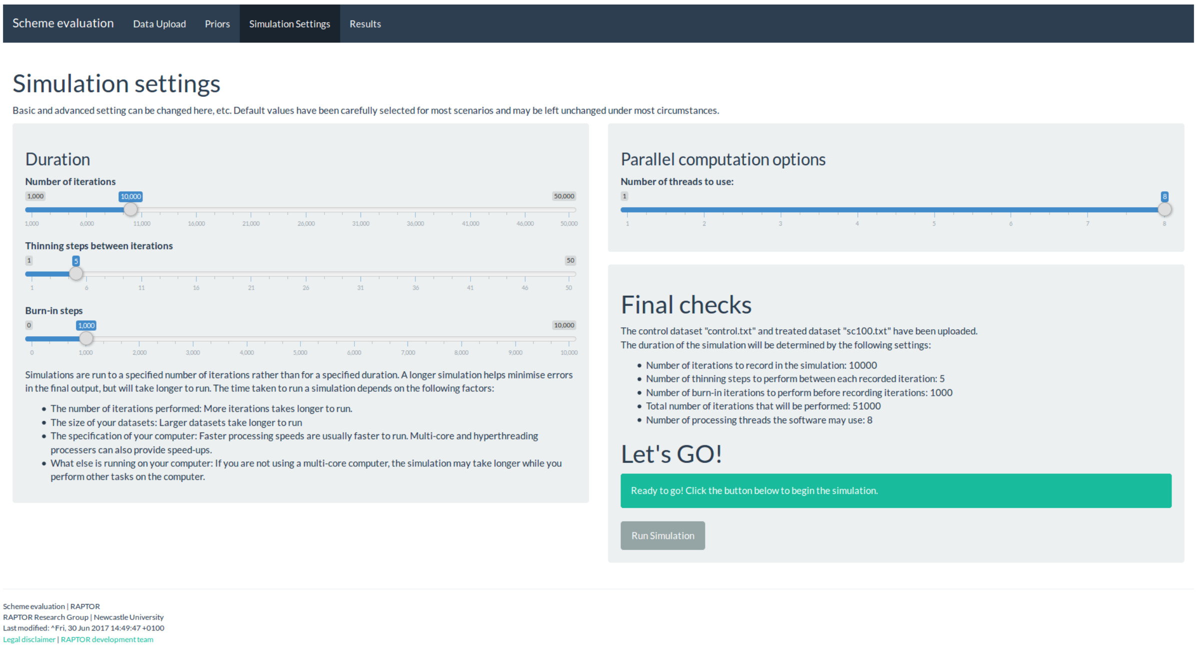
$$\gamma \sim \mathrm{Unif}(0.000, 50.000)$$

Prior distribution previews

tab-6724-2

Scheme evaluation | RAPTOR
RAPTOR Research Group | Newcastle University
Last modified: ^Fri, 30 Jun 2017 14:49:47 +0100
Legal disclaimer | RAPTOR development team

Figure A.2: The “Priors” tab of the scheme evaluation RAPTOR app.



Simulation settings

Basic and advanced setting can be changed here, etc. Default values have been carefully selected for most scenarios and may be left unchanged under most circumstances.

Duration

Number of iterations
1,000 10,000 50,000

Thinning steps between iterations
1 5 50

Burn-in steps
0 1,000 10,000

Simulations are run to a specified number of iterations rather than for a specified duration. A longer simulation helps minimise errors in the final output, but will take longer to run. The time taken to run a simulation depends on the following factors:

- The number of iterations performed: More iterations takes longer to run.
- The size of your datasets: Larger datasets take longer to run
- The specification of your computer: Faster processing speeds are usually faster to run. Multi-core and hyperthreading processers can also provide speed-ups.
- What else is running on your computer: If you are not using a multi-core computer, the simulation may take longer while you perform other tasks on the computer.

Parallel computation options

Number of threads to use:
1 8

Final checks

The control dataset "control.txt" and treated dataset "sc100.txt" have been uploaded.
The duration of the simulation will be determined by the following settings:

- Number of iterations to record in the simulation: 10000
- Number of thinning steps to perform between each recorded iteration: 5
- Number of burn-in iterations to perform before recording iterations: 1000
- Total number of iterations that will be performed: 51000
- Number of processing threads the software may use: 8

Let's GO!

Ready to go! Click the button below to begin the simulation.

Scheme evaluation | RAPTOR
RAPTOR Research Group | Newcastle University
Last modified: ^Fri, 30 Jun 2017 14:49:47 +0100
Legal disclaimer | RAPTOR development team

Figure A.3: The “Simulation Settings” tab of the scheme evaluation RAPTOR app.

tion 2.2. While this approach to trend has been demonstrated in the literature (see for instance [Fawcett and Thorpe, 2013]), there are limitations and so it is hoped an improved approach, such as the one demonstrated in Chapter 4 shall be incorporated soon, removing an requirement for expert prior information in order to model trend. Furthermore it is hoped to improve on the current method for imparting prior knowledge regarding the regression coefficients β_j, as the current method of specifying prior parameters may still be inaccessible to practitioners without a statistical background, see section 7.2.5 for details on future methods.

Simulation Settings

The “Simulation Settings” tab allows the user to specify how long they would like the Markov Chain Monte Carlo chain to run, in terms of number of iterations and degree of thinning. This has been included in order to allow the application to be flexible to the users needs, whereby if the user simply needs some quick results, or wishes to demonstrate the application and the potential results, the number of iterations can be lowered in order to speed up the analysis. The default settings of 50,000 iterations and a thinning rate of 5 have been tested on multiple datasets and shown to produce good convergence and consistent results, and so most users are encouraged not to change these for any serious analysis, and can increase them should the need for results not be urgent. The analysis of each treated site (i.e. the application of the obtained SPF) can be carried out independently, and so in order to improve computation time this process will run in parallel. The user has the option to adjust the number of threads used in the computation, however since the application runs on an external server, there is no computational requirement/cost to the user to running the analysis at the maximum 8 threads, and so there is no likely reason for most users to need to adjust this.

Results

Again as we would expect given the name, the "Results" tab provides the user with the output of the scheme evaluation analysis. This is provided in a variety of forms, firstly the "Summary" sub tab provides a breakdown of the change in collision totals between the before and after periods across the treated sites into change due to RTM, trend and treatment effect. This is provided firstly in the form of a table, which has rows corresponding to each treated site included in the analysis, and columns containing:

- site ID (i)
- collision total in the period before treatment ($y_{i,\text{BEF}}$)
- collision total in the period after treatment ($y_{i,\text{AFT}}$)
- total observed change in collisions from before to after the treatment ($y_{i,\text{BEF}} - y_{i,\text{AFT}}$)
- estimated change due to the RTM effect ($\hat{\rho}_i$)
- estimated change due to the trend effect ($\hat{\kappa}_i$)
- estimated change due to the treatment effect ($\hat{\tau}_i$)
- posterior mean of the collision rate ($E\left(\lambda_i | y_i\right)$)
- the estimated collision rate from the SPF (μ_i)

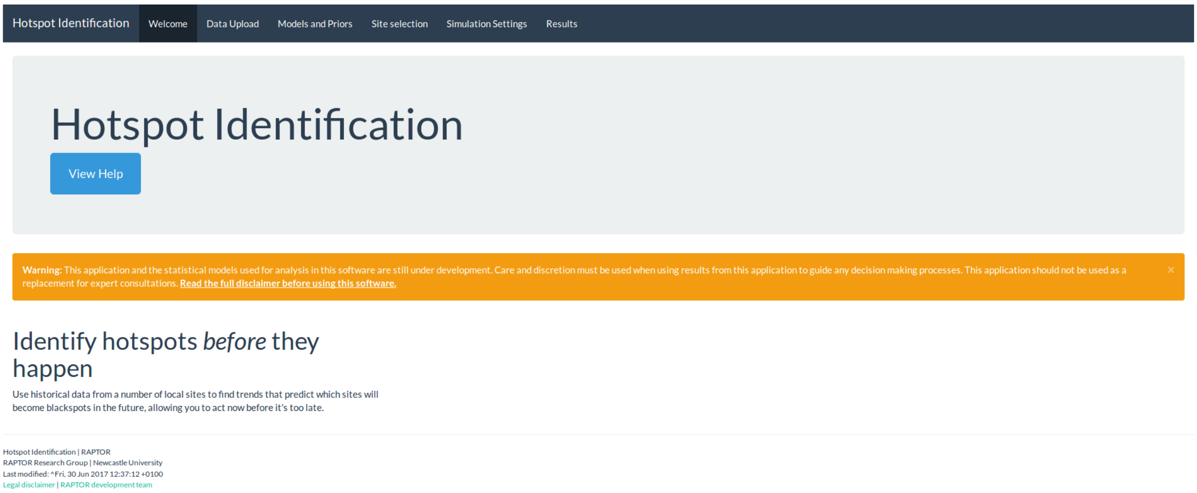
Clearly there is a lot of information in this table, and not all of it will be of interest to all practitioners, and so to better convey the main results, bar plots are provided below the table displaying the observed change from before to after treatment, as well as the estimates of the RTM, trend and treatment effects, both for the entirety of the treated sites, as well as on a site by site basis. The reasoning behind this is to allow practitioners to view the effectiveness of the countermeasure as whole, as well as see clearly where it has been most (and least) beneficial so as to best inform future decision making regarding countermeasure deployment/retention. All tables and plots can be downloaded and exported in a variety of formats for easy use in reports/presentations at the request of practitioners involved in the primary trials of the RAPTOR software. There are additional sub tabs giving tables of summary statistics of the posterior distribution for the underlying collision rate before treatment, $\lambda_{i,\mathrm{BEF}}$ and the estimated number of collisions from the SPF, μ_i, namely the mean, median and standard deviation of the posterior distributions, and a 95% credible interval, on a site-by-site basis. Clearly it is important to include these additional statistics in order for the analysis to be as honest as possible regarding the certainty of the results obtained in the scheme evaluation analysis. The final sub tab gives the same summaries of the posterior distributions of the regression coefficient vector $\boldsymbol{\beta} = \left(\beta_0, \ldots, \beta_{n_p}\right)^T$, which in addition to giving additional information regarding the level of certainty in the model (in this case the certainty in the SPF), it also provides information regarding the effect each covariate appears to have on the level of risk across the comparison site network, and indeed if such an effect really exists (which can be informally estimated easily by comparing the posterior standard deviation with the posterior mean, or conversely by looking at the 95% credible interval). Again the tables of results in all sub tabs can be directly exported in a variety of file formats for inclusion in documents etc.

