Modelling Terrorism Incidents in Iraq as a Log-Gaussian Cox Process

Alice Miranda Hankin

Bachelor of Science (Honours)
Department of Statistics
The University of Auckland
New Zealand

Abstract

This project aims to analyse terrorism trends in Iraq both spatially and temporally. Integrated Nested Laplace Approximation, a Bayesian technique, is used to fit log-Gaussian Cox models to data from the Global Terrorism Database. A spatial model fitted to data from 2017 gives strong evidence that both population density and the distance to the nearest primary road are highly correlated with terrorism intensity. A model with an autoregressive temporal component is then fit to the data from 2007 to 2018 inclusive; a weak temporal year-on-year correlation is found. Two models are fitted to the 2017 data - one to bombing-type attacks and the other to all other attack types, indicating that bombing attacks are significantly more correlated with road distance and less correlated with population than other attack types.

Contents

1 Introduction **4**

1.1 Terrorism in Iraq 4

1.2 A Brief Literature Review 5

1.2.1 Other Statistical Analysis of Terrorism Events 5

1.2.2 Aims of this project 6

2 Log-Gaussian Cox Processes, INLA, and NeSI **7**

2.1 Point Processes 7

2.2 Poisson Point Processes 8

2.2.1 Homogeneous Poisson Point Processes 8

2.2.2 Non-homogeneous Poisson Point Processes 10

2.3 Log-Gaussian Cox Processes 11

2.3.1 Cox Processes 11

2.3.2 Gaussian Random Fields 11

2.3.3 Parameters of the Gaussian Random Field 12

2.3.4 Log-Gaussian Cox Processes 12

2.3.5 Incorporating Other Information into the Model 13

2.4 INLA 15

2.4.1 What is INLA? 15

2.4.2 Bayes' Theorem 15

2.4.3 Hierarchical modelling 16

2.4.4 Implementing INLA in R 16

2.5 NeSI 17

3 Data **18**

3.1 Geography of Iraq 18

3.1.1 Population 19

3.1.2 Road Network Data 20

3.2 Terrorism data 21

3.2.1 Data source 21

3.2.2 Accessing and tidying the data 22

3.2.3 Plotting the data 23

4 Modelling **27**

4.1 A Spatial Model 27

4.1.1 Initial Model 27

4.1.2 Population covariate 31
4.1.3 Both covariates 34
4.2 A Spatio-Temporal Model 36
4.2.1 Year-on-year - AR(1) 36
4.2.2 Month-on-month - AR(1) Model 37
4.3 A Marked Model 40
4.3.1 Exploring data types 40
4.3.2 Two models - bombing versus all other attack types 41

5 Discussion **45**
5.1 Further work and limitations 46
5.1.1 Covariance Functions 46
5.1.2 Possible Other Covariates 46
5.1.3 Other Marked Models 47
5.1.4 Monthly Model 47
5.1.5 Spatial Accuracy of the Data 47
5.2 Conclusion 48

Chapter 1

Introduction

In 2017, terrorist activity was responsible for 0.05% of deaths worldwide, with over 26,000 deaths attributed to this cause. Although both the frequency of terrorist attacks and the coverage of such by the media has fluctuated over the past decade, terrorism remains a substantial public cause of concern to this day [1].

There is surprisingly little empirical information when it comes to research and countering terrorism [2]. There are, however, numerous reasons why it is crucial to have a comprehensive understanding of such. The allocation of money and resources to terrorism prevention and the assessment of existing prevention initiatives are all based on the analysis of data [3]. In other words, understanding patterns of terrorism is necessary to mitigate the risks of terrorism and protect the public [4].

This report will look at spatio-temporal data points corresponding to terrorist events in Iraq between 2007 and 2018. In particular, the locations of terrorist events will be modelled in order to infer, on a broad scale, what may be driving their spatial and temporal structure. Although it is impossible to measure all the economic, political, psychological, religious and social factors that nurture terrorism [5], a few key aspects (such as population density and road network access) that are quantifiable will be looked at, in order to see how they affect terrorist activity. Additionally, the different driving mechanisms behind the different classifications of attacks will be investigated.

1.1 Terrorism in Iraq

The Global Terrorism Index (GTI) is an annual report which aims to summarise global patterns in terrorism. It aims to provide a score for every country, rating the impact terrorism has had on it, based on the number of fatalities, injuries, and property damage a country has experienced due to terrorism in a given year. Between the years 2004 and 2017, the GTI has reported Iraq as being the most severely impacted by terrorist activity; following 2018, it is ranked second only to Afghanistan [6].

The word terrorism can be used to mean a variety of things, but the GTI defines it as "the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation" [7]. The combination of a political system alteration, regional upheavals, and the promotion of radical religious/political ideologies gave rise to increased terrorist activity in 2003 with the commencement of the Iraq war following the attacks on the World Trade Centre in September 2001 [5]. This combination of factors resulted in Iraq having seen the highest number of deaths caused by terrorism in any country between 2001 and 2019, with over 66 thousand deaths attributed to this cause [6].

The fact that Iraq ranks so highly in the Global Terrorism Index means that it is particularly of interest when looking to model terrorism data. Having a large dataset means that there is potentially more scope to investigate possible trends and relationships - perhaps more explanatory variables can be used to model the data than could be possible for a smaller dataset. Modelling the spatial and temporal locations of these events will not only allow one to understand the patterns of terrorist activity in the world's most terrorism-dense country, but it can also give insight into how terrorist activity manifests itself in other locations.

1.2 A Brief Literature Review

It is worth noting that the specific type of model used in this project is log-Gaussian Cox, a point process typically used to model clustered or non-homogeneous patterns. The INLA R package [8][9] will be used to estimate the parameters of this model. This is mentioned here, preceding a brief discussion about the importance of this project and how it differs from other similar research. The models and methods that will be used will be detailed further in chapter 2.

There have been a few other studies that analyse either a similar dataset or utilise a similar technique. Those that do are discussed briefly below.

1.2.1 Other Statistical Analysis of Terrorism Events

In 2012, Gao, Guo, Liao, Webb & Cutter analysed terrorist attacks throughout the world using scan statistics [10]. This is a method in which spatial clusters are detected by using circular windows that scan across the geographical area, testing if the number of events is higher than expected. Monte-Carlo methods are then used to find the statistical significance of the clusters [11]. Some limitations of Monte-Carlo methods will be mentioned in section 2.4.

Siebeneck, Medina, Yamada & Hepner used Geographic Information Science (GIS) techniques and cluster identification analyses to look at terrorist incidents in Iraq between 2004 and 2006 [12]. They looked at how terrorist cells emerged and evolved over the three-year study period, finding that the rate of growth of these clusters, both spatially and temporally, was gradual. The research determined that the incidents were mostly clustered around Northern Iraq in 2004 but spread to Western Iraq and Central Iraq by the end of 2006.

Townsley, Johnson & Ratcliffe [13] also looked at the space-time dynamics of terrorist activity - in particular, the use of improvised explosive devices - in Iraq over a three-month period in early 2004. They used a contingency table for the space-time difference of each pair of events.

In 2017, Clark and Dixon used INLA to model terrorism data in Iraq. Their goal was to demonstrate how the choice of model leads to differing conclusions on how violence spreads from 2003–2010. Rather than use a latent model (an example of which is the log-Gaussian Cox model), here, the spread of violence has been modelled using "theories of self-excitation", which gives the probability of an event occurring as a function of previous events [14].

None of the research mentioned above utilises the most recent terrorism data, in particular, data from 2017 and after. An emphasis on the different types of terrorist attack has also not been studied. As will be clarified in the following section, these are both very relevant to the aims of this project.

1.2.2 Aims of this project

Before 2019, log-Gaussian Cox models had not been used for terrorism research [15], although they have successfully been applied in similar areas such as the analysis of crime statistics [16][17]. In 2019, Python, Illian, Jones-Todd & Blangiardo [15] used a Bayesian hierarchical framework to model the lethality of terrorism, severity, and frequency of lethal terrorist attacks across the world between 2010 and 2015. The goal of this project is to expand on this research. In particular, the same basic framework will be used; however, it will be focusing specifically on Iraq in 2017 - the reason for this year being chosen is that it was when the Iraq war officially ended [18]. It is hoped that analysing this time period will return novel results, in contrast to those studies discussed above, which concentrated only on periods during the war. This project will also analyse different types of terrorist attack; in particular, bombing type attacks. A good reason for this is that Iraq is currently investing in counter-terrorism, which is actively working to decrease the number of suicide bombings [6]; research in this area is valuable.

Chapter 2

Log-Gaussian Cox Processes, INLA, and NeSI

The data to be analysed for this project is described in detail in chapter 3, but it reduces down to a list of terrorism events along with their location, time, and any other attributes of the attack (such as the type of attack). Some framework is needed for looking at this type of data. In this chapter, the idea of a point process is introduced. It is useful to think of this as a model for the intensity of terrorism, of which the observed data is one eventuality. A specific type of point process - the log-Gaussian Cox process - is then discussed in detail; this will be the specific model that will be used for the terrorism data. The parameters of the point process will be estimated using INLA [8]; some of the mathematics underlying the R package is also touched on. Finally, the way in which the R [9] code will be run via high-performance computing system NeSI is mentioned.

2.1 Point Processes

A point pattern is a collection of events occurring in time and/or space. This space could be one dimensional (for example, a period of one hour of time), two dimensional (for example, the country of New Zealand), or multi-dimensional. “Temporal”, “spatial”, or “spatio-temporal” are used to describe whether the data has time- or space-based coordinates.

For example, if the events we are interested in are the times of bus arrivals in a one hour period, first consider the hour-long period of time as a one-dimensional space. Denoting the space by a line and each arrival as a × symbol, one eventuality might look like Figure 2.1 below.

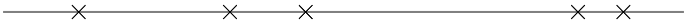


Figure 2.1: An example of a one-dimensional point pattern; the × symbols represent events over a one-dimensional space represented by the line

Alternatively, one might have locations of events (for example, terrorism events) in a two-dimensional space (for example, the country of Iraq), such as in Figure 2.2.

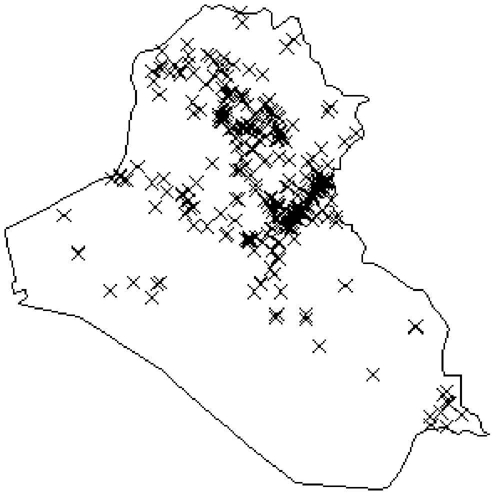


Figure 2.2: Locations of terrorism events in Iraq in 2019 as a two-dimensional point pattern; each × symbol represents a terrorist attack

For any given eventuality, such as that in Figure 2.2, we wish to estimate what the underlying intensity looks like. For this, some types of point process must first be introduced.

2.2 Poisson Point Processes

Let X be a discrete random variable such that $X \sim \text{Poisson}(\lambda)$. Then X has the following probability density function:

$$\mathbb{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

λ is called the rate or the intensity of the process, and is also the mean and variance of the distribution.

A Poisson point process is a point process that fits the criteria:

- The number of events in any region is distributed as per a Poisson distribution
- The number of events in any two disjoint regions are independent of each other

There are two types of Poisson point process - homogeneous and non-homogeneous.

2.2.1 Homogeneous Poisson Point Processes

Homogeneous Poisson processes are said to have complete spatial randomness; the rate is constant over the space [19].

Suppose we have some measure on the space. For a two-dimensional spatial process, this will be area. For other models, length, time or volume may be an appropriate measure. Define the rate λ as the average number of events per unit of space. Then, in a homogeneous Poisson point process, we have:

$$N(A) \sim \text{Poisson}\left(\lambda|A|\right)$$

where $|A|$ denotes the size of A with respect to the measure and $N(A)$ refers to the number of points in region A [19].

For example, consider a space of 9 units squared, and suppose $\lambda = 10$. Any area of x units squared should have a count of events that is Poisson$(10x)$ distributed. In particular, in any area of one unit squared, the number of events should be distributed following a Poisson distribution with a rate of 10.

We can use the `spatstat` [20] package in R [9] to demonstrate this using a simulation.

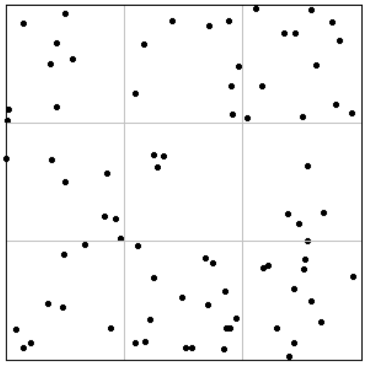


Figure 2.3: One realisation of a homogeneous Poisson point process with rate $\lambda = 10$ over an area of 9 units^2

8	8	12
7	3	5
8	16	11

Figure 2.4: Count of events per area of one unit squared in Figure 2.3

On the left-hand side, we have an area of 9 units squared and a rate of $\lambda = 10$. We can see the counts of events in each 1 unit squared area on the right-hand side. These counts follow a Poisson(10) distribution – we can see what a simulated Poisson(10) distribution looks like in Figure 2.5.

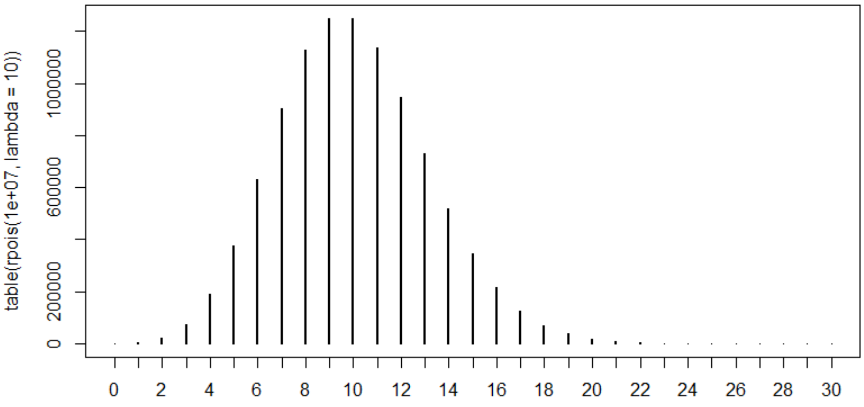


Figure 2.5: Simulating 10 million Poisson random variables with rate 10

We can see that a randomly generated value from the Poisson(10) distribution is most likely to be around 10; it has a very low (5%) chance of being greater than 17 or less than 4. This gives some verification that the values in Figure 2.4 do come from this distribution.

2.2.2 Non-homogeneous Poisson Point Processes

The rate λ in a non-homogeneous Poisson process is not constant over the space; instead, it is a function of the space [19]. Again, we can see what this looks like using `spatstat`.

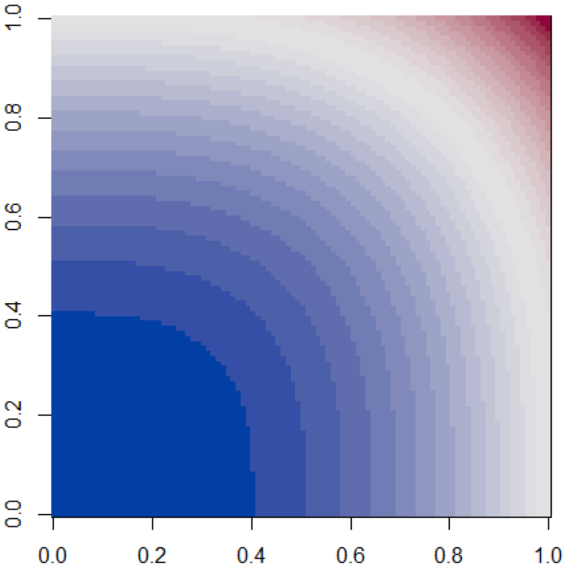


Figure 2.6: Intensity map given by $\lambda(\boldsymbol{s}) = 100(x^3 + y^3)$

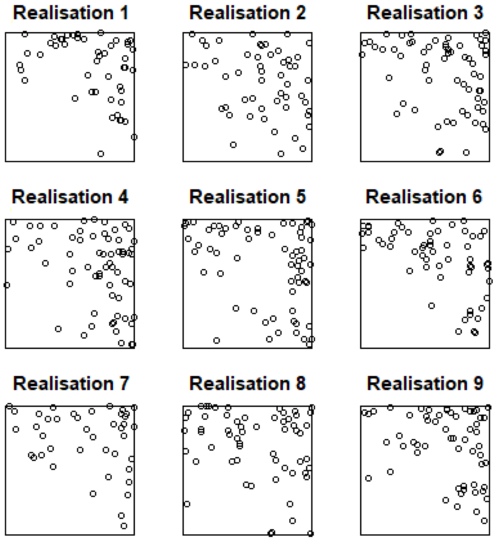


Figure 2.7: Nine simulations of non-homogeneous Poisson processes with density as given in Figure 2.6

The image on the left shows the intensity λ as a function of the space i.e. $\lambda(\mathbf{s})$ where $\mathbf{s} = (x, y)$ is a vector of the x and y coordinates. We have defined $\lambda(\mathbf{s}) = 100(x^3 + y^3)$. The blue areas have the lowest intensity and the red areas have the highest intensity.

The image on the right gives 9 different simulations of outcomes with this intensity function. We can see that the events represented by $\circ$ are densest in the upper right corner where there is high underlying intensity and least dense in the lower-left corner where there is low underlying intensity.

Recall that observed point patterns, like those in Figure 2.7, are just a single realisation of the point process in Figure 2.6. Given a single observed pattern (i.e. one of the realisations in Figure 2.7), we might aim to estimate the underlying latent process. For example, if we assumed that the underlying intensity was $\lambda(\mathbf{s}) = ax^3 + by^3$, then we could perform analysis to estimate the parameters a and b. However, in the case of our terrorist attacks, we have no knowledge of what the intensity function looks like; we have to introduce the idea of a doubly stochastic model.

2.3 Log-Gaussian Cox Processes

A log-Gaussian Cox process is a generalisation of the non-homogeneous Poisson point process, with two random components, the first being a Gaussian random field, and the second being the Poisson process discussed in section 2.2.2. This definition will be elaborated on here.

2.3.1 Cox Processes

We define a Cox process as a non-homogeneous Poisson process with a random intensity function.

The difference between a Cox process and a non-homogeneous Poisson process is that the latter has a deterministic rate, whereas the former has not. With a Cox process, the underlying intensity function (or field) is also viewed as random. Cox processes are also known as doubly stochastic processes [19]. In order to discuss the log-Gaussian Cox process, a particular type of Cox process, a Gaussian random field must first be defined.

2.3.2 Gaussian Random Fields

Multivariate Gaussian Distribution

A continuous random variable X has a Gaussian (or normal) distribution with mean μ and variance $\sigma^2 > 0$ if it has the probability density function

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

We write $X \sim \mathcal{N}(\mu, \sigma^2)$

If each element of $\mathbf{X} = (X_1, ..., X_n)^T$ has a Gaussian distribution. Then we say $\mathbf{X}$ is distributed as per a multivariate Gaussian distribution, and we denote this by:

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\mu} = (E(X_1), E(X_2), ..., E(X_n))^T$ and $\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}(X_i, X_j)$

$\mathbf{X}$ then has probability density function

$$(2\pi)^{-\frac{n}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

for positive definite $\boldsymbol{\Sigma}$; we call $\boldsymbol{\Sigma}$ the covariance matrix and $Q := \boldsymbol{\Sigma}^{-1}$ the precision matrix.

What is a Gaussian Random Field?

A random field f over a space S is a collection of random variables $\{f(s) \mid s \in S\}$ [21].

A random field G is Gaussian if for any set of locations $s_1, ..., s_n$ in S, $G(s_1), ..., G(s_n)$ has a multivariate Gaussian joint distribution.

2.3.3 Parameters of the Gaussian Random Field

Matérn covariance is one function that can be used to define the covariance matrix $\mathbf{\Sigma}$ for our Gaussian random field. The Matérn covariance function is used in this project since it fits in with the INLA methodology; this is elaborated on in section 2.4.

Suppose we have two points that are d units apart from each other. Then the covariance is given by:

$$C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{d}{0.5r}\right)^\nu K_\nu \left(\sqrt{2\nu}\frac{d}{0.5r}\right)$$

where $\Gamma(x)$ is the gamma function and $K_\nu(x)$ is the modified Bessel function [22].

Other reparametrisations instead use $\kappa := \sqrt{4\nu}/r$ or $\rho = 0.5r$ [23].

We call ν the smoothness parameter, σ the marginal standard deviation, and r the range. From here on out, ν will be treated as non-random. This is a way of simplifying the calculations and is the default in the programs that will be used here.

It is worth mentioning, for completeness, that the Matérn field is the solution $G(\mathbf{s})$ to the stochastic partial differential equation:

$$(\kappa^2 - \Delta)^{\alpha/2} G(\mathbf{s}) \overset{d}{=} W(\mathbf{s})$$

where $W(\mathbf{s})$ is spatial white noise and $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial s_i^2}$ is the Laplacian operator [24].

And so, in short, we have seen that a Gaussian random field can be characterised by its range and marginal variance parameters.

2.3.4 Log-Gaussian Cox Processes

A log-Gaussian Cox process (or LGCP) is a Cox process for which the underlying intensity function (sometimes called the latent field) is the exponential of a Gaussian Markov random field [19].

We model the intensity as:

$$\Lambda(\mathbf{s}) = \exp\{\beta_0 + G(\mathbf{s}) + \varepsilon\}$$

or alternatively

$$\log(\Lambda(\mathbf{s})) = \beta_0 + G(\mathbf{s}) + \varepsilon$$

Recall from section 2.2.2 that $\mathbf{s}$ is a vector corresponding to a location in space. We define β_0 to be a constant called the intercept, $G(\mathbf{s})$ to be a Gaussian random field, and ε to be the error term. The exponential function is used to ensure that the underlying field Λ is positive everywhere in the space.

We are interested in estimating the parameter β_0 (as well as possibly other β_is, coefficients of covariates - see section 2.3.5) as well as the parameters r and σ of the Gaussian random field. Note that in INLA (see section 2.4), r and σ are called hyperparameters, whereas the β_is are called parameters or fixed effects.

2.3.5 Incorporating Other Information into the Model

Recall that we also may have other information which we wish to incorporate into the model. Namely, we might have information about Iraq, information about each data point, or temporal data. The log-Gaussian Cox framework allows us to incorporate this information into the model.

The AR(1) model

An autoregressive(p) model is a type of random process where the value of the next variable depends linearly on the previous p values, plus some error term.

For the autoregressive(1) – which we shorten to AR(1) – model, the value of the next variable depends only on the previous value. The value of y at time t looks like: $y_t = \rho y_{t-1} + \varepsilon_t$ where ε_t is the error term, and ρ is a constant.

In our case, instead of a real value for y_t, we are dealing with Gaussian random fields. Our aim is to fit a log-Gaussian Cox model to each, say, month, and then use an AR(1) process to model the between-month correlation.

Let us use i as an index for the months (i.e. January is associated with $i = 1$ and so on). For month i, we have a model which looks like:

$$\Lambda_i(\mathbf{s}) = \exp\left\{\beta_0 + G_i(\mathbf{s}) + \varepsilon\right\}$$

Now, for $G_i(\mathbf{s})$, we use the autoregressive structure, i.e.

$$G_i(\mathbf{s}) = \rho G_{i-1}(\mathbf{s}) + \varepsilon'_i$$

where ρ is a constant and ε' has a multivariate normal distribution with mean $\mathbf{0}$.

We then define the correlation between two points in time and space as the correlation in time multiplied by the correlation in space [25].

A Marked Point Process

It may also be interesting to consider other information along with the time or location of each point.

For example, with a dataset related to terrorism, we may also be interested in the type of attack.

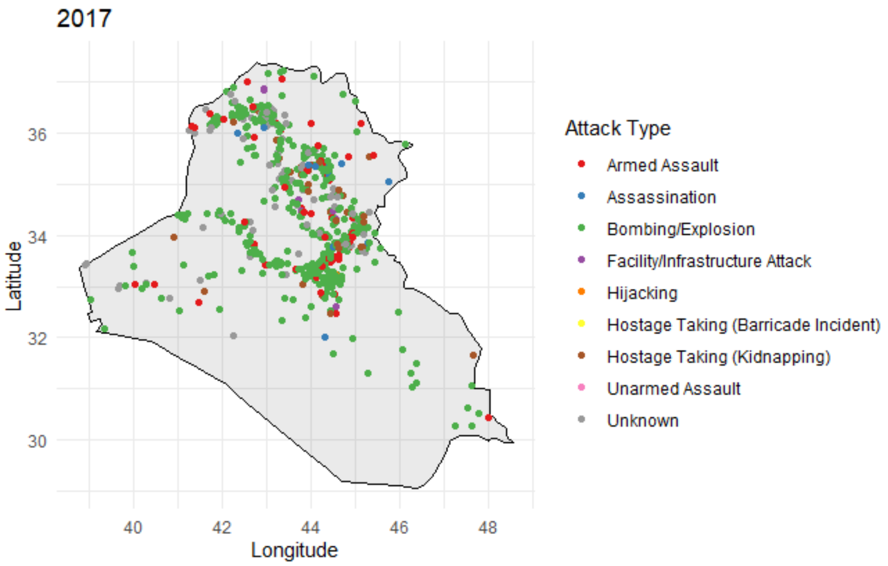


Figure 2.8: Locations of terrorist events in Iraq 2017, coloured by a categorical mark

This is another characteristic attached to each point - we call it a mark, and call the point process a marked point process. In this project, models for each of the marks will be fitted independently, however, it is possible to fit models which incorporate some dependence between the marks.

A Point Process with Covariates

A covariate is other information about the space that we can use as an extra factor in modelling - in this case - the terrorism data. This can be continuous or discrete.

For example, we might be interested in how population affects our data. We may expect that if the population density is high, then there is likely to be more terrorist activity, compared to an area with low population density. Looking at the following plot, we can see that this seems to be the case.

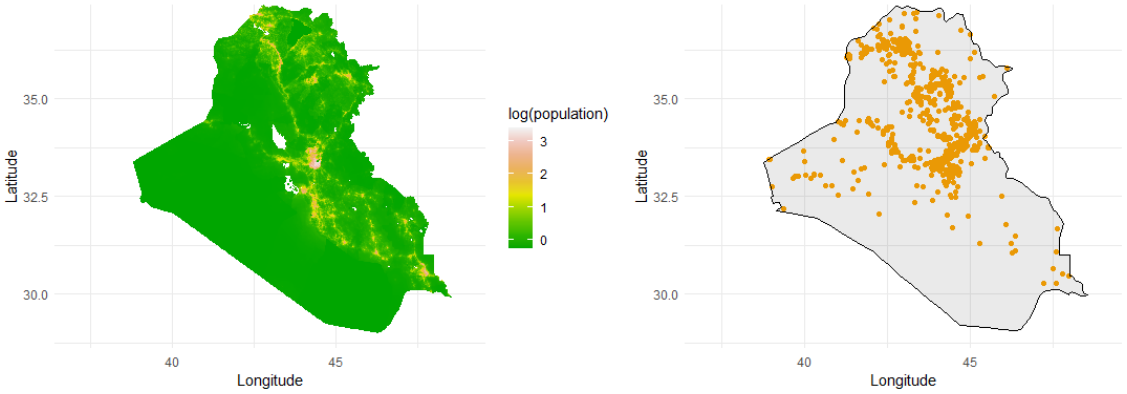


Figure 2.9: Terrorist attacks (right) and population on a log scale (left) in Iraq, 2017

Suppose that $Population(\mathbf{s})$ returns the population at point $\mathbf{s}$. Now, as discussed in the following section (2.4), the methods that will be used do discretise the space, so defining population at an infinitely small point is not necessary. Then, the model will be:

$$\Lambda(\mathbf{s}) = \exp\{\beta_0 + \beta_1 \cdot Population(\mathbf{s}) + G(\mathbf{s}) + \varepsilon\}$$

Of course, we can combine our covariate, mark, and temporal components into one model if we so wish - they are not mutually exclusive.