A.1.3 Hotspot Prediction

The hotspot prediction application carries out the hotspot prediction analysis described in Chapter 5. Because this technique is largely novel and more statistically complex (as opposed to the scheme evaluation application which makes use of the already established and statistically straightforward FB technique), making use of a software application to implement the method is more important than ever. The overall structure of the app remains the same, with the stages of the analysis broken into separate tabs describing the data upload procedure, inclusion of prior information, the settings for the MCMC scheme, and the displaying of results.

Welcome Tab

Appropriately, the first screen displayed upon opening the hotspot prediction application is the “Welcome” tab, which provides a description of the purpose of the app, briefly describing the advantages of a hotspot prediction approach over hotspot identification (as discussed in Section 5.1), alongside a disclaimer for the usage of the application, and a link to a user guide written to allow users to navigate the app independently.



Hotspot Identification

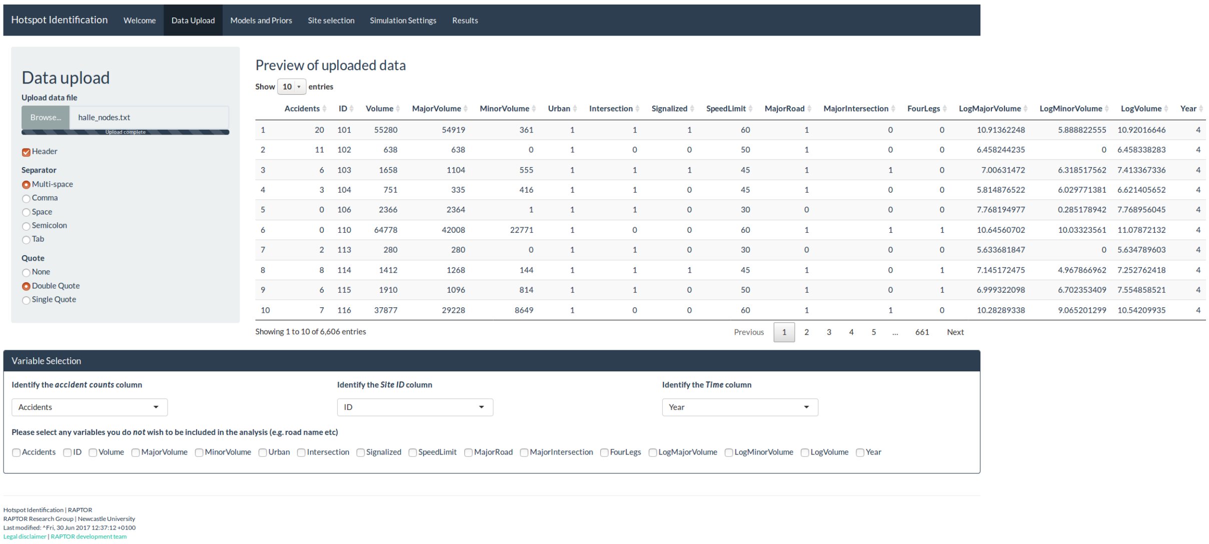
Warning: This application and the statistical models used for analysis in this software are still under development. Care and discretion must be used when using results from this application to guide any decision making processes. This application should not be used as a replacement for expert consultations. **Read the full disclaimer before using this software.**

Identify hotspots *before* they happen

Use historical data from a number of local sites to find trends that predict which sites will become blackspots in the future, allowing you to act now before it's too late.

Hotspot Identification | RAPTOR
RAPTOR Research Group | Newcastle University
Last modified: ^Fri, 30 Jun 2017 12:37:12 +0100
Legal disclaimer | RAPTOR development team

Figure A.4: The "Welcome" tab of the hotspot prediction RAPTOR app.



Data upload

Upload data file: halle_nodes.txt

☑ Header

Separator
● Multi-space
○ Comma
○ Space
○ Semicolon
○ Tab

Quote
○ None
● Double Quote
○ Single Quote

Preview of uploaded data

Show 10 entries

	Accidents	ID	Volume	MajorVolume	MinorVolume	Urban	Intersection	Signalized	SpeedLimit	MajorRoad	MajorIntersection	FourLegs	LogMajorVolume	LogMinorVolume	LogVolume	Year
1	20	101	55280	54919	361	1	1	1	60	1	0	0	10.91362248	5.888822555	10.92016646	4
2	11	102	638	638	0	1	0	0	50	1	0	0	6.458244235	0	6.458338283	4
3	6	103	1658	1104	555	1	1	1	45	1	1	0	7.00631472	6.318517562	7.413367336	4
4	3	104	751	335	416	1	1	0	45	1	0	0	5.814876522	6.029771381	6.621405652	4
5	0	106	2366	2364	1	1	1	0	30	0	0	0	7.768194977	0.285178942	7.768956045	4
6	0	110	64778	42008	22771	1	0	0	60	1	1	1	10.64560702	10.03323561	11.07872132	4
7	2	113	280	280	0	1	1	0	30	0	0	0	5.633681847	0	5.634789603	4
8	8	114	1412	1268	144	1	1	1	45	1	0	1	7.145172475	4.967866962	7.252762418	4
9	6	115	1910	1096	814	1	1	0	50	1	0	1	6.999322098	6.702353409	7.554858521	4
10	7	116	37877	29228	8649	1	0	0	60	1	1	0	10.28289338	9.065201299	10.54209935	4

Showing 1 to 10 of 6,606 entries

Variable Selection

Identify the *accident counts* column: Accidents

Identify the *Site ID* column: ID

Identify the *Time* column: Year

Please select any variables you do *not* wish to be included in the analysis (e.g. road name etc)

□ Accidents □ ID □ Volume □ MajorVolume □ MinorVolume □ Urban □ Intersection □ Signalized □ SpeedLimit □ MajorRoad □ MajorIntersection □ FourLegs □ LogMajorVolume □ LogMinorVolume □ LogVolume □ Year

Hotspot Identification | RAPTOR
RAPTOR Research Group | Newcastle University
Last modified: ^Fri, 30 Jun 2017 12:37:12 +0100
Legal disclaimer | RAPTOR development team

Figure A.5: The "Data Upload" tab of the hotspot prediction RAPTOR app.

Data Upload

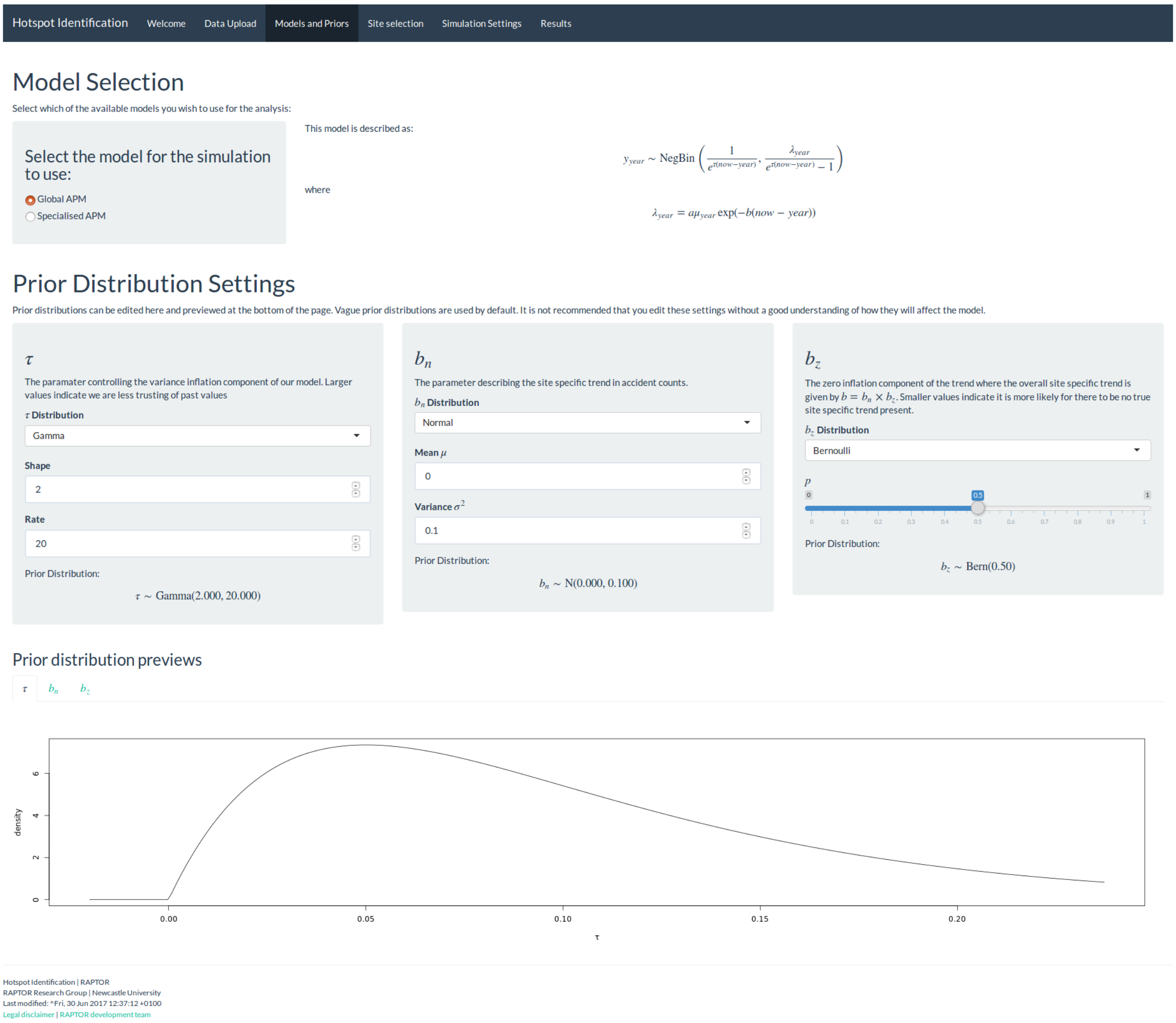
The "Data Upload" tab largely resembles the corresponding tab from the scheme evaluation app. Perhaps the first key difference is the single field for uploading data, since for the hotspot prediction analysis there is no comparison data, and so only one dataset is required. The mechanism for uploading data is exactly the same as for scheme evaluation with regards to data types etc. The sole difference between the dataset used for hotspot prediction and for scheme evaluation is that in the case of hotspot prediction, clearly there are multiple observations on each site, and each row should correspond to a given site in a given time period (so there are nn_{years} rows, as opposed to n rows in the case of scheme evaluation). Additionally therefore there should be an additional column denoting the time period which the observation corresponds to (meaning there should be $n_p + 3$ columns, corresponding to the n_p covariates used in the SPF, site ID, collision count, and time indicator). As with scheme evaluation, there is a set of select box menus for the user to specify which columns correspond to the collision counts, site IDs and time indicators (although the application will again attempt to set these from the dataset column headings). Furthermore there is the option to deselect covariates from the analysis should they not be suitable (e.g. the name of the road the site is situated on), and state whether certain numeric variables should be treated as categorical or continuous (e.g. speed limit).