2.4 INLA

Due to the doubly-stochastic nature of Cox processes, Bayesian inference is a natural way of modelling LGCPs [19]. This is due to the hierarchical structure of Bayesian models; this will be discussed further in section 2.4.3. Two efficient techniques for Bayesian inference are Markov Chain Monte Carlo (MCMC) and Integrated Nested Laplace Approximation (INLA) [26].

MCMC is widely regarded as a slow but exact method, whereas INLA is fast but approximate [26]. In 2013, Taylor and Diggle found a trade-off between faster computation and the errors of approximation; both MCMC and INLA are helpful tools for spatial data analysis. They found that MCMC was more accurate and flexible, however, INLA provided access to a much wider class of latent Gaussian models [26].

This project will be using INLA, however many of the same methods could easily be applied to an MCMC framework instead; the drawback being computational time as well as issues related to overlapping data points, which will be discussed in section 4.1.1.

2.4.1 What is INLA?

Integrated Nested Laplace Approximation, or INLA, was developed in 2009 as an alternative to Markov Chain Monte Carlo methods such as JAGS and BUGS. It is designed to focus specifically on models with an underlying intensity that is a Gaussian Markov random field [22]. Rather than use a simulation as MCMC does, INLA is based on Laplace approximations, making it more computationally efficient and more accessible than MCMC methods [27].

2.4.2 Bayes' Theorem

Bayes' Theorem says:

$$p(\theta \mid y) = \frac{p(y \mid \theta) \cdot p(\theta)}{\int p(y \mid \theta) \cdot p(\theta) \cdot d\theta}$$

Each of these terms have names:

- $p(\theta \mid y)$ the posterior
- $p(\theta)$ the prior
- $p(y \mid \theta)$ the likelihood
- $p(y) = \int p(y \mid \theta) \cdot p(\theta) \cdot d\theta$ the marginal

Bayesian analysis is one way of determining parameters θ given data y. Recall that our aim is to find parameters for the log-Gaussian Cox model, given the terrorism data. The likelihood function is determined by the model and is shown in Equation 2.1. INLA's default priors are to be used for the analysis since prior sensitivity analysis is outside the scope of this report.

2.4.3 Hierarchical modelling

INLA deals with the doubly stochastic Cox process using a Bayesian framework.

The structure of our model looks like:

- Data: the set of observed locations of terrorist attacks
  Our data is modelled by a likelihood function (see Equation 2.1). We have optional parameters $\beta_0, \beta_1, ...$ which act as the coefficients for covariates. These are assumed to be conditionally independent given the latent field.
- Latent field
  This describes the dependency structure of the data - we assume this is a Gaussian random field, and denote this $G(\mathbf{s})$.
- Hyperparameters
  These are the parameters (r and σ) which control the latent field.

For the basic model

$$\log(\Lambda(\mathbf{s})) = \exp\{G(\mathbf{s}) + \varepsilon\}$$

we have the likelihood function

$$\log(f(y \mid r, \sigma)) = |S| - \int_S \Lambda(\mathbf{s}) d\mathbf{s} + \sum_{s_i \in Y} \Lambda(s_i) \qquad (2.1)$$ [28]

Where Y is the set of locations of terrorist events and S is our space (the 2-dimensional shape of Iraq).

In order to calculate our posterior distributions for the parameters, the likelihood must first be found, and as such, we are required to calculate the integral of the intensity function. This is very expensive to compute numerically and often impossible to compute analytically [28]. INLA uses various analytical approximations as well as numerical integration to calculate the posterior distribution, Laplace approximation, hence the name, being a component. Part of this involves discretising the space in order to make the integrals possible to calculate, which is why INLA can deal with overlapping data points, as well as covariates which cannot be calculated on an infinitesimally small scale. The details of this are out of the scope of this report but can be found in Martino & Rieber, 2020 [29].

2.4.4 Implementing INLA in R

The programming language R [9] will be used for this project. There is an INLA package for R, which can be downloaded from `https://www.r-inla.org/` [8]. The package `inlabru` will also be used, which provides wrappers for many INLA commands [30].

2.5 NeSI

Computing the fit for a spatio-temporal model involves a lot of computer power - several gigabytes of memory and storage are required. New Zealand eScience Infrastructure (NeSI) provides a platform to access high-performance computing (HPC) in New Zealand [31].

In practice, this works by the user connecting their local machine to the NeSI cluster, enabling them to send files back and forth. R code, or similar, can be run via SLURM (Simple Linux Utility for Resource Management) scripts, which runs the code on the NeSI servers. The output of the R code is then saved to the cluster and can be transferred back to a local machine.

Using NeSI has required attainment of command-line skills, as well as Linux operating system knowledge. It has, of course, provided an incredibly valuable resource to this project; much of the code used here would not be able to be run without its help.

The code for this project can be found at `https://github.com/alicemhankin/inlabru-terrorism`.

Chapter 3

Data

At this point, the model to be fitted has been discussed. However, we do not yet know about the data to be used. This chapter will discuss a little bit about the geography of Iraq before discussing the source of our main dataset and then completing some exploratory data analysis.

3.1 Geography of Iraq

In this section, two covariates are introduced - population and distance to road. These were selected since these are likely to be highly correlated with the intensity of terrorism. According to Python et al in 2021, high population density, access by roads, and closeness to large cities increases the ease by which the communicated message travels through the audience. This means that these locations generally constitute more attractive targets to terrorists [32]. The distance-to-road measure may also be relevant to terrorists, as being in a road-dense area could make it easier to evacuate the crime scene faster.

3.1.1 Population

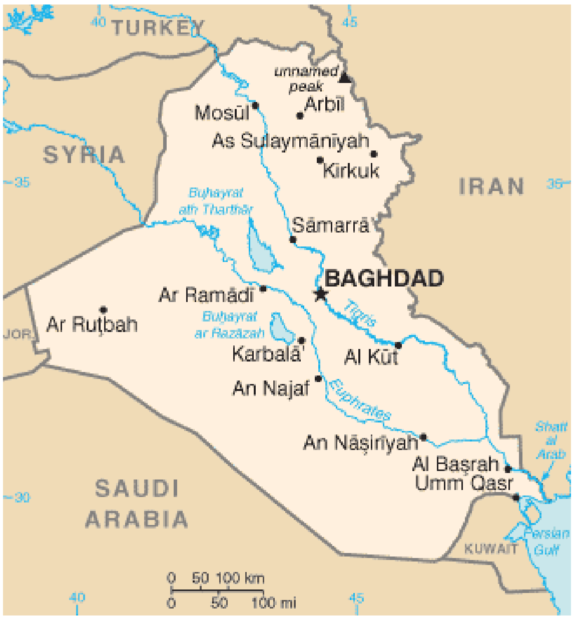


Figure 3.1: A map of Iraq [33] showing the capital city, Baghdad, as well as the Tigris-Euphrates river system

In order to put the spatial data in context, it is worth discussing the main geographical features of Iraq. The Tigris-Euphrates river system runs through the centre of the country. A major component of Iraq's economy comes from agriculture [34], so much of the population is centred around here [35]. In particular, one-third of the population of Iraq live in capital city Baghdad. To the south-west of the river lies arid desert, which comprises around two-fifths of the country [34].

We can source population data from `https://data.humdata.org` for every year since 2000. This spatial distribution is based on the country total adjusted to match the corresponding UNPD estimate [36].

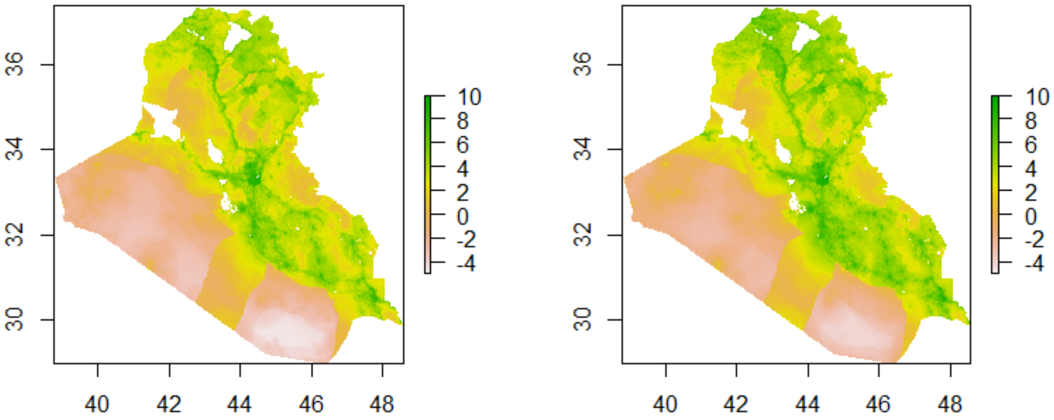


Figure 3.2: Population of Iraq in 2000 (left) and 2019 (right) on a log scale

Here, using R, the population on a logarithmic scale has been plotted. We have the population density in the year 2000 on the left and the population density in the year 2019 on the right. We can see that, comparing Figure 3.2 to Figure 3.1, there is a very low population in the desert area, and the population is most dense near Baghdad, and along where the Tigris-Euphrates river runs. It is apparent that there is very little change over the years. We can also plot the difference in population between these two years.

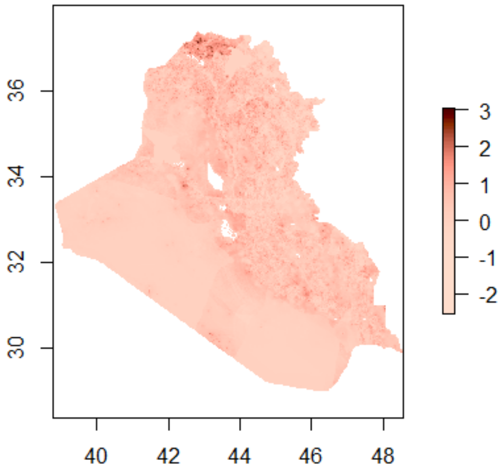


Figure 3.3: Difference in log(population) between the years 2000 and 2019

It seems as if there has been the highest change in population near the very north of the country. Other than this, the areas of dense population seem to have the highest increase in population over time.

3.1.2 Road Network Data

We can source road network data as a shapefile from `https://data.humdata.org` and plot this in R.

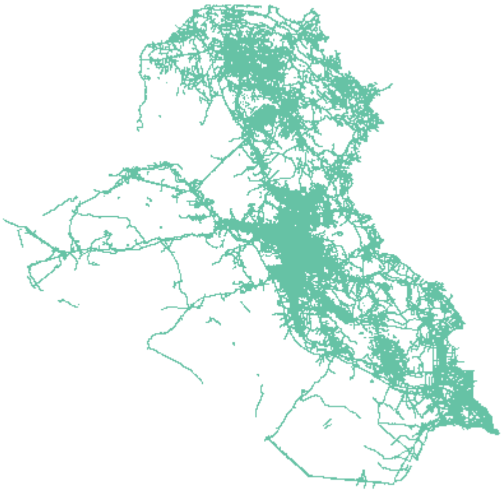


Figure 3.4: Every road in Iraq

This shows every road in Iraq in green. As expected, by comparing this to Figure 3.2, we can see that is a high density of roads where there is a high population. We are not interested in using this full data. Primarily, this is due to it being essentially a proxy for population. Also, the fact that there are over 82,000 roads recorded here means that coming up with a measure of road density would be difficult. This is discussed further in section 5.1.2.

We can also filter by the roads that are designated “primary roads”. There are 3,106 of these.

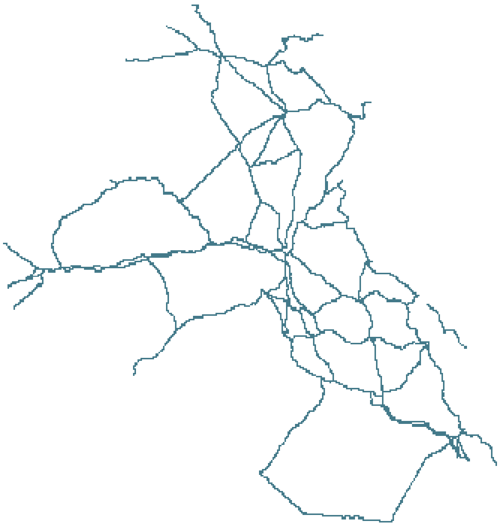


Figure 3.5: Primary roads in Iraq

Next, the terrorist data itself will be introduced.

3.2 Terrorism data

3.2.1 Data source

The data is sourced from the Global Terrorism Database (GTD), which defines itself as the most comprehensive unclassified database on terrorist attacks [7].

The approach that the GTD adopts is to collect/structure the data in a way that makes it useful to as many people as possible. As such, they act in the interest of inclusion, with the view that the user can filter out unwanted entries. This being said, they collect the data based on six criteria.

The event must have all three of:

- The incident must be intentional
- The incident must entail some level of violence or immediate threat of violence

- The perpetrators of the incidents must be sub-national actors (i.e. state terrorism is not included)

As well as two of:

- The act must be aimed at attaining a political, economic, religious, or social goal
- There must be evidence of an intention to coerce, intimidate, or convey some other message to a larger audience (or audiences) than the immediate victims
- The action must be outside the context of legitimate warfare activities

These events are flagged by machine learning algorithms that look through relevant news articles but are then manually reviewed to ensure they satisfy the inclusion criteria. The way that incidents are recorded is that events occurring in both the same geographic and temporal points are regarded as a single event, but if either the time or location is different, they will be regarded as separate [7].

3.2.2 Accessing and tidying the data

The data is downloadable from the GTD website as an .xlsx file. After importing into R, and filtering to be left with only the events that occurred within Iraq, there remain 26,593 events from between 1975 and 2019.

We can use the `visdat` [37] package in order to have a look at the amount of missing data, and the data types we have. The column names here are too small to read, but the image works as an overview.