Priors

The "Priors" tab again allows the user to specify any expert prior information they have by adjusting the prior distributions for some of the parameters in the model. These parameters are:

- The probability parameter in the *Bernoulli* distribution for the zero inflation component b_Z of the local trend parameter b. This parameter effectively controls the prior probability that there is no site-specific deviation from the global trend fitted in the SPF. It is expected therefore that if the user believes their network to be fairly homogenous, and thus have largely similar temporal trends, this parameter will be set closer to 1, and vice-versa should they believe their network will be diverse in trends. The default is 0.5, representing no prior knowledge.

- The mean and variance of the $Normal$ component b_N of the local trend parameter b. This again can be adjusted to reflect how wildly the user believes the trend is likely to vary across the network. Since the parameter b corresponds to site-specific deviations from the global trend, which itself can be thought of informally as the "average" trend across the network, it is unlikely for the mean of these deviations to be far from zero (although this could be the case should there be anomalous sites



Model Selection

Select which of the available models you wish to use for the analysis:

Select the model for the simulation to use:

● Global APM
○ Specialised APM

This model is described as:

$$y_{year} \sim \mathrm{NegBin}\left(\frac{1}{e^{\tau(now-year)}}, \frac{\lambda_{year}}{e^{\tau(now-year)}-1}\right)$$

where

$$\lambda_{year} = a\mu_{year}\exp(-b(now-year))$$

Prior Distribution Settings

Prior distributions can be edited here and previewed at the bottom of the page. Vague prior distributions are used by default. It is not recommended that you edit these settings without a good understanding of how they will affect the model.

τ

The paramater controlling the variance inflation component of our model. Larger values indicate we are less trusting of past values

τ Distribution: Gamma

Shape: 2

Rate: 20

Prior Distribution:

$$\tau \sim \mathrm{Gamma}(2.000, 20.000)$$

b_n

The parameter describing the site specific trend in accident counts.

b_n Distribution: Normal

Mean μ: 0

Variance σ^2: 0.1

Prior Distribution:

$$b_n \sim \mathrm{N}(0.000, 0.100)$$

b_z

The zero inflation component of the trend where the overall site specific trend is given by $b = b_n \times b_z$. Smaller values indicate it is more likely for there to be no true site specific trend present.

b_z Distribution: Bernoulli

p: 0.5

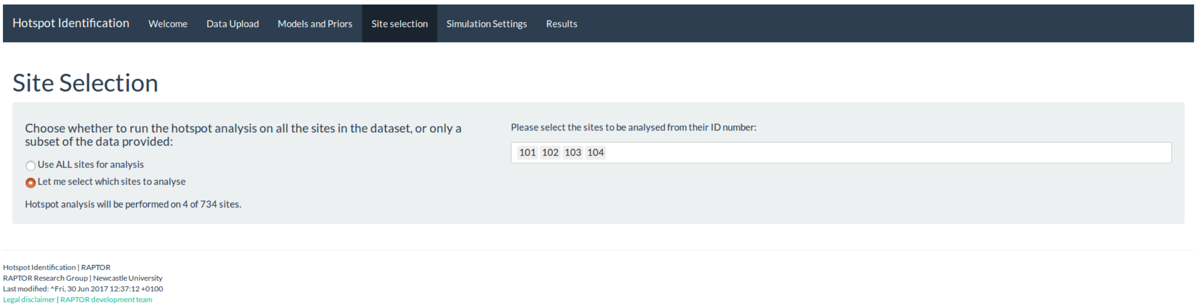
Prior Distribution:

$$b_z \sim \mathrm{Bern}(0.50)$$

Prior distribution previews

Hotspot Identification | RAPTOR
RAPTOR Research Group | Newcastle University
Last modified: ^Fri, 30 Jun 2017 12:37:12 +0100
Legal disclaimer | RAPTOR development team

Figure A.6: The “Priors” tab of the hotspot prediction RAPTOR app.



Site Selection

Choose whether to run the hotspot analysis on all the sites in the dataset, or only a subset of the data provided:

○ Use ALL sites for analysis

● Let me select which sites to analyse

Hotspot analysis will be performed on 4 of 734 sites.

Please select the sites to be analysed from their ID number:

101 102 103 104

Hotspot Identification | RAPTOR
RAPTOR Research Group | Newcastle University
Last modified: ^Fri, 30 Jun 2017 12:37:12 +0100
Legal disclaimer | RAPTOR development team

Figure A.7: The "Site Selection" tab of the hotspot prediction RAPTOR app.

which have significantly different trends which distort the global trend parameter). The variance of b_N however can be adjusted to represent how wildly the user expects the trends to vary across the network. While one would expect a decrease in p to correspond to an increase in the variance of b_Z, since both would indicate a heterogeneous trend across the network, this may not be the case since for example it could be believed that trends will deviate from the global trend often, but not by a large amount and vice-versa, and so it is worthwhile allowing both parameters to be specified. The defaults for these again reflect no prior knowledge, and so a mean of 0 and a variance of 10 are selected.

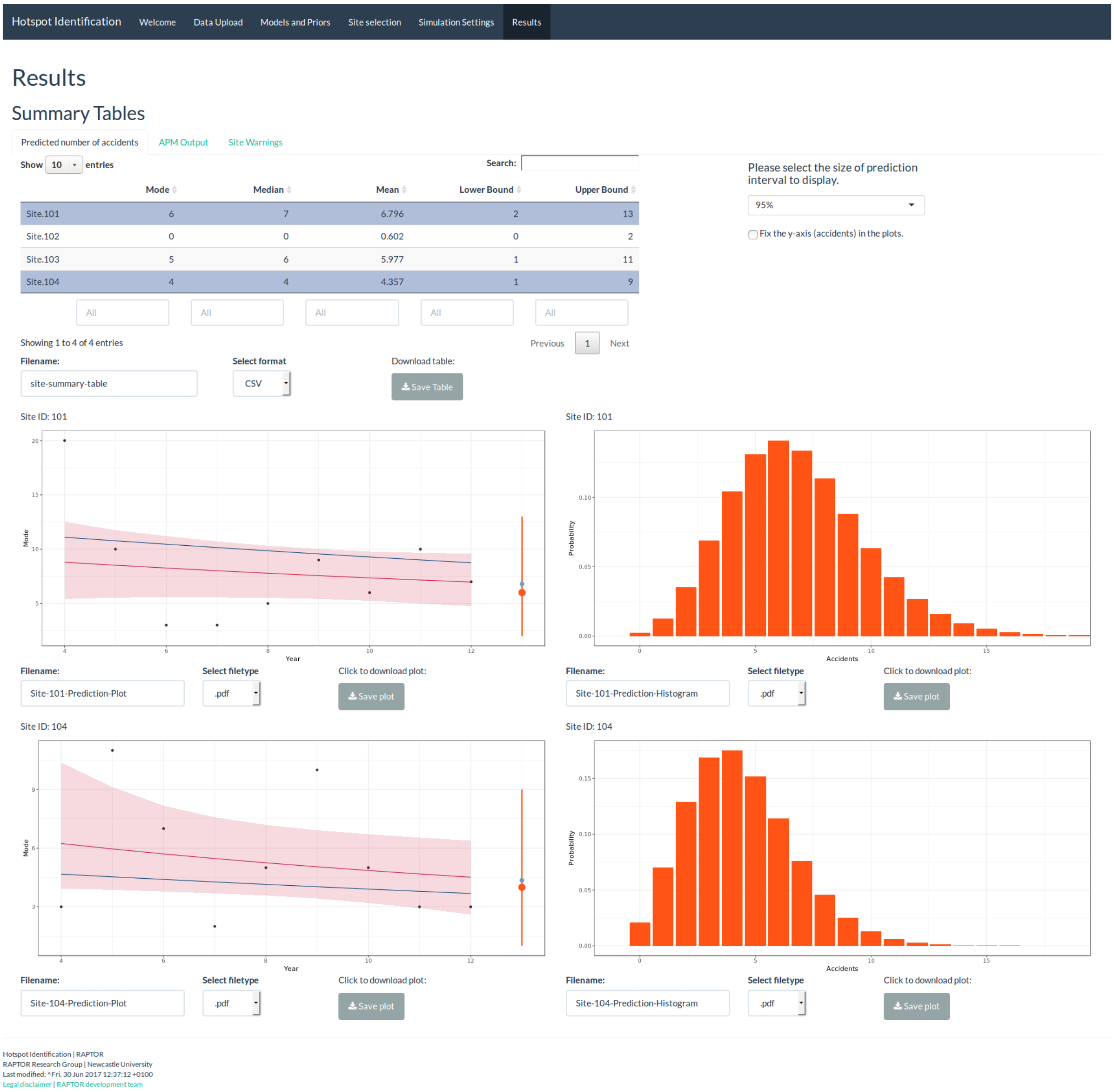
The defaults for these prior distributions reflect those used in standard analyses as described in Chapter 5, meaning that even if the expert did not wish to include any prior information, the analysis should still run well. Due to the increased likelihood of datasets used for hotspot prediction being significantly larger than those used for scheme evaluation (the number of potential hotspots should always comfortably outnumber the chosen hotspots) we choose to fit the SPF using maximum likelihood estimation, meaning the time taken to obtain the SPF is negligible, and allowing for individual sites to be analysed quickly rather than the entire network. Because of this non-Bayesian SPF, there is no need to specify prior distributions for the regression coefficient parameters β_j, as was the case for scheme evaluation.

Site Selection

As discussed in the previous section, there is significant potential for datasets used in hotspot prediction analyses to be very large (for instance the Halle dataset discussed in Chapter 5 has 734 × 9 = 6606 individual data points, and is just analysing one city, whereas it is not uncommon for datasets of entire states or even countries to require analysis). It is important however to ensure that the model runtime remains feasible and appealing to potential users, which could quickly not be the case should the datasets become too large. To help avoid this issue, as discussed previously, the SPF is fitted using maximum likelihood estimation, meaning the time taken to compute the estimates of the regression parameters is negligible, significantly reducing computation time. This has the added benefit of taking advantage of the fact that all model parameters at a site are conditionally independent of the parameters at all other sites, given the regression coefficients. Because of this, once the regression coefficients have been specified, a hotspot prediction analysis can be run on each site separately, without the need for all sites to be analysed at once. This therefore provides the huge advantage for quick analyses to be run on a small number of target sites much quicker, whilst still retaining information from across the network for use in the SPF. The ability for the user to select which sites they wish to be analysed is included in the "Site Selection" tab, where there is a list of 2 radio buttons, one option to select to analyse all sites (accompanied by a warning as to the potential runtime) and another to choose a subset of sites, which generates a box from which the user can select sites to analyse by their site IDs.

Results

As with the scheme evaluation app there are a variety of sub tabs within the "Results" tab, each offering a different aspect to the output of the analysis. The first sub tab, named "Predicted number of accidents" gives the overall results of the hotspot prediction analysis in both numerical and graphical form. A table of summary statistics describing the posterior predictive distribution for the number of collisions in the next time period for each site included in the analysis. The summary statistics displayed are the mode, median and mean of the posterior predictive distribution, along with a credible interval. The size of the credible interval is determined by the select input menu to the right of the summary table which contains the option to specify a 50%, 90%, 95% or 99% credible interval. Making use of a small pre-selected list of interval sizes (as opposed to allowing the user to specify and width they wish) saves a significant amount of computation time, and the interval sizes selected are likely the only ones users would consider of interest (and are heavily the most prevalent in research/practice). Clicking on a row produces a pair of plots, the first being the same as those shown in Figure 5.2, showing the posterior



Results

Summary Tables

Show 10 entries

Search:

	Mode	Median	Mean	Lower Bound	Upper Bound
Site.101	6	7	6.796	2	13
Site.102	0	0	0.602	0	2
Site.103	5	6	5.977	1	11
Site.104	4	4	4.357	1	9
	All	All	All	All	All

Showing 1 to 4 of 4 entries

Filename: site-summary-table

Select format: CSV

Download table:

Please select the size of prediction interval to display.

95%

□ Fix the y-axis (accidents) in the plots.

Site ID: 101

Filename: Site-101-Prediction-Plot

Select filetype: .pdf

Click to download plot:

Site ID: 101

Filename: Site-101-Prediction-Histogram

Select filetype: .pdf

Click to download plot:

Site ID: 104

Filename: Site-104-Prediction-Plot

Select filetype: .pdf

Click to download plot:

Site ID: 104

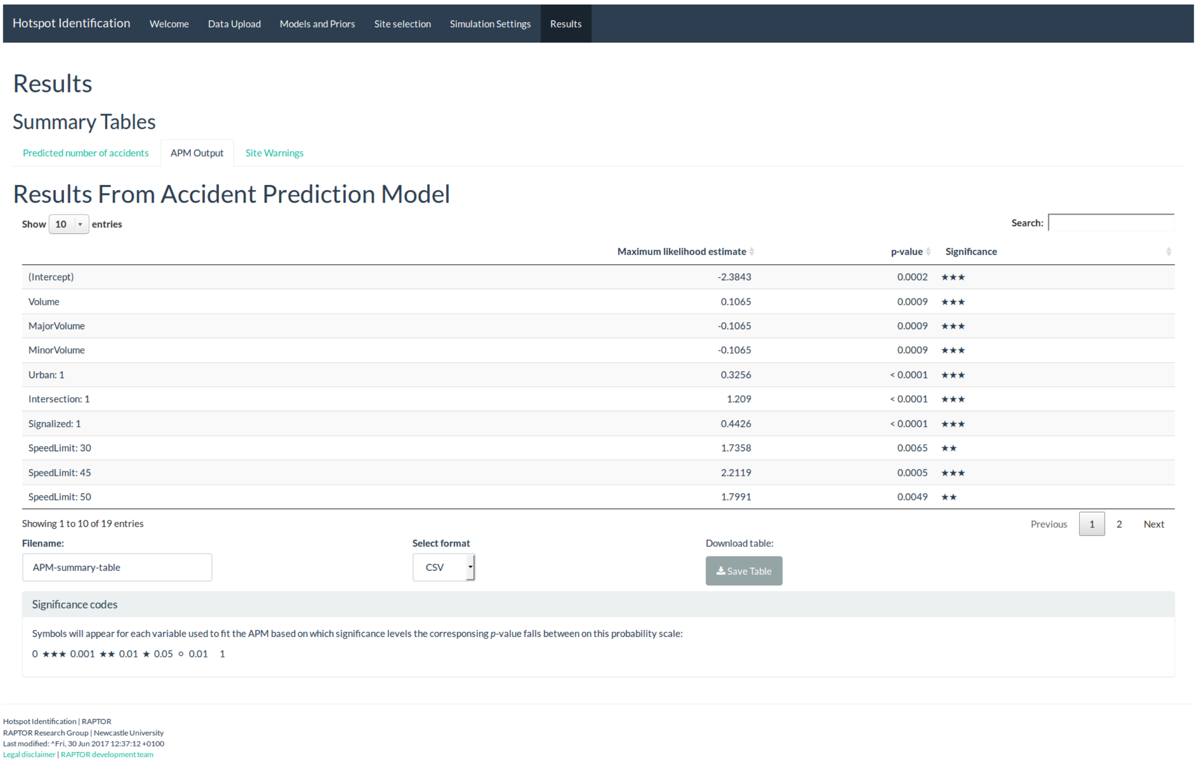
Filename: Site-104-Prediction-Histogram

Select filetype: .pdf

Click to download plot:

Hotspot Identification | RAPTOR
RAPTOR Research Group | Newcastle University
Last modified: ^Fri, 30 Jun 2017 12:37:12 +0100
Legal disclaimer | RAPTOR development team

Figure A.8: The "Predicted number of accidents" sub tab of the "Results" tab of the hotspot prediction RAPTOR app.



Results

Summary Tables

Results From Accident Prediction Model

Show 10 entries

Search:

	Maximum likelihood estimate	p-value	Significance
(Intercept)	-2.3843	0.0002	★★★
Volume	0.1065	0.0009	★★★
MajorVolume	-0.1065	0.0009	★★★
MinorVolume	-0.1065	0.0009	★★★
Urban: 1	0.3256	< 0.0001	★★★
Intersection: 1	1.209	< 0.0001	★★★
Signalized: 1	0.4426	< 0.0001	★★★
SpeedLimit: 30	1.7358	0.0065	★★
SpeedLimit: 45	2.2119	0.0005	★★★
SpeedLimit: 50	1.7991	0.0049	★★

Showing 1 to 10 of 19 entries

Filename: APM-summary-table

Select format: CSV

Download table:

Significance codes

Symbols will appear for each variable used to fit the APM based on which significance levels the corresponsing *p*-value falls between on this probability scale:

0 ★★★ 0.001 ★★ 0.01 ★ 0.05 ∘ 0.01 1

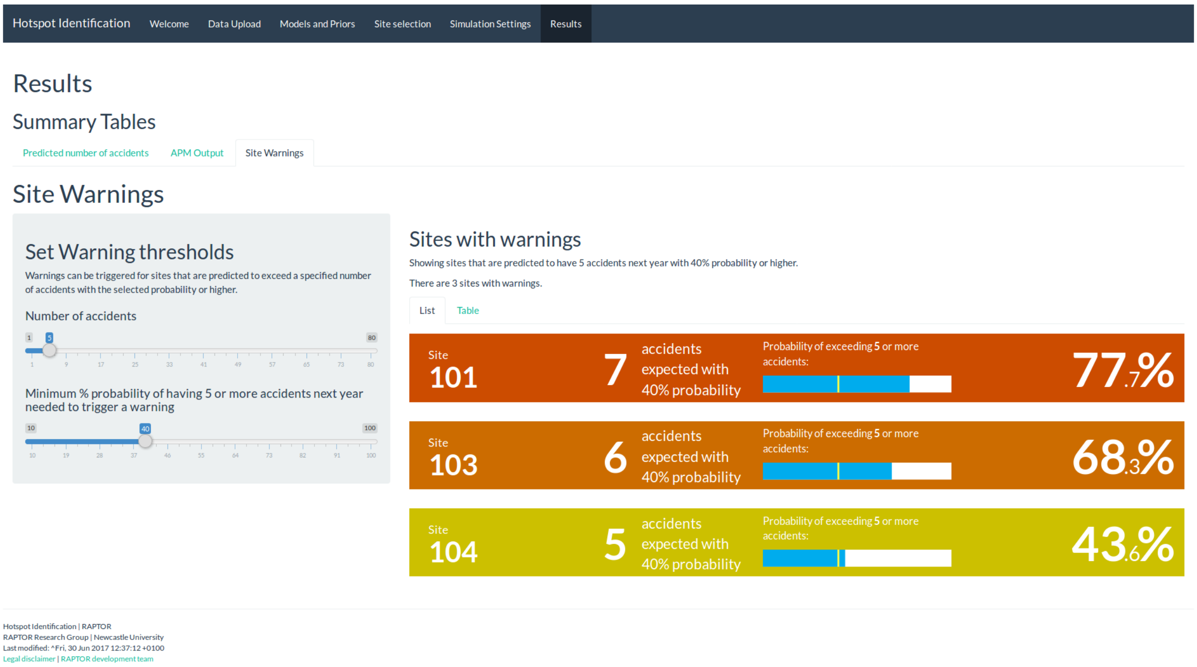
Hotspot Identification | RAPTOR
RAPTOR Research Group | Newcastle University
Last modified: ^Fri, 30 Jun 2017 12:37:12 +0100
Legal disclaimer | RAPTOR development team