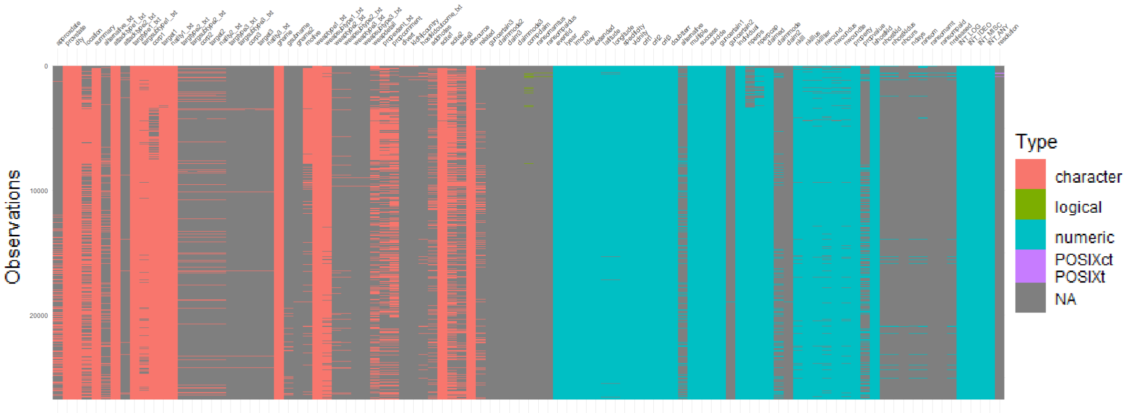


Figure 3.6: Visualisation of dataset showing data type of each entry and number of NA entries

We can see that there is a lot of missing data, however, most of these columns are not relevant to our investigation, or are empty by design. For example, we have three separate columns for target nationality, for the cases where there are multiple targets involved.

We have a mixture of character vectors and numerical vectors. Often, the character vectors are actually factors (for example, location, weapon type, attack type). Columns such as the number of people killed, time of attack, and attack location are all numerical.

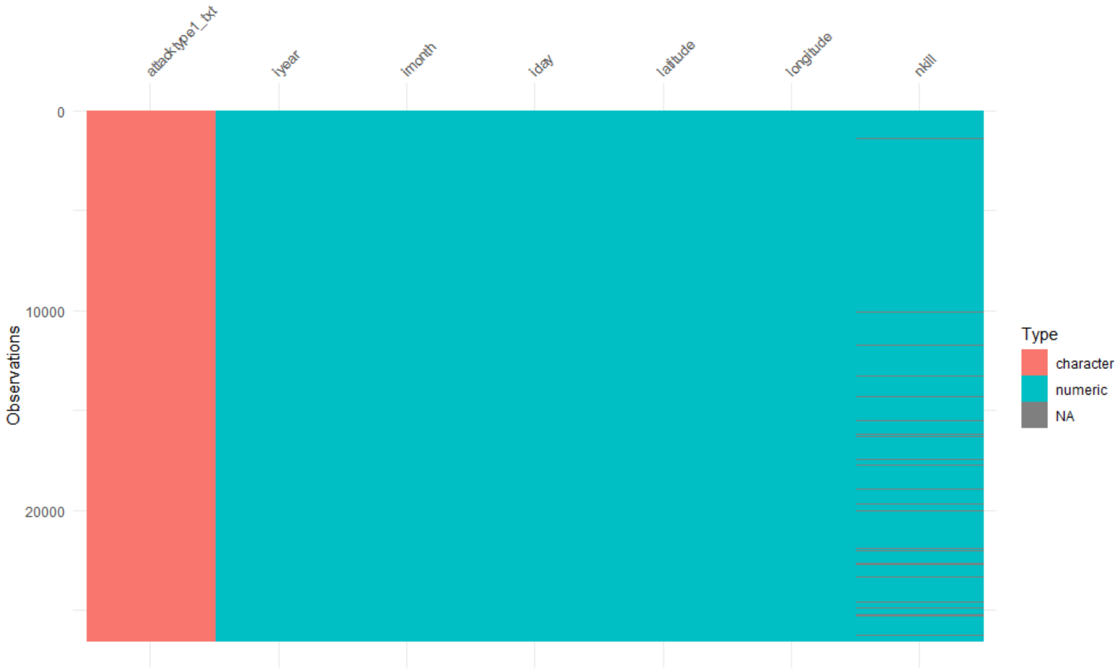


Figure 3.7: Visualisation of a subset of the dataset showing data type of each entry, and number of NA entries

If we subset to only the time, place and two of the marks (attack type and number of fatalities), there are not many missing values at all. In fact, there are no missing values in any column excluding nkill, in which 3% of the data is missing.

3.2.3 Plotting the data

We can look at the amount of data we have for each year:

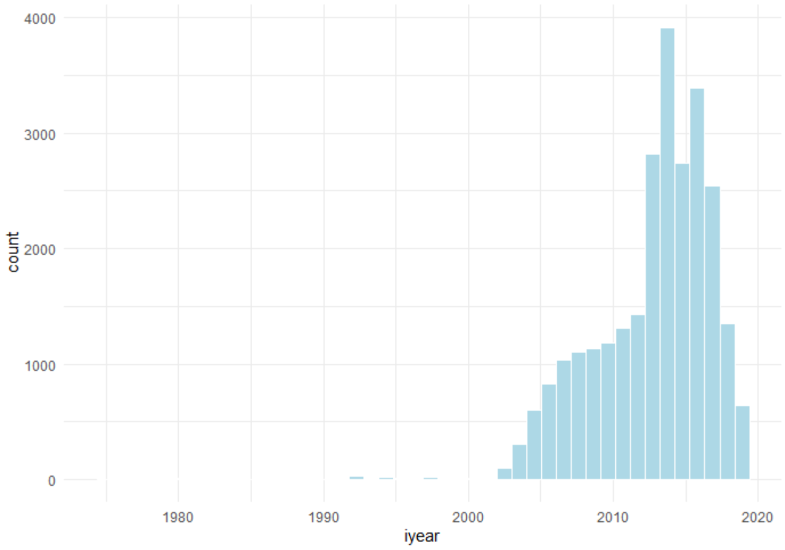


Figure 3.8: Histogram of the number of terrorist attacks in Iraq by year

For all years before 2003, we have less than 30 recorded events each year. We can see there is a steady increase in the number of events up to 2012, there is a spike from 2013-2017 when the number of events per year is over 2500. It then dies back down to the same as it was before the spike. It is worth noting that these figures may not be entirely representative of terrorism numbers over this time period; there may have been changes in the way the numbers were reported.

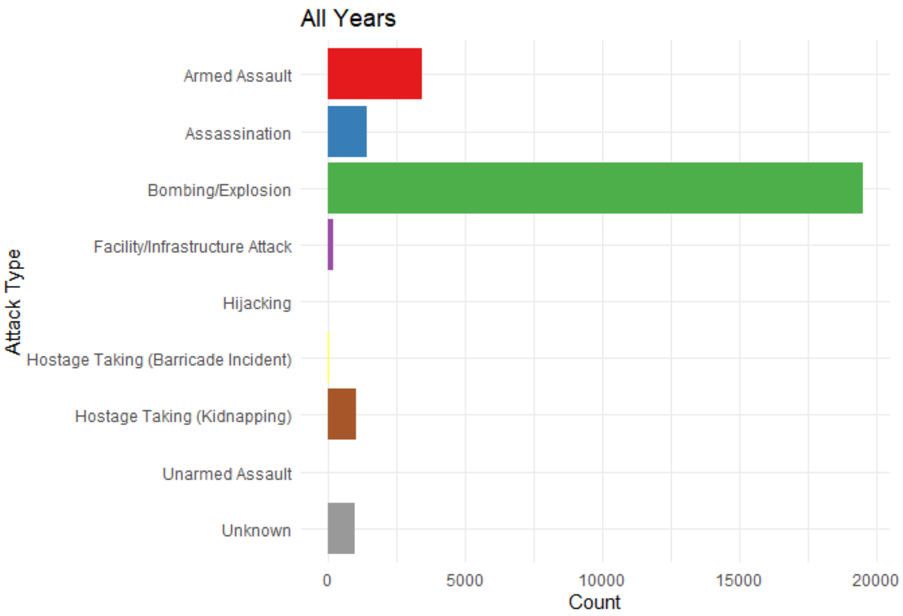


Figure 3.9: Histogram of terrorist attacks in Iraq by attack type

The most prevalent primary attack type is bombing/explosion with 19,529 of the 26,593 events being classified in this category. All other attack types have fewer than 3,500 events, the most popular being armed assault and the least being unarmed assault at only 9 events. There are 974 events that are unknown.

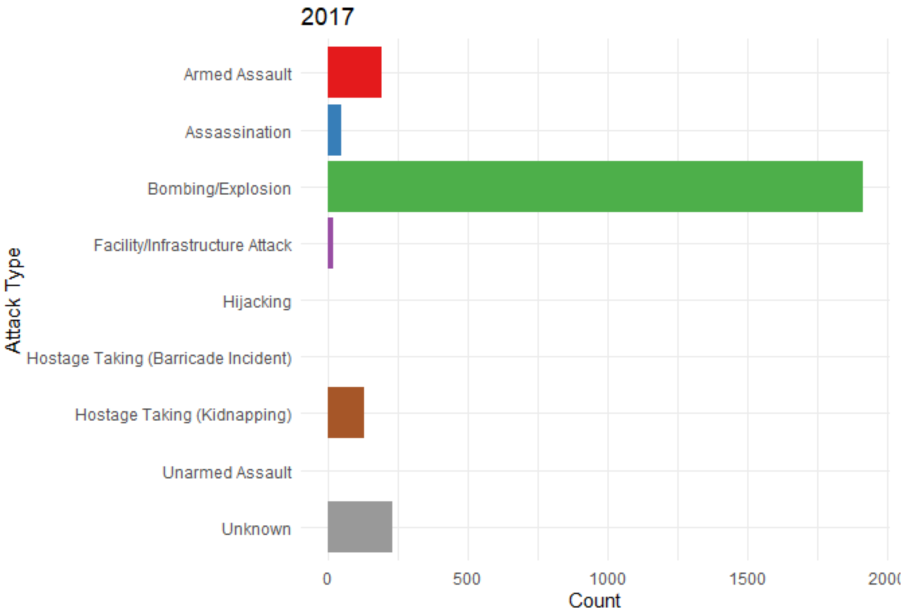


Figure 3.10: Histogram of terrorist attacks in Iraq, 2017, by attack type

The data for 2017 has a very similar distribution, with 1917 of the 2543 events being in the Bombing/Explosion category, which is the most prevalent. There is a lower proportion of events being categorised as armed assault, with the proportion being 7.5%, down from 12.9% in the full dataset.

The `rnaturalearth` package [38] allows us to be able to plot spatial data with reference to the map of the world.

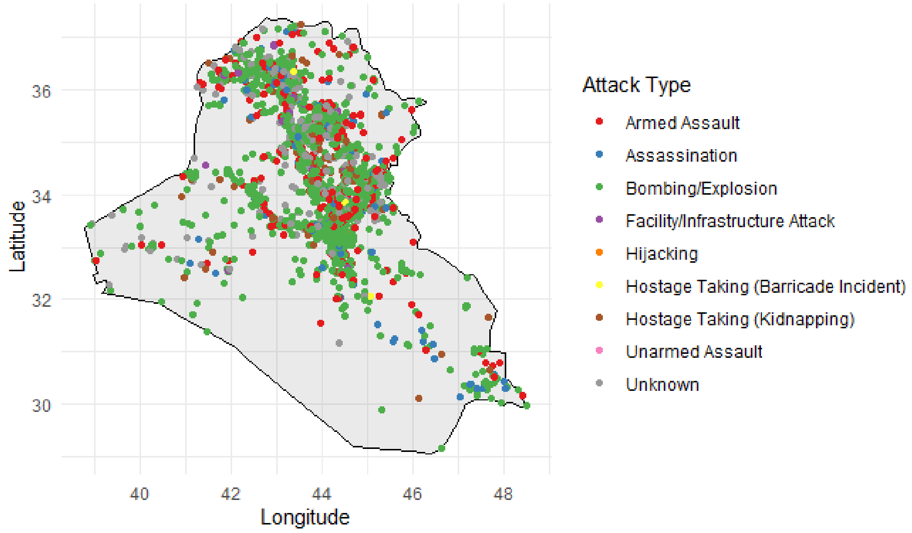
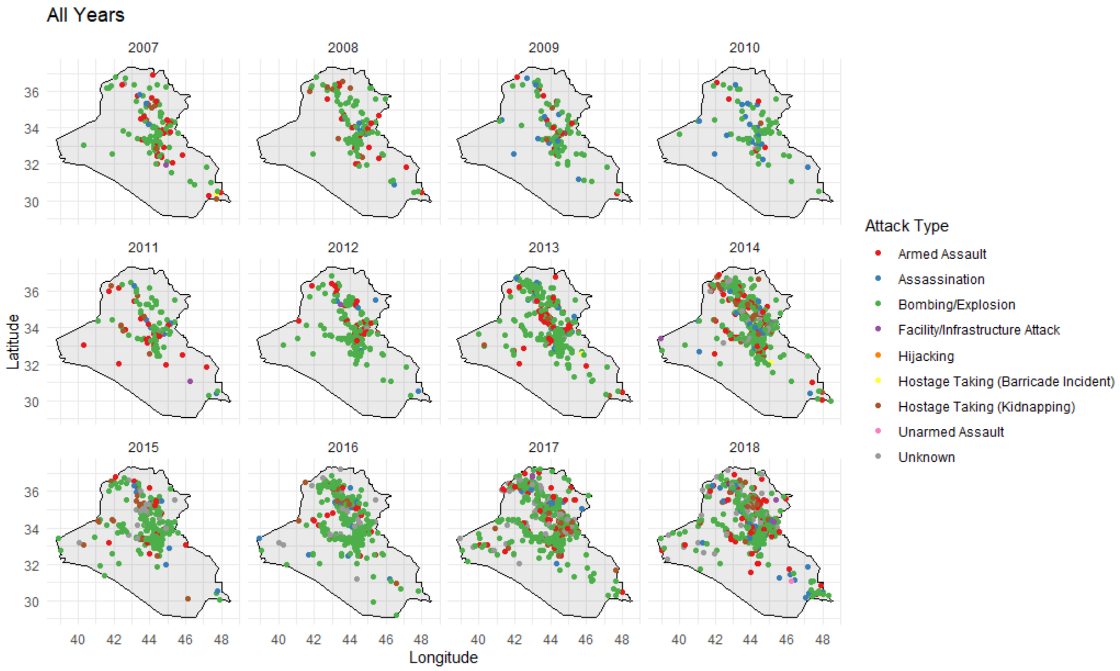


Figure 3.11: Location of terrorist events in Iraq by attack type

Comparing this to Figure 3.2, it seems very likely that the density of terrorist events is highly correlated with population. In particular, there is a cluster of events along the Tigris-Euphrates river and near Baghdad (see Figure 3.1). A plot showing the population and density of attacks in 2017 can be seen in Figure 2.9.