Figure A.9: The "APM Output" sub tab of the "Results" tab of the hotspot prediction RAPTOR app.

means for λ and μ over time at each site, along with the observed collision counts and posterior predictive mean and mode (specified by the orange and blue dots respectively), along with the specified predictive interval (given by the orange line). Alongside this plot there is the full posterior predictive distribution for the number of collisions in the future time period, obtained via the same mechanism as specified in Chapter 5. As with the scheme evaluation app, the summary table and all plots can be downloaded and exported in a variety of formats using the buttons below each.

The second sub tab in the results section is "APM Output" which, as for the scheme evaluation app, gives summaries of the covariate coefficients used to formulate the SPF. As mentioned previously, the SPF is fitted using maximum likelihood estimation, and so rather than summary information of the posterior distributions, we instead have a table giving the maximum likelihood estimate and corresponding p-value (and significance) for each covariate. The significance column assigns a star rating to each covariate's p-value to allow for easy interpretation by practitioners:

- $0 < p \leq 0.001$ - 3 stars, extremely significant
- $0.001 < p \leq 0.01$ - 2 stars, highly significant



Results

Summary Tables

Site Warnings

Set Warning thresholds

Warnings can be triggered for sites that are predicted to exceed a specified number of accidents with the selected probability or higher.

Number of accidents: 5

Minimum % probability of having 5 or more accidents next year needed to trigger a warning: 40

Sites with warnings

Showing sites that are predicted to have 5 accidents next year with 40% probability or higher.

There are 3 sites with warnings.

Site 101 — 7 accidents expected with 40% probability — Probability of exceeding **5** or more accidents: 77.7%

Site 103 — 6 accidents expected with 40% probability — Probability of exceeding **5** or more accidents: 68.3%

Site 104 — 5 accidents expected with 40% probability — Probability of exceeding **5** or more accidents: 43.6%

Hotspot Identification | RAPTOR
RAPTOR Research Group | Newcastle University
Last modified: ^Fri, 30 Jun 2017 12:37:12 +0100
Legal disclaimer | RAPTOR development team

Figure A.10: The “Site Warnings” sub tab of the “Results” tab of the hotspot prediction RAPTOR app.

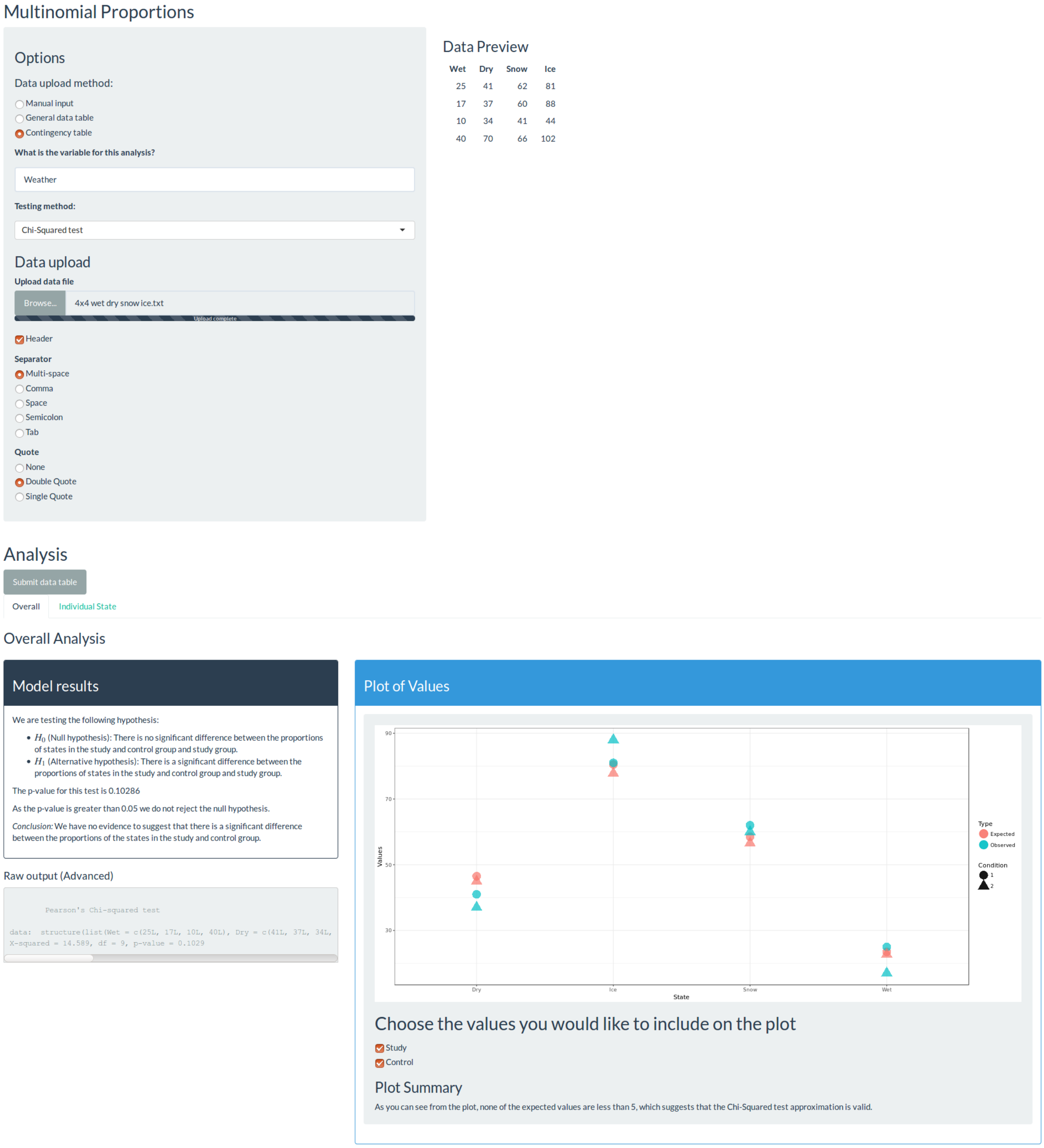
- $0.01 < p \leq 0.05$ - 1 star, moderately significant
- $0.05 < p \leq 0.01$ - 1/2 star (denoted by a white circle in the app), potentially significant
- $0.01 < p \leq 1$ - 0 stars, not significant

A guide on interpreting the output from the SPF, and how to act accordingly, e.g. if a lot of the covariates are deemed to be non-significant, is given in the app’s user guide. Again this table can be exported and downloaded using the buttons below the table.

The final sub tab is the “Site Warnings” sub tab which interprets the output of the hotspot prediction analysis in a way that will be highly useful for most practitioners. As discussed in Chapters 1 and 5, usually hotspots are determined as locations which exceed given safety threshold criteria, usually based around the number of collisions (perhaps of a given severity). We can therefore appeal to the predictive posterior distribution to estimate the probability of exceeding the threshold in the future time period as the posterior predictive density beyond this value. In the site warnings tab the user can specify the threshold number of collisions beyond which they would consider a site a hotspot using the top slider bar in the left hand panel, and the predictive probability of each site exceeding this threshold is given. In the case that a lot of sites have been analysed, there is the option for the user to specify a minimum probability, below which they would not be interested in viewing sites (for instance a practitioner is unlikely to care which sites have only a 3% chance of exceeding the threshold) so as to allow the user to easily identify which sites are likely to be a problem. To further aid this process, the list of exceedance probabilities are colour coded from red indicating a probability close to 100%, to green denoting probabilities closer to 0%.

A.1.4 Causation Factor Analysis

A potential limitation of the scheme evaluation and hotspot prediction analyses discussed thus far, is that whilst they are effective at analysing overall collision rates at locations, they do little to explain why the collisions occur, which is clearly an important consideration for practitioners as deciding on the most important countermeasure to deploy is as important as determining where to deploy it. One approach could be for the practitioner to carry out a standard hotspot prediction analysis using the hotspot prediction application, but only include collisions of a specific type. For instance if the practitioner wished to know where to deploy a set of speed cameras, they could carry out a hotspot prediction analysis, but only include collisions which were related to speeding, thereby obtaining the predicted number of speeding collisions at sites across the network in the next time period. While this approach is valid in principle, it can be cumbersome and inefficient if the practitioner wishes to investigate multiple possible causation factors, which may have multiple levels, at multiple sites. In such situations, it may be sufficient for the practitioner to investigate where (if anywhere) there is an unusual pattern in the number of collisions relating to possible causation factors. The investigation into proportions of collision totals due to differing types and levels of causation factors is well documented in literature (see for example [Das et al., 2015], [Shrestha and Shrestha, 2017], [Hao et al., 2016]), and it is the purpose of this application to make it easy for practitioners to carry out their own analyses. This is particularly pertinent as it relates to geographical variables such as weather condition or time of day, as opposed to the fixed factors we have previously considered when developing SPFs. Such an analysis can be carried out using the Multinomial Proportions app, shown in Figure A.11, which using various statistical hypothesis tests to investigate the possibility of imbalances between collision totals due to different causation factors between groups of sites. Data is provided in the form of an $m \times n$ contingency table, where m is the number of groups being compared, and n is the number of levels, or "states" of the causation factor being investigated (e.g. if the factor was "weather conditions", states could be "wet","dry", "snow" and "ice"). In order for any analysis to occur we clearly therefore require $m > 1$ and $n > 1$, and an error message is displayed should the input data fail to achieve this (most likely indicating a data formatting error). This table can either be entered manually (using the "Manual input" radio button) us-



Multinomial Proportions

Options

Data upload method:

○ Manual input
○ General data table
● Contingency table

What is the variable for this analysis?