]

Figure 3.12: Location of terrorist events in Iraq by attack type and year

We can see that the points do tend to follow the same pattern each year. This is indicative of a high ρ parameter for the AR(1) model mentioned in section 2.3.5. This will be discussed further in section 4.2.1.

Chapter 4

Modelling

We have discussed the data we will be analyzing in chapter 3 and the techniques which we plan to use in chapter 2. And so, in this chapter, several models will be fitted. Initially, we use a spatial model - i.e. the temporal information is disregarded. Then, a spatio-temporal model is considered before a model which takes into account attack type.

4.1 A Spatial Model

4.1.1 Initial Model

First, a basic model will be fitted to the data. Consider all the terrorism events in 2017 as one realisation of a two-dimensional point process.

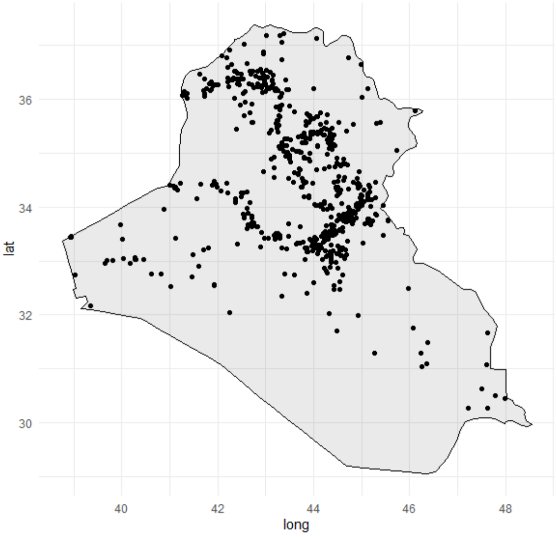


Figure 4.1: Location of terrorist events in Iraq in 2017

Note that our x and y axes are longitude and latitude; this is the format that the `inlabru` package requires the data to be in, and is consistent with the format of the covariates which will be used in later models.

The model we use is log-Gaussian Cox, i.e. the formula is given by:

$$\Lambda(\mathbf{s}) = \exp\{\beta_0 + G(\mathbf{s}) + \varepsilon\}$$

Recall from section 2.3.3 that we wish to estimate β_0 as well as the parameters of the Gaussian random field, range r and standard deviation σ.

As discussed previously, INLA has a system for estimating the parameters of a log-Gaussian Cox process. An excerpt from the summary of the fit object that is provided by INLA is given by:

```
Fixed effects:
           mean    sd 0.025quant 0.5quant 0.975quant mode kld
Intercept 1.411 0.357       0.69    1.418      2.098 1.43   0

Model hyperparameters:
                mean    sd 0.025quant 0.5quant 0.975quant mode
Range for field 1.44 0.119       1.22     1.43       1.69 1.42
Stdev for field 1.52 0.093       1.35     1.51       1.71 1.51
```

`Intercept` refers to the intercept β_0. `Range` and `Stdev` are the hyperparameters for the Gaussian random field (r and σ respectively).

This is perhaps easier to visualise and discuss alongside the posterior density plots:

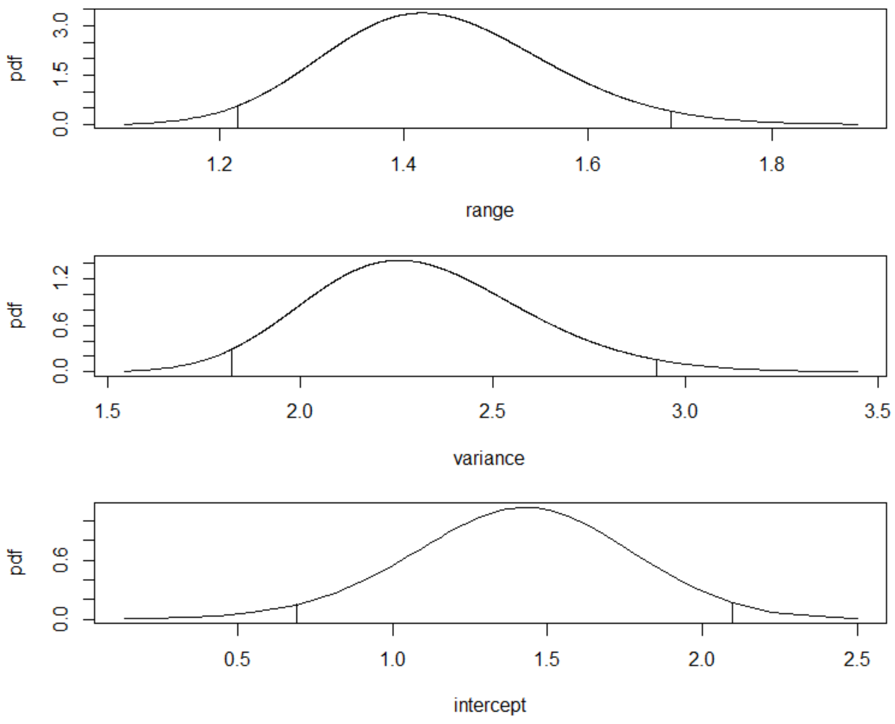


Figure 4.2: Density plots for our posterior range, variance, and intercept

Note that here we have a “variance” parameter, which is calculated as the square of the standard deviation - i.e σ^2.

This output indicates that the expected value for β_0 is 1.411 with a 95% credible interval of (0.69, 2.098). As a comparison, if we run the model with no Gaussian random field term – i.e. our model looks like $\Lambda(\mathbf{s}) = \exp\{\beta_0\}$ – we get an intercept term of $\beta_0 = 2.792$. This indicates that our Gaussian random field is a useful component to have in our model - there is evidence of non-homogeneity in our data.

This is also apparent due to the size of the variance parameter, which has a 95% credible interval of (1.235, 1.71). Since zero is not contained in the interval, this indicates that we have a significant amount of evidence for non-homogeneity i.e. local effects or clustering.

Our expected value for the range of our Gaussian random field is 1.44 with a credible interval of (1.22, 1.69). It is easier to understand what this means by plotting the covariance function, as well as the correlation.

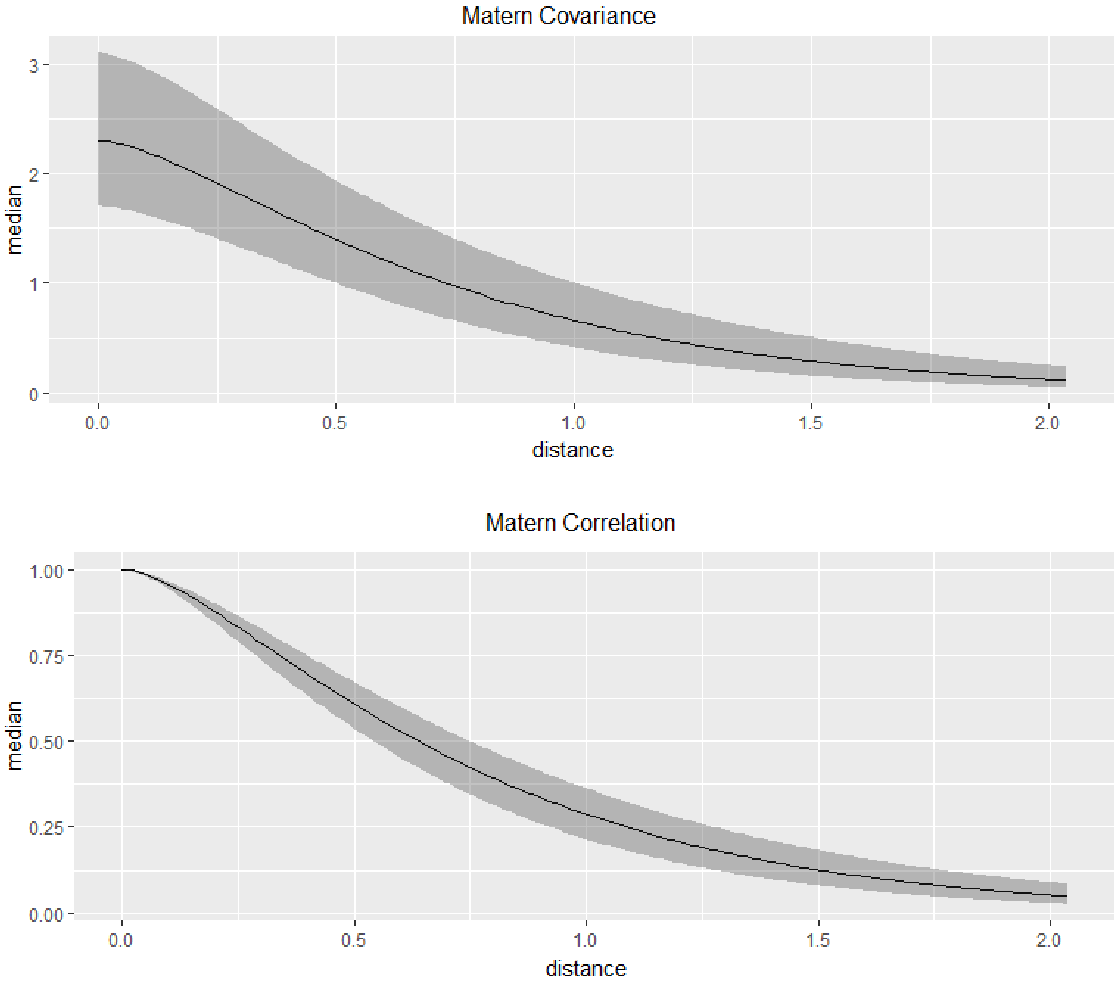


Figure 4.3: Correlation and covariance of the Gaussian random field

We can see on the vertical axis that our credible interval for the variance is highlighted in dark grey on the covariance function. We recall that the range is the value at which the distance at which spatial autocorrelation is small - we can see our mean range value of 1.44 corresponds to the correlation plot above. This means the existence of one terrorism event has very little influence on the existence of any other terrorism events which are more than 1.44 degrees of latitude/longitude (or 69 miles) away.

We can plot the mean Gaussian field as predicted by the model. Note that the following plot is on the log scale. As expected, the regions of higher density correspond to a higher expected value of the random field. In particular, areas alongside the river and by the capital city Baghdad which can be seen in the geographical map of Iraq in Figure 3.1.

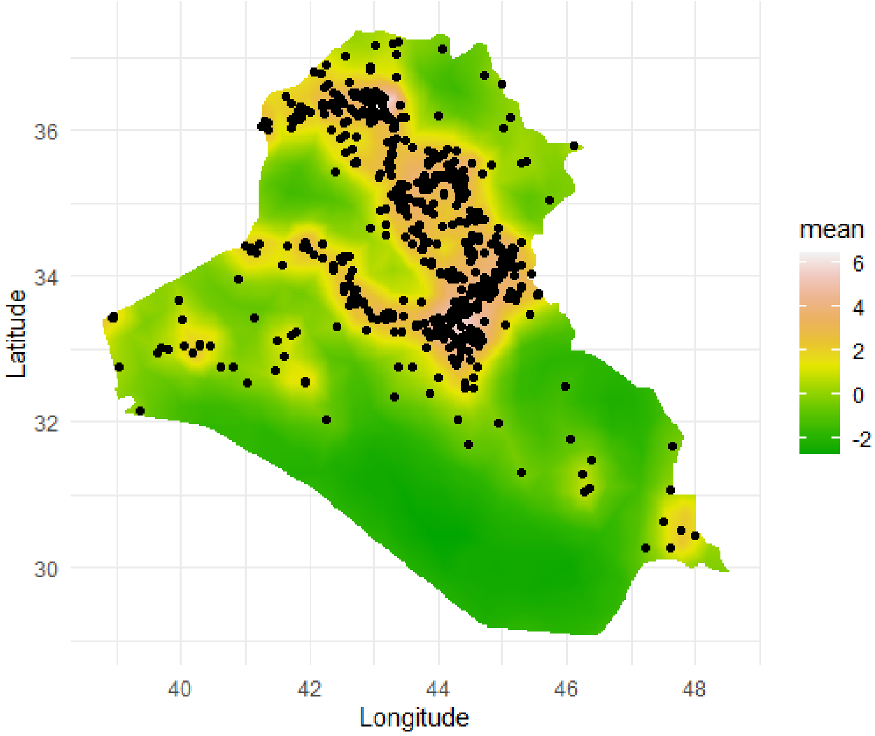


Figure 4.4: Estimated mean of the Gaussian Random Field

Overlapping data points

Now, one assumption of a Poisson point process is that no two points can occur in the same location. This is not the case for our dataset. In fact, the 2,543 points in this dataset occur in only 689 unique locations. We can see the distribution of points below:

```
Number of events        1   2   3   4   5   6   7   8   9  10  11  13  14
Number of locations   484  86  25  21  16   5   8   7   6   4   2   1   1

 15  16  17  18  19  20  22  23  25  29  30  31  32  34  37 395 496
  4   1   1   1   1   1   1   1   1   2   2   1   1   1   1   1   1
```

Figure 4.5: Table of the number of points versus the number of locations

This essentially says that there is one location that has 496 events occurring there and there is one location associated with 395 events. On the other hand, there are 484 locations with a unique event associated with them.

Because INLA discretises the space, the overlapping data points do not make a difference to the fitted model. To test this, a similar model to the one discussed above is created, which utilises the `jitter()` function in R. This adds a small amount of noise to the longitude and latitude of our points. For example, we can increase both the latitude and longitude of each data point by a randomly generated value from a Normal(0, 0.01) distribution. Running this code did not change the estimates for the mean value of our hyperparameters to within three decimal places. It makes sense to utilise INLA's inherent functionality to deal with overlapping points from here on out.

Matérn covariance and the model

There are a few issues with this fit. The fact that our point distribution is bimodal may be a problem. Consider the country of Iraq split into areas of 1×1 degree latitude/longitude. Then in any given area, there is likely to be either many points or very few points; the following histogram demonstrates this.