Weather

Testing method:

Chi-Squared test

Data upload

Upload data file

Browse... 4x4 wet dry snow ice.txt

Upload complete

☑ Header

Separator

● Multi-space
○ Comma
○ Space
○ Semicolon
○ Tab

Quote

○ None
● Double Quote
○ Single Quote

Data Preview

Wet	Dry	Snow	Ice
25	41	62	81
17	37	60	88
10	34	41	44
40	70	66	102

Analysis

Submit data table

Overall | Individual State

Overall Analysis

Model results

We are testing the following hypothesis:

- H_0 (Null hypothesis): There is no significant difference between the proportions of states in the study and control group and study group.
- H_1 (Alternative hypothesis): There is a significant difference between the proportions of states in the study and control group and study group.

The p-value for this test is 0.10286

As the p-value is greater than 0.05 we do not reject the null hypothesis.

Conclusion: We have no evidence to suggest that there is a significant difference between the proportions of the states in the study and control group.

Raw output (Advanced)

```

	Pearson's Chi-squared test

data:  structure(list(Wet = c(25L, 17L, 10L, 40L), Dry = c(41L, 37L, 34L,
X-squared = 14.589, df = 9, p-value = 0.1029
```

Plot of Values

Choose the values you would like to include on the plot

☑ Study
☑ Control

Plot Summary

As you can see from the plot, none of the expected values are less than 5, which suggests that the Chi-Squared test approximation is valid.

Figure A.11: The Multinomial Proportions RAPTOR app, using an uploaded contingency table and displaying output from the "Overall Analysis" sub tab.

ing the numeric input boxes in the table, and editing the corresponding headings, or by uploading an existing table (with the same formats as for the other applications being supported) using the “Contingency table” radio button. The method for analysing the data can then be selected from the “Testing method” select input menu, the choices being to carry out a “Chi-squared test” ([Greenwood and Nikulin, 1996]) or a “Fisher’s Exact test” ([Fisher, 1922]), with the choice largely being due to the size of the numbers in the contingency table. It is well documented that the accuracy of Pearson’s chi-squared test diminishes severely if the minimum number in a group becomes too small (with 5-10 being the recommended smallest value), and so in such a case it is recommended for users to select the option to carry out Fisher’s exact test (a warning notice appears should a dataset not be suitable for Pearson’s chi-squared test).

The “Analysis” section consists of two sub-tabs, the “Overall” sub tab providing the results for the overall dataset, carrying out the hypothesis test for all m groups in the dataset:

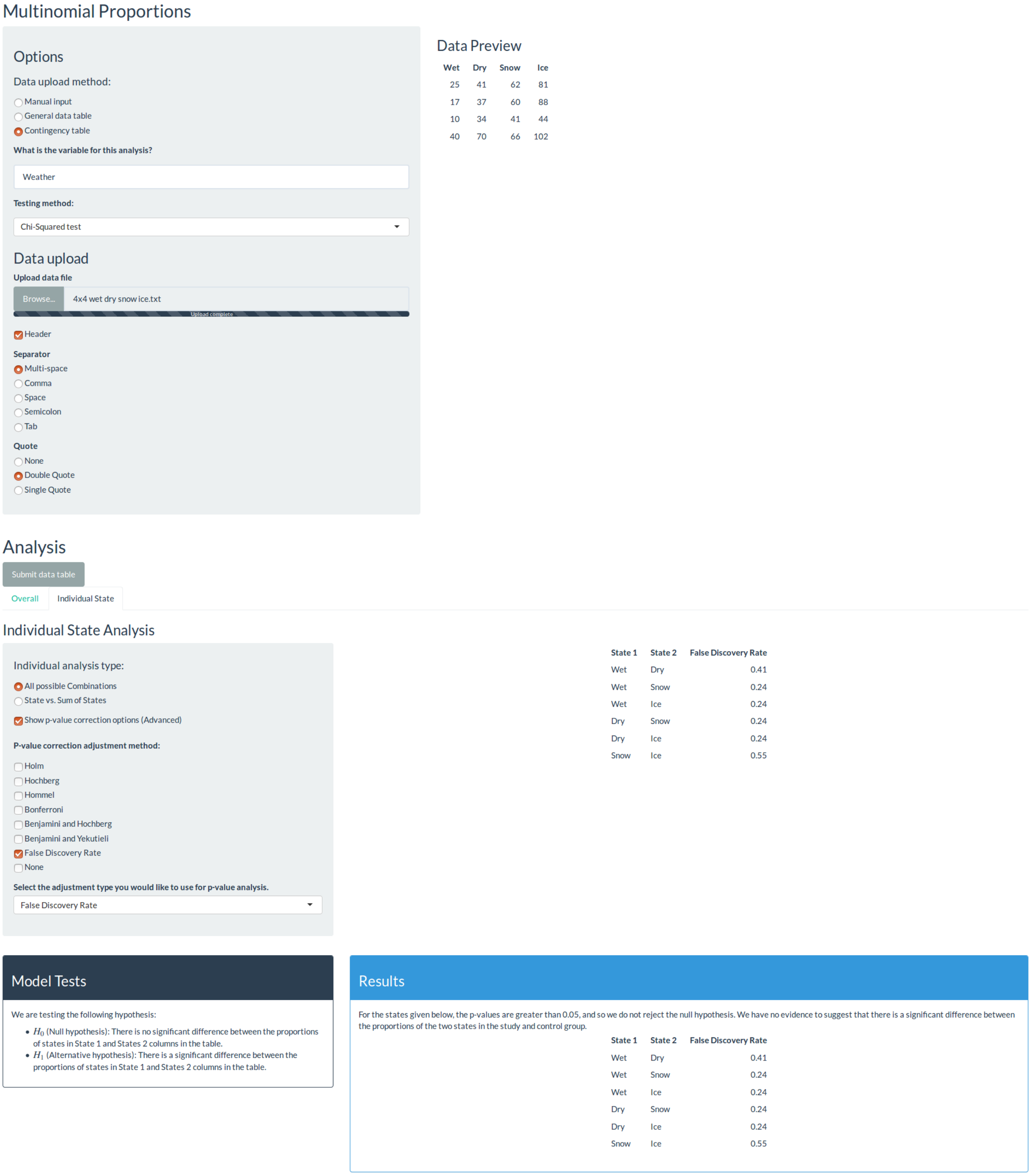
- H_0: There is no significant difference between the proportions of each state across all m groups.
- H_1: There is (at least) a significant difference between the proportions of each state across all m groups,

The hypotheses are given in the “Model Results” box along with the resulting p-value and interpretation. The explicit output from `R` is given below in the “Raw output” box, although this is specified as “Advanced” and should only be of use to users experienced with `R` or other similar statistical software. The results are presented graphically in the “Plot of Values” panel to the right, visually displaying the number of collisions expected under the null hypothesis, and those observed, for each group in each of the states. This plot is designed to help the user easily notice where any significant discrepancies are and hence which states (if any) lead to the model results provided. The “Plot Summary” statement provides information on whether the appropriate test was used, based on the expected values displayed. The “Individual State” sub tab then allows the user to compare the proportions of one state against the remaining data in order to best ascertain which states give output which is not in accordance with the null hypothesis. This can be done in two ways:

- All possible combinations: Compares proportions between every pair of states and tests the null hypothesis that there is no difference in proportion between each pair of states.
- State vs sum of states: Compares the proportion of one state against the proportions for the other states combined (i.e. for state i the collisions for each group is summed

across all $j \neq i$ states, and the resulting proportion compared with that of state i, with the null hypothesis again being that there is no difference between the proportions.

A statistical problem with these comparisons is that they involve running multiple hypothesis tests on the same dataset ($\frac{n(n-1)}{2}$ comparisons for the "all possible combinations" analysis, and n comparisons for the "state vs sum of states" analysis), which leads to increased risk of a false positive (i.e. a significant result being given purely due to chance) occurring. To counteract this, adjusted p-values are obtained, the method by which p-values can be adjusted can be selected by the user by checking the "Show p-value correction options" box, and checking the correction method they would like to use, the options available being the options available in the in-built `p.adjust` R command: Holm, Hochberg, Hommel, Bonferroni, Bonferroni and Hochberg, Benjamini and Yekutieli, False Discovery Rate, or none. More information on these corrections can be found in [Wright, 1992].



Multinomial Proportions

Options

Data upload method:

○ Manual input
○ General data table
● Contingency table

What is the variable for this analysis?

Weather

Testing method:

Chi-Squared test

Data upload

Upload data file

Browse... 4x4 wet dry snow ice.txt

Upload complete

☑ Header

Separator

● Multi-space
○ Comma
○ Space
○ Semicolon
○ Tab

Quote

○ None
● Double Quote
○ Single Quote

Data Preview

Wet	Dry	Snow	Ice
25	41	62	81
17	37	60	88
10	34	41	44
40	70	66	102

Analysis

Submit data table

Overall | Individual State

Individual State Analysis

Individual analysis type:

● All possible Combinations
○ State vs. Sum of States

☑ Show p-value correction options (Advanced)

P-value correction adjustment method:

☐ Holm
☐ Hochberg
☐ Hommel
☐ Bonferroni
☐ Benjamini and Hochberg
☐ Benjamini and Yekutieli
☑ False Discovery Rate
☐ None

Select the adjustment type you would like to use for p-value analysis.

False Discovery Rate

State 1	State 2	False Discovery Rate
Wet	Dry	0.41
Wet	Snow	0.24
Wet	Ice	0.24
Dry	Snow	0.24
Dry	Ice	0.24
Snow	Ice	0.55

Model Tests

We are testing the following hypothesis:

- H_0 (Null hypothesis): There is no significant difference between the proportions of states in State 1 and States 2 columns in the table.
- H_1 (Alternative hypothesis): There is a significant difference between the proportions of states in State 1 and States 2 columns in the table.

Results

For the states given below, the p-values are greater than 0.05, and so we do not reject the null hypothesis. We have no evidence to suggest that there is a significant difference between the proportions of the two states in the study and control group.

State 1	State 2	False Discovery Rate
Wet	Dry	0.41
Wet	Snow	0.24
Wet	Ice	0.24
Dry	Snow	0.24
Dry	Ice	0.24
Snow	Ice	0.55

Figure A.12: The Multinomial Proportions RAPTOR app, using an uploaded contingency table and displaying output from the "Individual State" sub tab.

Bibliography

[Flo,] Florida Department of Transport. https://www.fdot.gov/.

[NSR,] Northumbria Safer Roads Initiative. http://www.safespeedforlife.co.uk/.

[Abdel-Aty, 2003] Abdel-Aty, M. (2003). Analysis of driver injury severity levels at multiple locations using ordered probit models. *Journal of safety research*, 34(5):597–603.

[Agbelie, 2016] Agbelie, B. R. (2016). Random-parameters analysis of highway characteristics on crash frequency and injury severity. *Journal of traffic and transportation engineering (English edition)*, 3(3):236–242.

[Akaike, 1998] Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer.

[Al-Bdairi and Hernandez, 2017] Al-Bdairi, N. S. S. and Hernandez, S. (2017). An empirical analysis of run-off-road injury severity crashes involving large trucks. *Accident Analysis & Prevention*, 102:93–100.

[Anastasopoulos, 2016] Anastasopoulos, P. C. (2016). Random parameters multivariate tobit and zero-inflated count data models: addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. *Analytic methods in accident research*, 11:17–32.

[Bivand et al., 2013] Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, Second edition*. Springer, NY.

[Buchanan Computing,] Buchanan Computing. Accsmap. http://www.buchanancomputing.net/accsmap.html.

[Carnis and Blais, 2013] Carnis, L. and Blais, E. (2013). An assessment of the safety effects of the french speed camera program. *Accident Analysis & Prevention*, 51:301–309.

[Casella and George, 1992] Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

[Chatfield, 2016] Chatfield, C. (2016). *The analysis of time series: an introduction.* Chapman and Hall/CRC.

[Chin and Quddus, 2003] Chin, H. C. and Quddus, M. A. (2003). Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis & Prevention*, 35(2):253–259.

[Cressie, 1992] Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, 4(5):613–617.

[Das et al., 2015] Das, S., Sun, X., Wang, F., and Leboeuf, C. (2015). Estimating likelihood of future crashes for crash-prone drivers. *Journal of Traffic and Transportation Engineering (English edition)*, 2(3):145–157.

[De Pauw et al., 2014] De Pauw, E., Daniels, S., Brijs, T., Hermans, E., and Wets, G. (2014). An evaluation of the traffic safety effect of fixed speed cameras. *Safety science*, 62:168–174.

[Deka and Quddus, 2014] Deka, L. and Quddus, M. (2014). Network-level accident-mapping: distance based pattern matching using artificial neural network. *Accident Analysis & Prevention*, 65:105–113.

[Department for Transport, 2016] Department for Transport (2016). Roads investment. the roads funding package.

[Deublein et al., 2015] Deublein, M., Schubert, M., Adey, B. T., and García de Soto, B. (2015). A bayesian network model to predict accidents on swiss highways. *Infrastructure Asset Management*, 2(4):145–158.

[Diggle and Gratton, 1984] Diggle, P. J. and Gratton, R. J. (1984). Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2):193–212.

[El-Basyouny and Sayed, 2012] El-Basyouny, K. and Sayed, T. (2012). Measuring direct and indirect treatment effects using safety performance intervention functions. *Safety science*, 50(4):1125–1132.

[Farid et al., 2016] Farid, A., Abdel-Aty, M., Lee, J., Eluru, N., and Wang, J.-H. (2016). Exploring the transferability of safety performance functions. *Accident Analysis & Prevention*, 94:143–152.

[Fawcett and Thorpe, 2013] Fawcett, L. and Thorpe, N. (2013). Mobile safety cameras: Estimating casualty reductions and the demand for secondary healthcare. *Journal of Applied Statistics*, 40(11):2385–2406.

[Fawcett et al., 2017] Fawcett, L., Thorpe, N., Matthews, J., and Kremer, K. (2017). A novel bayesian hierarchical model for road safety hotspot prediction. *Accident Analysis & Prevention*, 99:262–271.

[FIA Foundation,] FIA Foundation. UN decade of action. `https://www.fiafoundation.org/our-work/road-safety-fund/un-decade-of-action/`.

[Fisher, 1922] Fisher, R. A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94.

[Forbes et al., 2011] Forbes, C., Evans, M., Hastings, N., and Peacock, B. (2011). *Statistical distributions*. John Wiley & Sons.

[Galton, 1886] Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

[García-Gallego et al., 2011] García-Gallego, A., Georgantzís, N., Navarro-Martínez, D., and Sabater-Grande, G. (2011). The stochastic component in choice and regression to the mean. *Theory and decision*, 71(2):251–267.

[Garthwaite et al., 2005] Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701.

[Gelfand and Vounatsou, 2003] Gelfand, A. E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15.

[George and McCulloch, 1993] George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

[Ghosh et al., 1994] Ghosh, M., Rao, J., et al. (1994). Small area estimation: an appraisal. *Statistical science*, 9(1):55–76.

[Gomes et al., 2017] Gomes, M. J. T. L., Cunto, F., and da Silva, A. R. (2017). Geographically weighted negative binomial regression applied to zonal level safety performance models. *Accident Analysis & Prevention*, 106:254–261.

[Greenwood and Nikulin, 1996] Greenwood, P. E. and Nikulin, M. S. (1996). *A guide to chi-squared testing*, volume 280. John Wiley & Sons.

[Greibe, 2003] Greibe, P. (2003). Accident prediction models for urban roads. *Accident Analysis & Prevention*, 35(2):273–285.

[Guo et al., 2019] Guo, X., Wu, L., Zou, Y., and Fawcett, L. (2019). Comparative analysis of empirical bayes and bayesian hierarchical models in hotspot identification. *Transportation Research Record*, page 0361198119849899.

[Haleem and Abdel-Aty, 2010] Haleem, K. and Abdel-Aty, M. (2010). Examining traffic crash injury severity at unsignalized intersections. *Journal of safety research*, 41(4):347–357.

[Hanson et al., 2013] Hanson, C. S., Noland, R. B., and Brown, C. (2013). The severity of pedestrian crashes: an analysis using google street view imagery. *Journal of transport geography*, 33:42–53.

[Hao et al., 2016] Hao, W., Kamga, C., and Wan, D. (2016). The effect of time of day on driver's injury severity at highway-rail grade crossings in the united states. *Journal of Traffic and Transportation Engineering (English edition)*, 3(1):37–50.

[Hastings, 1970] Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

[Hauer, 1980] Hauer, E. (1980). Bias-by-selection: Overestimation of the effectiveness of safety countermeasures caused by the process of selection for treatment. *Accident Analysis & Prevention*, 12(2):113–117.

[Hauer, 1986] Hauer, E. (1986). On the estimation of the expected number of accidents. *Accident Analysis & Prevention*, 18(1):1–12.

[Hauer, 1991] Hauer, E. (1991). Comparison groups in road safety studies: an analysis. *Accident Analysis & Prevention*, 23(6):609–622.

[Hewett et al., 2019] Hewett, N., Fawcett, L., Thorpe, N., Matthews, J., and Kremer, K. (2019). Sensitivity to prior specification when evaluating road safety countermeasures. *Transportation Research Record*. Submitted and under review.

[Heydari et al., 2014] Heydari, S., Miranda-Moreno, L. F., and Liping, F. (2014). Speed limit reduction in urban areas: A before–after study using bayesian generalized mixed linear models. *Accident Analysis & Prevention*, 73:252–261.

[Hossain and Muromachi, 2013] Hossain, M. and Muromachi, Y. (2013). A real-time crash prediction model for the ramp vicinities of urban expressways. *IATSS research*, 37(1):68–79.

[Hou et al., 2018] Hou, Q., Tarko, A. P., and Meng, X. (2018). Analyzing crash frequency in freeway tunnels: A correlated random parameters approach. *Accident Analysis & Prevention*, 111:94–100.

[Høye, 2015] Høye, A. (2015). Safety effects of section control-an empirical bayes evaluation. *Accident analysis & prevention*, 74:169–178.

[Ihueze and Onwurah, 2018] Ihueze, C. C. and Onwurah, U. O. (2018). Road traffic accidents prediction modelling: An analysis of anambra state, nigeria. *Accident; analysis and prevention*, 112:21–29.

[Imprialou and Quddus, 2017] Imprialou, M. and Quddus, M. (2017). Crash data quality for road safety research: current state and future directions. *Accident Analysis & Prevention*.

[Imprialou et al., 2014] Imprialou, M.-I. M., Quddus, M., and Pitfield, D. E. (2014). High accuracy crash mapping using fuzzy logic. *Transportation research part C: emerging technologies*, 42:107–120.

[Jiang et al., 2014] Jiang, X., Abdel-Aty, M., and Alamili, S. (2014). Application of poisson random effect models for highway network screening. *Accident Analysis & Prevention*, 63:74–82.