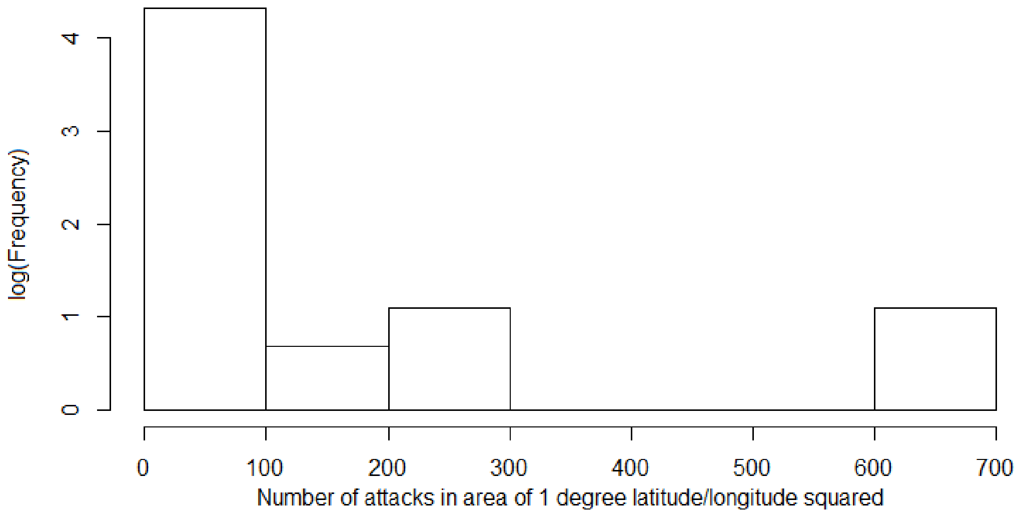


Figure 4.6: Histogram of the number of terrorist attacks in 2017 by longitude and latitude; the frequency is on a log scale.

We can see that the distribution does in fact have two modes. It is possible that type of data is not well suited to be modelled by a Gaussian random field. This indicates that a covariate would be a good idea to incorporate into the model. This would explain away much of the rapid changes in attack intensity.

4.1.2 Population covariate

We are interested in including covariates to explain some of the variation in our spatial data. Recall from section 3.1.1 that we had a high population density in the north and east of the country. In fact, our population map in Figure 3.2 has similar features to that of our Gaussian field in Figure 4.4.

Using our population dataset as a covariate, we wish to create a new model with formula as follows:

$$\Lambda(\mathbf{s}) = \exp\{\beta_0 + \beta_1 \cdot \mathit{Population}(\mathbf{s}) + G(\mathbf{s}) + \varepsilon\}$$

We are interested in estimating β_0 as well as GRF parameters r and σ as in the previous chapter. We also wish to estimate β_1, which is the coefficient for population, a spatial covariate.

Here, our population is measured in thousands of people per square kilometre. We then scale this using the `scale()` command in R - this subtracts the mean and divides by the standard deviation to standardise the data. This was done in order to aid our inference - make our estimates converge faster and increase the stability. It is not important to be able to interpret the coefficients, only to check whether 0 is in the credible interval, so there is no harm in scaling the data in this way.

An excerpt from the summary of the fit object is as follows:

```
           mean    sd 0.025quant 0.5quant 0.975quant  mode kld
pop       0.043 0.004      0.035    0.043      0.052 0.043   0
Intercept 0.577 0.027      0.523    0.577      0.630 0.577   0

Model hyperparameters:
                 mean    sd 0.025quant 0.5quant 0.975quant mode
Range for field 2.797 0.447      2.047    2.751      3.801 2.65
Stdev for field 0.446 0.044      0.366    0.444      0.538 0.44
```

Note that `pop` refers to the coefficient for the population covariate β_1. As before, `Intercept` refers to β_0 and `Range` and `Stdev` are the hyperparameters for the Gaussian random field. These relate to the smoothness of the field.

β_1 has a mean value of 0.043 with a 95% credible interval of (0.027, 0.630). We can see that the expected value is over 10 standard deviations away from zero, indicating that this covariate is in fact significant, and it was useful incorporating into our model - it explains a significant amount of the spatial variation in the data. We can see the posterior distribution of β_1 below:

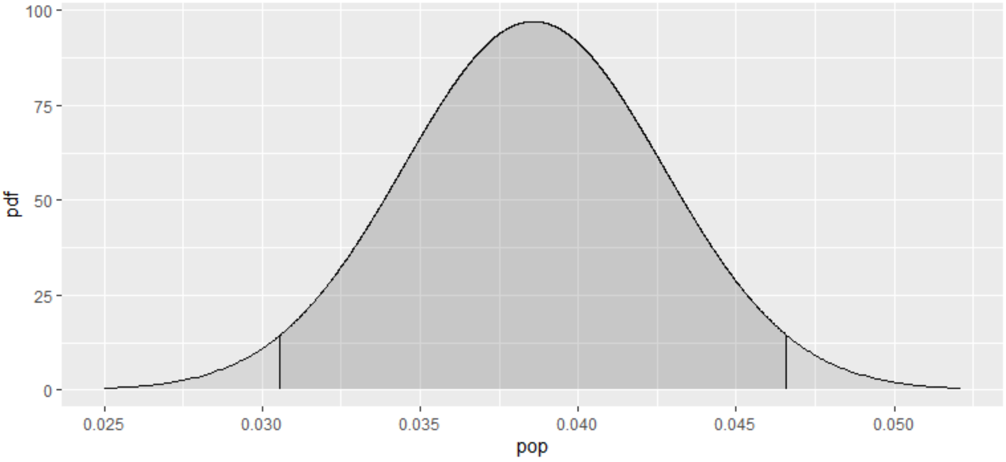


Figure 4.7: Posterior distribution for scaled population covariate

Our mean range and standard deviation, as parameters of our Gaussian random field, are 2.797 and 0.446, with standard deviations of 0.447 and 0.044 respectively. Compared to our model which does not include the covariate, we have a much higher range and much lower standard deviation. This is visible when we plot the mean Gaussian field.

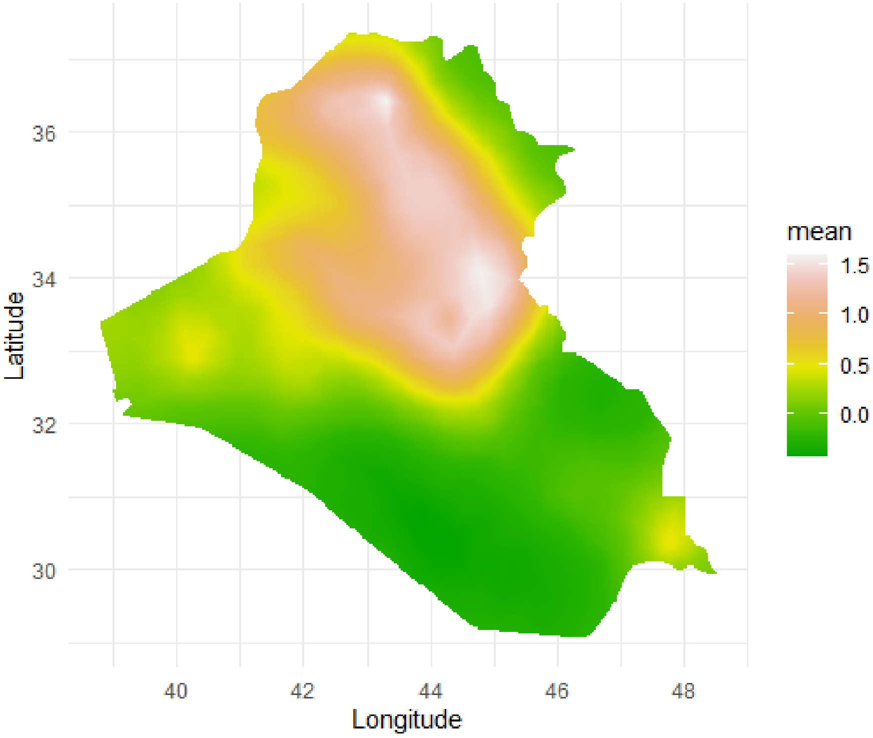


Figure 4.8: Estimated mean of the Gaussian random field (on a log scale) for a fit with population covariate

We can see that the random field is more uniform, which we would expect since the covariate has explained away a lot of the spatial variation. This is shown in the range being bigger - now observations are correlated with other observations between 2.047 and 3.801 degrees of longitude/latitude away, compared to 1.44 for the previous model.

4.1.3 Both covariates

Recall from 3.1.2that we had a dataset containing all the primary roads. If we compare the road layout in Figure 3.4 and our Figure above, we can see that there is likely to be a correlation between the two.

Let's create a covariate which we call "distance to road". This covariate takes the distance from each point to the nearest primary road (see Figure 3.5) - call this x. Our distance to road covariate is given by $|x - m|$ where m is the maximum distance to road in degrees. We then scale the data by subtracting the mean and dividing by the standard deviation.

Our new model looks like:

$$\Lambda(\mathbf{s}) = \exp\{\beta_0 + \beta_1 \cdot \mathit{Road}(\mathbf{s}) + \beta_2 \cdot \mathit{Population}(\mathbf{s}) + G(\mathbf{s}) + \varepsilon\}$$

As in the previous chapter, we are interested in estimating coefficients for our spatial covariates (β_1, β_2), the intercept term (β_0) and the parameters of the Gaussian random field (r and σ).

An excerpt from the summary of the fit object is below. Note that `pop` refers to our population covariate (i.e. this is β_2) and `distance` refers to our distance-to-road covariate (i.e. this is β_1).

```
           mean    sd 0.025quant 0.5quant 0.975quant  mode kld
pop       0.036 0.004      0.027    0.036      0.044 0.036   0
distance  0.531 0.029      0.473    0.531      0.588 0.531   0
Intercept 0.495 0.027      0.442    0.495      0.549 0.495   0

                 mean    sd 0.025quant 0.5quant 0.975quant  mode
Range for field 3.399 0.618      2.378    3.331      4.797 3.191
Stdev for field 0.375 0.042      0.299    0.373      0.465 0.369
```

Recall that our units are essentially dimensionless - although β_1 and β_2 both appear small, this is due to rescaling; neither of their credible intervals contain zero. In fact, β_2 is 9 standard deviations away from zero and β_1 is 18 standard deviations away. We can plot their posterior distributions:

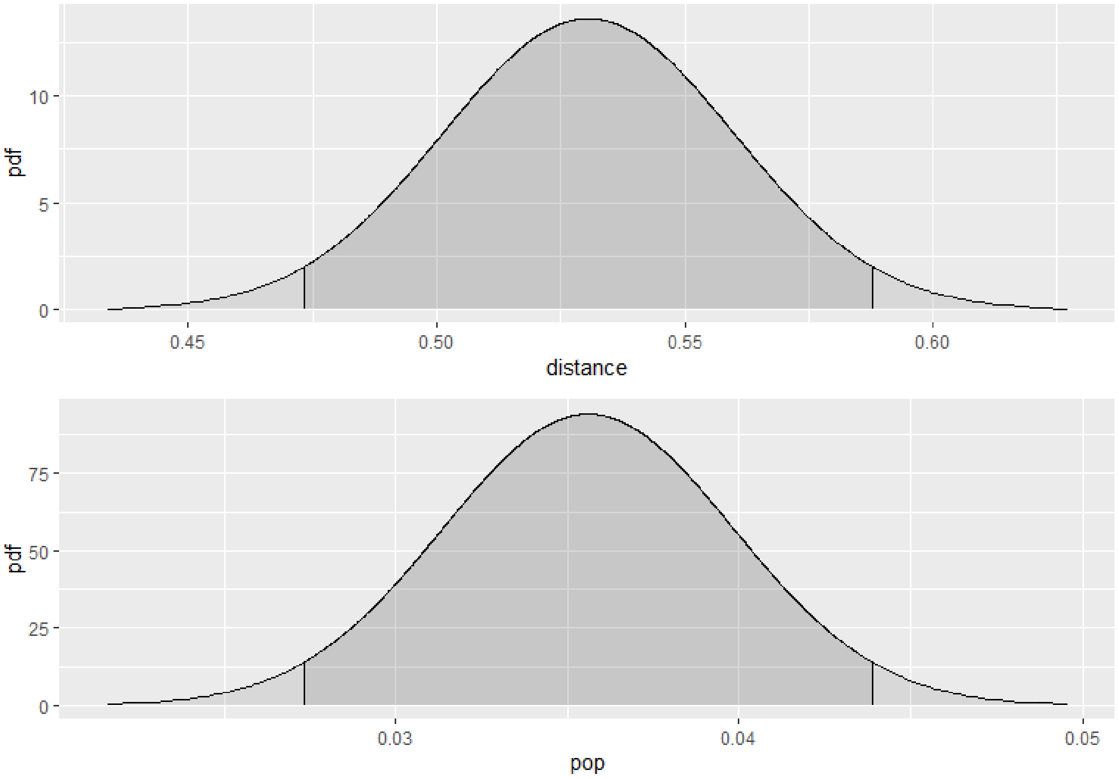


Figure 4.9: Posterior densities for population and distance to road covariates

This means that both covariates are significant. If there were to be a change in population at a point (or, alternatively, a new primary road built near a point), we may expect to see a higher rate of terrorism. Recall, however, that we cannot make statements about causation, only correlation.

In terms of hyperparameters, the range has an expected value of 3.399 and the standard deviation has an expected value of 0.375. This is a higher range and smaller standard deviation than for the one-covariate model, i.e. we have a “flatter” random field. This again is evidence that our covariates are explaining much of the clustering behaviour of points.

The mean posterior Gaussian random field is plotted in Figure 4.10 below.

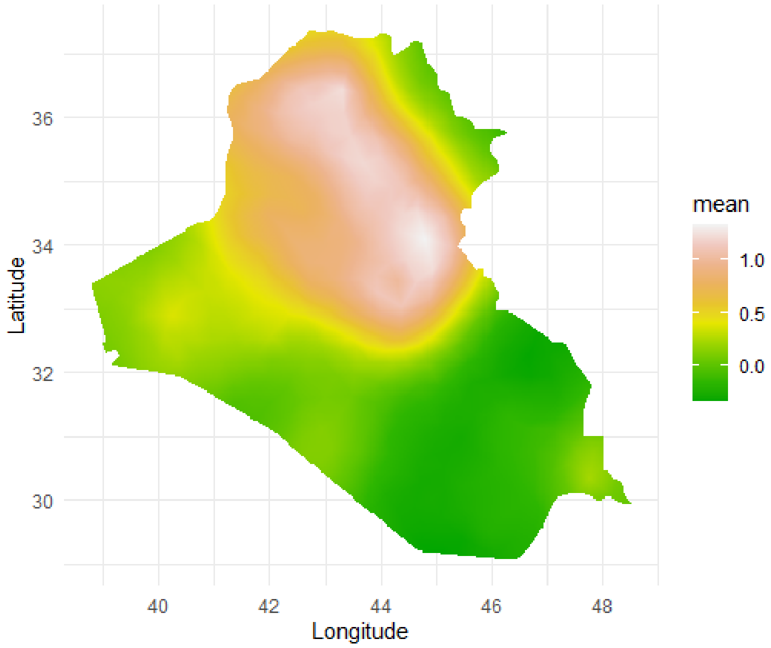


Figure 4.10: Expected random field (on a log scale) for our covariate model - i.e. the predicted field after the effect of population and primary roads have been removed

4.2 A Spatio-Temporal Model

4.2.1 Year-on-year - AR(1)

We are interested in fitting an AR(1) model to the entire dataset. Consider a model

$$\Lambda_i(\mathbf{s}) = \exp\{\beta_0 + \beta_1 \cdot Road(\mathbf{s}) + \beta_2 * Population(\mathbf{s}) + G_i(\mathbf{s}) + \varepsilon\}$$

Where $G_i(\mathbf{s}) = \rho \cdot G_{i-1}(\mathbf{s}) + \varepsilon_i'$ and i is the number of years since 2008.

One assumption of this model is that our road data and population data are constant. Historic information about roads in Iraq is not easily accessible, and we have seen from section 3.1.1 that there is minimal change in population over time. This assumption is made in the interest of simplicity.

If we run the INLA code, we get the following as an excerpt from the summary of the fit object:

```
Fixed effects:
           mean    sd 0.025quant 0.5quant 0.975quant  mode kld
distance  0.579 0.011      0.557    0.579      0.601 0.579   0
pop       0.029 0.003      0.024    0.029      0.034 0.029   0
Intercept 0.554 0.010      0.534    0.554      0.574 0.554   0

Model hyperparameters:
                     mean    sd 0.025quant 0.5quant 0.975quant  mode
```

```
Range for field     2.611 0.295      2.105     2.584       3.259 2.519
Stdev for field     0.356 0.023      0.309     0.357       0.398 0.363
GroupRho for field 0.365 0.277     -0.278     0.411       0.780 0.527
```

It appears that our ρ term has a mean value of 0.365, a credible interval of (-0.278, 0.780), and a mode of 0.527. It is easier to understand this with a graph since it shows the left-skewed shape of the posterior distribution.

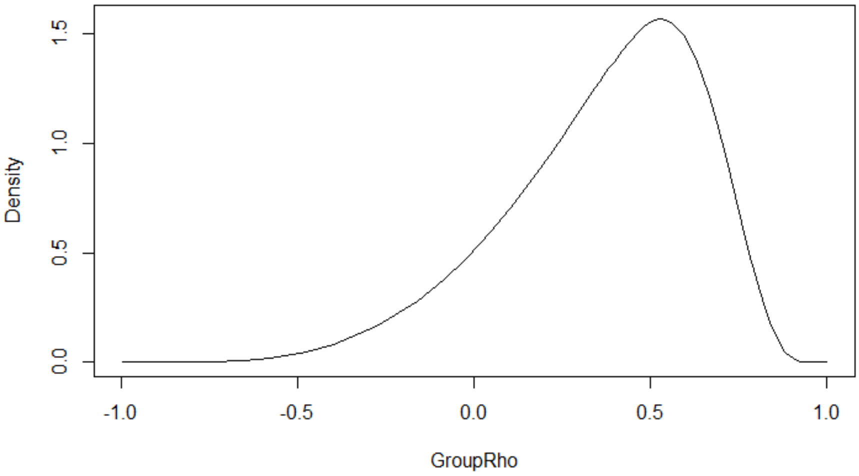


Figure 4.11: Posterior density of AR(1) parameter ρ for spatio-temporal model

The credible interval does contain zero, which is around one and a half standard deviations away from the mean, which means that our ρ parameter is not significant. This indicates that this is due to the strong temporal dependence of the covariates - the Gaussian random field (which models the residuals after the effect of the covariates is taken out) is only weakly correlated over time. We can say that, year-on-year, the structure that is not explained by the population and road covariates changes over time.

4.2.2 Month-on-month - AR(1) Model

Seeing how the Gaussian random field changes over time is also of interest. Rather than fitting twelve independent models, one for each month of 2017, an autoregressive model can instead be used.

Here, the model looks like:

$$\Lambda_i(\mathbf{s}) = \exp\{\beta_0 + \beta_1 \cdot \mathit{Road} + \beta_2 * \mathit{Population} + G_i(\mathbf{s}) + \varepsilon\}$$

Where $G_i(\mathbf{s}) = \rho \cdot G_{i-1}(\mathbf{s}) + \varepsilon_i'$ and G_i is the Gaussian random field for month i

INLA has the capability to incorporate grouped models such as the AR(1) model; the output of fitting this model to our data looks like:

```
Fixed effects:
           mean    sd 0.025quant 0.5quant 0.975quant  mode kld
distance  0.539 0.029      0.482    0.539      0.596 0.539   0
pop       0.036 0.004      0.028    0.036      0.044 0.036   0
Intercept 0.500 0.027      0.446    0.500      0.554 0.500   0

Model hyperparameters:
                    mean    sd 0.025quant 0.5quant 0.975quant  mode
Range for field    3.903 0.665      2.780    3.838      5.383 3.707
Stdev for field    0.263 0.021      0.223    0.262      0.306 0.261
GroupRho for field 0.100 0.435     -0.764    0.144      0.802 0.382
```

Here, `GroupRho` indicates the estimate of our ρ parameter; the other fixed effects and hyperparameters have the same meanings as in the previous model.

We can see that our estimates for distance from road, population, and intercept all look very similar to the model without the temporal component. In fact, the predicted means of the temporal model are all well within half of one standard deviation of the spatial model, which is what we would expect to see.

Recall that the mean range and standard deviation of the GRF for the spatial model were 3.399 and 0.375 respectively. In our spatio-temporal model, we have a slightly higher mean range (around one standard deviation away from the spatial model's estimate) and a much lower marginal standard deviation (lower than the 2.5% quantile for the spatial model). This relates to the first month - January - indicating that our Gaussian random field in this month has a much lower peak intensity compared to the peak intensity in the whole of 2017. This is expected, due to the fact we have much fewer points. Our range being similar means that the existence of one terrorism event influences the intensity of the surrounding area up to the same distance away for both January 2017, and the whole of 2017.

Our ρ is positive, though rather small, with a mean value of 0.1 and a standard deviation of 0.435. A plot of the posterior density is:

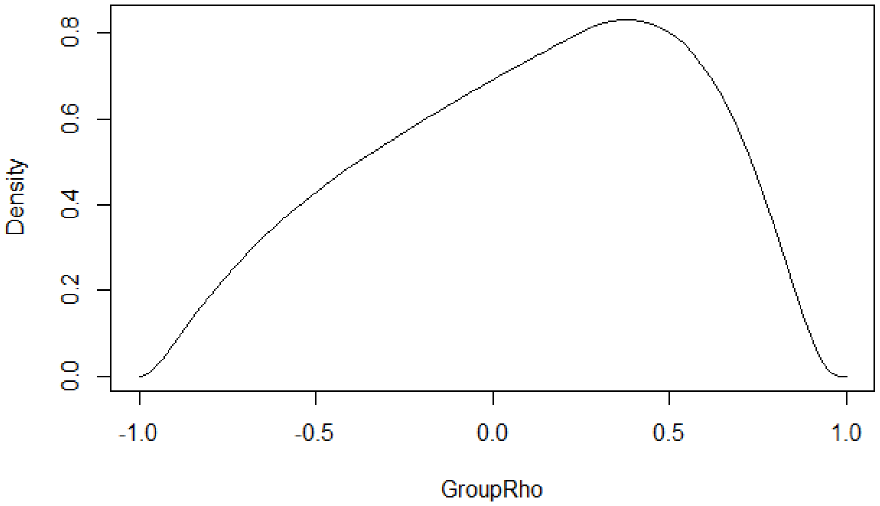


Figure 4.12: Posterior density of AR(1) parameter ρ for the spatio-temporal model, 2017

It is the case that 0 is very much in the credible interval for the parameter ρ. This indicates that the covariates we have - population and distance to the nearest road - are explaining away most of the variation, meaning that our Gaussian random field is only weakly correlated over time. It is interesting to compare this to Figure 4.11; we can see that there seems to be more temporal correlation over the years compared to the months. It is plausible that this is due to the higher number of data points in the yearly model causing the underlying field to be "spikier" - i.e. having areas of very high and very low density.

The predicted Gaussian random fields for each month can be plotted:

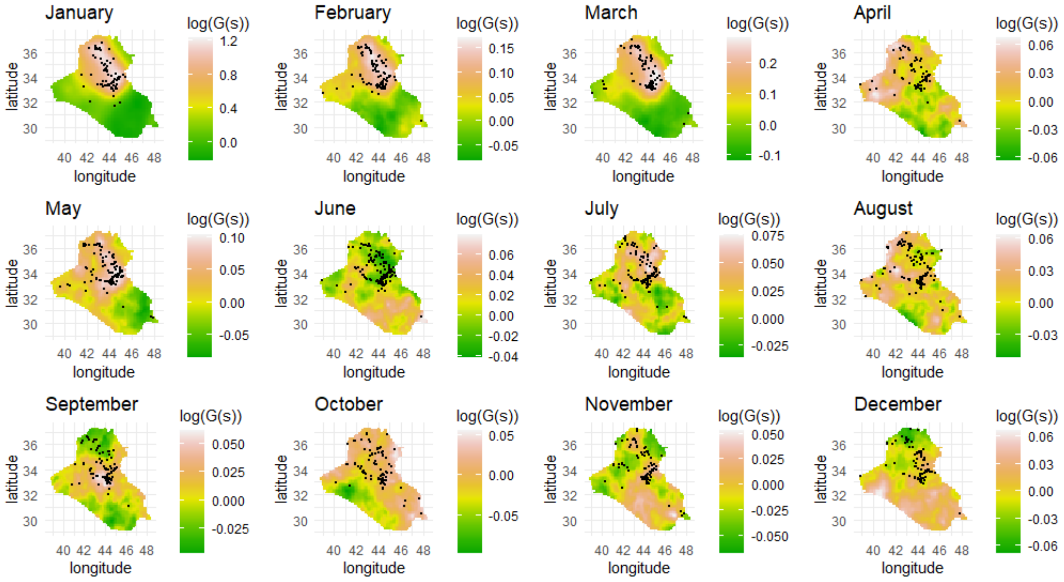


Figure 4.13: Mean predicted Gaussian random field for each month in 2017

Looking at the Gaussian random fields as predicted by the AR(1) model, it is apparent that, although we can see quite a strong pattern in the first few months, this tapers down to a pretty much uniform field by the end of the year. This is due to our predicted value of ρ being close to zero.

June is an important year in terms of terrorism, being the year that the US began targeting the Islamic State terrorist group [39]. It is not unreasonable to suggest that the spatial structure changes in this month, although it is perhaps not obvious from Figure 4.13. This is discussed further in section 5.1.4.

We have looked at autoregressive models for the temporal component. In the following section, the different attack types will be considered as a mark.

4.3 A Marked Model

4.3.1 Exploring data types

Recall that in our dataset, we have a mark that gives each event an attack type. This is a factor that can take nine different values.

Let us apply the covariate model discussed in the previous section to each of the marks separately. Here, we are just looking at the data from 2017. This is to be used as an exploration primarily, so the values of the parameters are not important; instead, only the random fields will be plotted. Recall that the effect of the covariates (and the intercept) is not included in the random field, which is why the fields do not appear to match the points exactly.

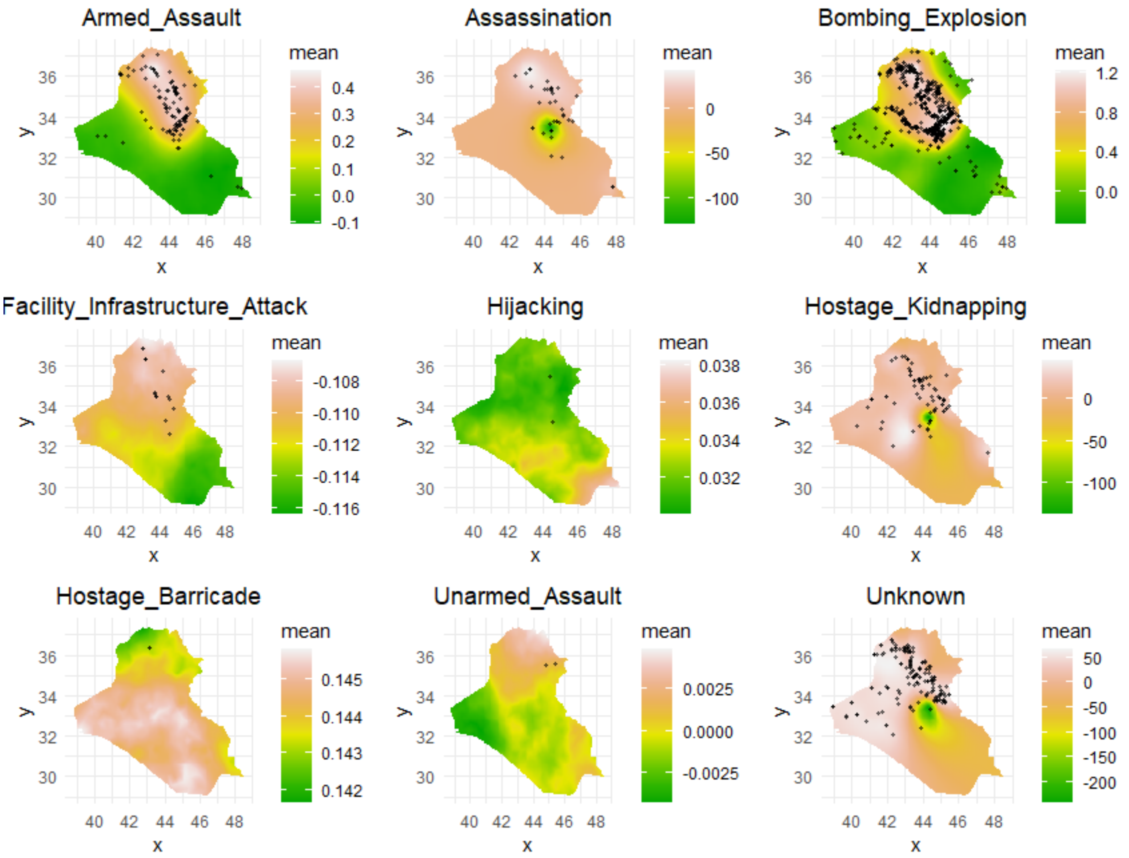


Figure 4.14: Expected random field for several marks, 2017

Now, it is quite hard to see the number of points in each plot due to overprinting, however, we already plotted this in Figure 3.10; there are significantly more points in the bombing category than any other category. The plots where the fields look strange or patchy are mostly just those which have very few points.