[Jones and Kenward, 2003] Jones, B. and Kenward, M. G. (2003). *Design and analysis of cross-over trials*. Chapman and Hall/CRC.

[Kitali and Sando, 2017a] Kitali, A. E. and Sando, P. T. (2017a). A full bayesian approach to appraise the safety effects of pedestrian countdown signals to drivers. *Accident Analysis & Prevention*, 106:327–335.

[Kitali and Sando, 2017b] Kitali, A. E. and Sando, P. T. (2017b). A full bayesian approach to appraise the safety effects of pedestrian countdown signals to drivers. *Accident Analysis & Prevention*, 106:327–335.

[Lan et al., 2009] Lan, B., Persaud, B., Lyon, C., and Bhim, R. (2009). Validation of a full bayes methodology for observational before–after road safety studies and application to evaluation of rural signal conversions. *Accident Analysis & Prevention*, 41(3):574–580.

[Li and Graham, 2016] Li, H. and Graham, D. J. (2016). Heterogeneous treatment effects of speed cameras on road safety. *Accident Analysis & Prevention*, 97:153–161.

[Li et al., 2017] Li, L., Gayah, V. V., and Donnell, E. T. (2017). Development of regionalized spfs for two-lane rural roads in pennsylvania. *Accident Analysis & Prevention*, 108:343–353.

[Li et al., 2013] Li, Z., Wang, W., Liu, P., Bigham, J. M., and Ragland, D. R. (2013). Using geographically weighted poisson regression for county-level crash modeling in california. *Safety science*, 58:89–97.

[Liu et al., 2017] Liu, J., Khattak, A. J., and Wali, B. (2017). Do safety performance functions used for predicting crash frequency vary across space? applying geographically weighted regressions to account for spatial heterogeneity. *Accident Analysis & Prevention*, 109:132–142.

[Mahalanobis, 1936] Mahalanobis, P. C. (1936). On the generalized distance in statistics. National Institute of Science of India.

[Matthews et al., 2019] Matthews, J., Fawcett, L., Thorpe, N., Hewett, N., and Kremer, K. (2019). The problem of, and a possible solution to, comparison site selection in scheme evaluation. *Transportation Research Record*. Submitted and under review.

[Matthews et al., 2018] Matthews, J., Newman, K., Green, A., Fawcett, L., Thorpe, N., and Kremer, K. (2018). A decision support toolkit to inform road safety investment decisions. In *Proceedings of the Institution of Civil Engineers–Municipal Engineer*, volume 172, pages 53–67. Thomas Telford Ltd.

[Miller, 1984] Miller, A. J. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)*, pages 389–425.

[Montella, 2010] Montella, A. (2010). A comparative analysis of hotspot identification methods. *Accident Analysis & Prevention*, 42(2):571–581.

[Morris et al., 2014] Morris, D. E., Oakley, J. E., and Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, 52:1–4.

[Neal, 2011] Neal, R. M. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2.

[Ntzoufras, 2011] Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*, volume 698. John Wiley & Sons.

[of Transportation,] of Transportation, U. D. FHWA Road Safety Audit Guidelines. https://safety.fhwa.dot.gov/rsa/guidelines/appendix_a.htm.

[O'Hagan and Oakley,] O'Hagan, T. and Oakley, J. E. SHELF: the Sheffield Elicitation Framework. http://www.tonyohagan.co.uk/shelf/.

[Papadimitriou et al., 2013] Papadimitriou, E., Eksler, V., Yannis, G., and Lassarre, S. (2013). Modelling the spatial variation of road safety in greece. In *Proceedings of the Institution of Civil Engineers-Transport*, volume 166. Thomas Telford Ltd.

[Park and Abdel-Aty, 2015] Park, J. and Abdel-Aty, M. (2015). Development of adjustment functions to assess combined safety effects of multiple treatments on rural two-lane roadways. *Accident Analysis & Prevention*, 75:310–319.

[Park and Abdel-Aty, 2016] Park, J. and Abdel-Aty, M. (2016). Evaluation of safety effectiveness of multiple cross sectional features on urban arterials. *Accident Analysis & Prevention*, 92:245–255.

[Park et al., 2017a] Park, J., Abdel-Aty, M., and Wang, J.-H. (2017a). Time series trends of the safety effects of pavement resurfacing. *Accident Analysis & Prevention*, 101:78–86.

[Park et al., 2017b] Park, J., Abdel-Aty, M., and Wang, J.-H. (2017b). Time series trends of the safety effects of pavement resurfacing. *Accident Analysis & Prevention*, 101:78–86.

[Park and Sahaji, 2013] Park, P. Y. and Sahaji, R. (2013). Safety diagnosis: Are we doing a good job? *Accident Analysis & Prevention*, 52:80–90.

[Persaud and Lyon, 2007] Persaud, B. and Lyon, C. (2007). Empirical bayes before–after safety studies: lessons learned from two decades of experience and future directions. *Accident Analysis & Prevention*, 39(3):546–555.

[Plummer, 2003] Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling.

[PTV Group, a] PTV Group. PTV Visum. http://vision-traffic.ptvgroup.com/en-us/products/ptv-visum/.

[PTV Group, b] PTV Group. Traffic and Logistrics Software & Technology. https://www.ptvgroup.com/en/.

[Quddus, 2008a] Quddus, M. A. (2008a). Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of london crash data. *Accident Analysis & Prevention*, 40(4):1486–1497.

[Quddus, 2008b] Quddus, M. A. (2008b). Time series count data models: an empirical application to traffic accidents. *Accident Analysis & Prevention*, 40(5):1732–1741.

[R Core Team, 2019] R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[RAC Foundation, 2016] RAC Foundation (2016). The effectiveness of average speed cameras in Great Britain.

[Roberts et al., 1997] Roberts, G. O., Gelman, A., Gilks, W. R., et al. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120.

[Sacchi et al., 2014] Sacchi, E., Sayed, T., and El-Basyouny, K. (2014). Collision modification functions: Incorporating changes over time. *Accident Analysis & Prevention*, 70:46–54.

[Schlüter et al., 1997] Schlüter, P., Deely, J., and Nicholson, A. (1997). Ranking and selecting motor vehicle accident sites by using a hierarchical bayesian model. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(3):293–316.

[Schwarz et al., 1978] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

[Shirazi et al., 2016] Shirazi, M., Lord, D., and Geedipally, S. R. (2016). Sample-size guidelines for recalibrating crash prediction models: recommendations for the highway safety manual. *Accident Analysis & Prevention*, 93:160–168.

[Shrestha and Shrestha, 2017] Shrestha, P. P. and Shrestha, K. J. (2017). Factors associated with crash severities in built-up areas along rural highways of nevada: A case study of 11 towns. *Journal of Traffic and Transportation Engineering (English edition)*, 4(1):96–102.

[Spiegelhalter et al., 2002] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

[Sun and Sun, 2015] Sun, J. and Sun, J. (2015). A dynamic bayesian network model for real-time crash prediction using traffic speed conditions data. *Transportation Research Part C: Emerging Technologies*, 54:176–186.

[Traffic Safety By Sweden,] Traffic Safety By Sweden. Vision Zero. `http://www.visionzeroinitiative.com`.

[United Nations,] United Nations. United Nations decade of action for road safety 2011-2020. `http://www.un.org/en/roadsafety`.

[U.S. Department of Transportation,] U.S. Department of Transportation. Highway safety manual. `http://www.highwaysafetymanual.org/Pages/default.aspx`.

[Wang et al., 2009] Wang, C., Quddus, M., and Ison, S. (2009). The effects of area-wide road speed and curvature on traffic casualties in england. *Journal of transport geography*, 17(5):385–395.

[Wang et al., 2017] Wang, J.-H., Abdel-Aty, M., and Wang, L. (2017). Examination of the reliability of the crash modification factors using empirical bayes method with resampling technique. *Accident Analysis & Prevention*, 104:96–105.

[Wang et al., 2015] Wang, L., Abdel-Aty, M., Shi, Q., and Park, J. (2015). Real-time crash prediction for expressway weaving segments. *Transportation Research Part C: Emerging Technologies*, 61:1–10.

[Wood, 2005] Wood, G. (2005). Confidence and prediction intervals for generalised linear accident models. *Accident Analysis & Prevention*, 37(2):267–273.

[Wood and Donnell, 2016] Wood, J. and Donnell, E. T. (2016). Safety evaluation of continuous green t intersections: A propensity scores-genetic matching-potential outcomes approach. *Accident Analysis & Prevention*, 93:1–13.

[Wood et al., 2015] Wood, J. S., Gooch, J. P., and Donnell, E. T. (2015). Estimating the safety effects of lane widths on urban streets in nebraska using the propensity scores-potential outcomes framework. *Accident Analysis & Prevention*, 82:180–191.

[World Health Organisation, 2018] World Health Organisation (2018). Global status report on road safety.

[Wright, 1992] Wright, S. P. (1992). Adjusted p-values for simultaneous inference. *Biometrics*, pages 1005–1013.

[Xie et al., 2014] Xie, K., Wang, X., Ozbay, K., and Yang, H. (2014). Crash frequency modeling for signalized intersections in a high-density urban road network. *Analytic methods in accident research*, 2:39–51.

[Yanmaz-Tuzel and Ozbay, 2010] Yanmaz-Tuzel, O. and Ozbay, K. (2010). A comparative full bayesian before-and-after analysis and application to urban road safety countermeasures in new jersey. *Accident Analysis & Prevention*, 42(6):2099–2107.

[Yu et al., 2014a] Yu, H., Liu, P., Chen, J., and Wang, H. (2014a). Comparative analysis of the spatial analysis methods for hotspot identification. *Accident Analysis & Prevention*, 66:80–88.

[Yu et al., 2014b] Yu, H., Liu, P., Chen, J., and Wang, H. (2014b). Comparative analysis of the spatial analysis methods for hotspot identification. *Accident Analysis & Prevention*, 66:80–88.