Now, one point that is important to consider is whether it is fair to be combining all these terrorism events together in one model, as we have been. After all, there is no reason to believe that just because there is a high density of, say, bombing in one area, it does not mean there is a high density of, say, armed assault. Now, since bombing is the most prevalent attack type, it may be possible that due to the sheer quantity of bombing points, this mark unduly influences the model.

This indicates that it might be interesting to do further analysis on whether there is a different random field underlying the bombing events and all the other types of events.

4.3.2 Two models - bombing versus all other attack types

In order to justify putting our bombing events in with the other marks, we should fit a model to both the "bombing" data points, and the "other" datapoints, and note the differences. We can fit two independent models in INLA, and compare them.

Fitting the model

$$\Lambda(\mathbf{s}) = \exp\{\beta_0 + \beta_1 \cdot Road(\mathbf{s}) + \beta_2 * Population(\mathbf{s}) + G(\mathbf{s}) + \varepsilon\}$$

to just the data tagged with the "Bombing/Explosion" label gives us, as an excerpt from the summary of the fit object:

```
Fixed effects:
             mean    sd 0.025quant 0.5quant 0.975quant  mode kld
pop         0.036 0.004      0.027    0.036      0.045 0.036   0
distance    0.498 0.032      0.435    0.498      0.561 0.498   0
Intercept   0.463 0.030      0.403    0.463      0.523 0.463   0

Model hyperparameters:
                 mean    sd 0.025quant 0.5quant 0.975quant  mode
Range for field 3.134 0.569      2.196    3.071      4.426 2.939
Stdev for field 0.368 0.041      0.293    0.366      0.456 0.361
```

Doing the same for all the rest of the data gives us:

```
Fixed effects:
             mean    sd 0.025quant 0.5quant 0.975quant  mode kld
pop         0.068 0.010      0.048    0.068      0.087 0.068   0
distance    0.367 0.043      0.283    0.367      0.451 0.367   0
Intercept   0.327 0.043      0.244    0.327      0.411 0.327   0

Model hyperparameters:
                 mean    sd 0.025quant 0.5quant 0.975quant  mode
Range for field 3.597 0.708      2.445    3.514      5.216 3.344
Stdev for field 0.322 0.041      0.248    0.319      0.409 0.315
```

We can see that the range is very similar for both models - the posterior mean of the range for the bombing model is within one standard deviation away from the posterior mean of the other model. Also, the posterior mean of the marginal standard deviation of the bombing model is slightly more than one standard deviation away from the posterior mean of the model for the other attack types. There is not a significant difference between the hyperparameters for the two models. As a visualisation of this, we can plot the posterior distributions:

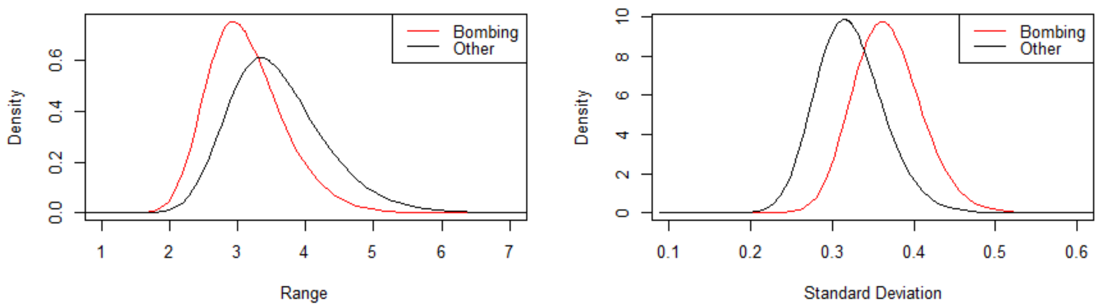


Figure 4.15: Posterior distributions for the hyperparameters of our GRF, bombing attack type versus all other attack types

Now, on the other hand, there is a difference in the parameters - our population and distance to road coefficients - for both datasets. While the bombing model has a credible interval of $(0.027, 0.045)$ for β_2, the other attack type model has a credible interval of $(0.048, 0.087)$ - they do not overlap. This indicates that bombing or explosion type terrorist attacks are significantly less correlated with population than other attack types.

Similarly, bombing type attacks are more highly correlated with the distance to road than other attacks with a credible interval for β_1 of $(0.435, 0.561)$ compared to $(0.283, 0.451)$. The posterior means for β_1 are around three standard deviations apart from each other. We can say that bombing attacks are driven more by road than population, compared to all other attack types. To visualise this difference, the posterior densities can be plotted.

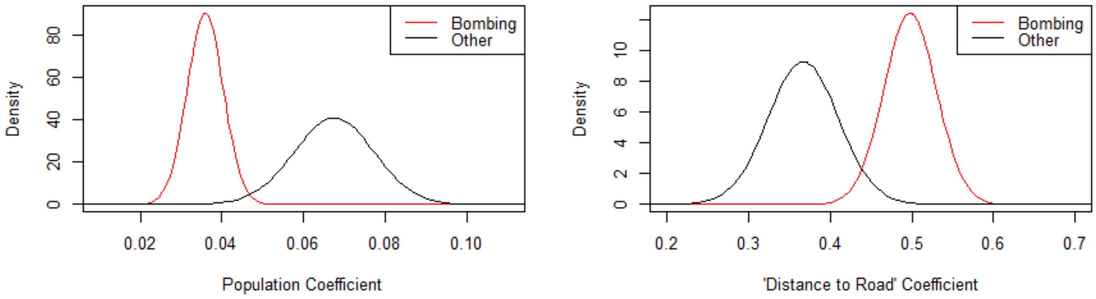


Figure 4.16: Posterior distributions for the parameters of our model, bombing attack type versus all other attack types

Here, it is very apparent that the credible intervals are significantly different. We can also plot the mean predicted Gaussian random fields:

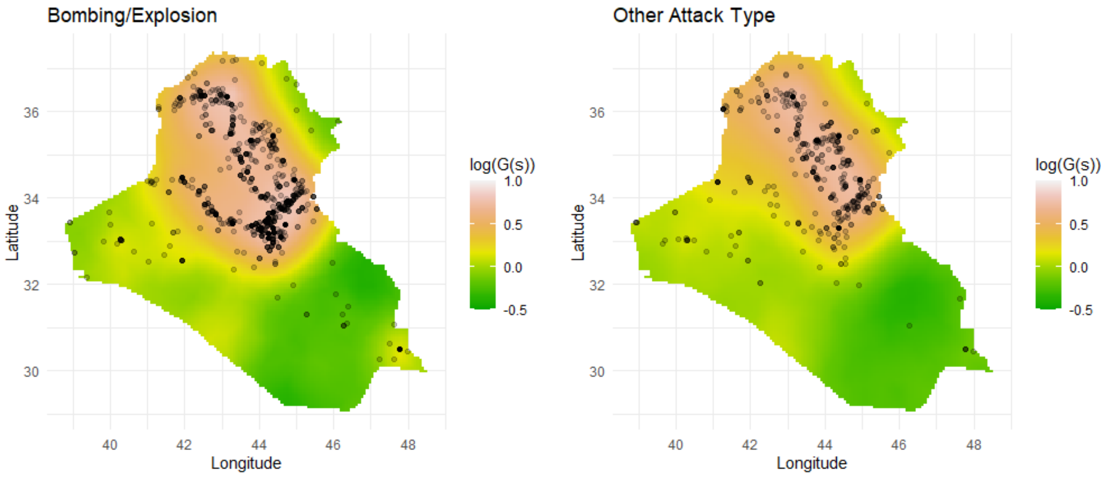


Figure 4.17: Expected random field for bombing versus all other attack types, 2017

The patterns have similar areas of high and low density. It is hard to visually make out the differences in the parameters β_0 and β_1 but the difference in values for the range are perhaps more apparent.

Several log-Gaussian Cox models have now been fitted to the data. In the following chapter, an overview will be given of the work this far, before a discussion of some limitations of the project as well as future work that could be done.

Chapter 5

Discussion

In chapter 1, an introduction to terrorism in Iraq was given, alongside an overview of the aims of this project. Then, in chapters 2 and 3, the data to be used was discussed, as well as the model to be fit and the methods for doing so. Finally, in chapter 4, the models were fitted to the data. Here, a brief overview of the results is given.

First, a spatial model was fitted to the terrorism data from 2017. Evidence of clustering behaviour was found; i.e. there was evidence that the latent field is non-homogeneous. The next terrorism event is likely to occur in a location where there have been previous terrorism events. It was found by looking at the hyperparameters of our model that the correlation between the underlying intensity of terrorism between locations in space more than 1.44 units of longitude/latitude away is negligible. The results of this model indicated that including a covariate would be informative due to the bimodal distribution of data points. In this section, the issue of overlapping data points was also addressed.

Next, a model with a population covariate was introduced. It was determined that the spatial distribution of terrorist attacks is significantly correlated with population; the predicted latent field has a higher mean intensity in areas that have a dense population. This is unsurprising - if a terrorist aims to affect as many people as possible, naturally, the intensity of attacks is denser in areas with a high population. Due to the population covariate explaining a significant amount of the spatial variation, the Gaussian random field - which, recall, models the residual field after the effect of the covariates are removed - was found to be flatter (i.e. have higher range and lower standard deviation) than the model with no covariates.

A model with two covariates - the distance to the nearest primary road as well as population - was then fitted. It was found that both covariates were highly significant; the underlying intensity of terrorism is high in areas with a small distance to the nearest primary road and a high population. This indicates that the distance-to-road measure does not act as a proxy for population, as we had suggested could be the case in section 3.1.2. This indicates that, when discussing where to allocate anti-terrorism resources, it is not only areas of dense population that need to be considered, but also primary roads which may not be densely populated.

A spatio-temporal model with an autoregressive(1) component modelling the year-on-year correlation was then fit to the years 2007-2018. Again, the population and distance-to-road parameters were highly significant, as expected. The parameter ρ, which gives the temporal correlation after the effect of the covariates have been removed, was not found to be significantly greater than 0. This indicates that the Gaussian random fields which soak up the residual correlation are not strongly correlated over time. Although the credible interval did not contain 0, the estimate indicated a positive expected value for ρ, indicating a weak temporal persistence. A similar model, this time with the autoregressive temporal component used to model the month-on-month correlation, was also fit to the data from 2017. The ρ parameter was also not found to be significantly greater than 0.

Finally, a marked model was fitted to the data. It was found that bombing type attacks are significantly less related to population density and more related to the distance to the nearest primary road, compared to all other terrorist attack types. A plausible explanation for this is the use of sticky bombs, which come under this classification of attack [7], which were "the assassination tool of choice in Iraq" [40] and have seen widespread use in Iraq since 2004 [41]. These are a type of explosive device which is attached to motor vehicles [42]. The prevalence of these vehicle-related bombs could well be connected to this difference.

5.1 Further work and limitations

5.1.1 Covariance Functions

For this project, the Matérn covariance function (see section 2.3.3) has been used to describe the random field. Since the measure used is the Euclidean distance between points, the Matérn covariance is isotropic, i.e. it assumes that the covariance between two points separated by d units longitudinally is the same as two points separated by d units laterally. From Figure 3.11, it is apparent that, although there appears to be little visual difference between the clustering of points in the north/south as opposed to the east/west, there does appear to be bands of terrorist attacks working diagonally, from the north-west down to the south-east. We can hypothesise that this is related to the direction of the river systems in Iraq (see Figure 3.1). This indicates that it might be interesting to use a non-isotropic covariance function in order to distinguish the covariance between points separated in the north-west to south-east direction and points separated in the north-east to south-west direction.

5.1.2 Possible Other Covariates

As discussed in section 4.2, there was only weak evidence of a temporal component in the form of an auto-regressive model being useful for our log-Gaussian Cox model. However, looking at Figure 4.10 alludes to the fact that there does seem to be some pattern in the Gaussian random field that has not been accounted for in our spatial model. This indicates that there may well be some covariate that could have been included in the model. As discussed above - and in fact, this is apparent if we compare the map of river systems in Iraq in Figure 3.1 and our Gaussian random field in Figure 4.10 - it seems that there is likely a correlation between the distance to a river and the underlying Gaussian random field. The assumption was made initially that this pattern would be able to be explained by population; however, this does not appear to be the case, plausibly due to the incredibly high relative density of people living in Baghdad. A “distance to the nearest river” covariate may well be a useful additional component to the model.

Another alternative covariate would be another measure of road density. In Figure 3.4, it is apparent that this also matches the same shape of our Gaussian random fields. The way road density is measured in this project - finding the distance to the nearest primary road - leaves out some of the data. In particular, a spot on a primary road in the middle of the desert with no other roads in the vicinity is given the same associated value as a point in the centre of Baghdad. Perhaps finding another way to measure the density of roads would be fruitful. A suggestion would be to count the number of roads (of any type) in a certain radius of the location.

5.1.3 Other Marked Models

In section 4.3, marked models were analysed. The most feasible method was to fit two independent models to the data - one for bombing type attacks and one for all other attacks. Although this method lent some powerful results, for further research, a joint model could be fitted to the data. This would allow the similarities between the Gaussian random fields to be quantified and allow for better comparison between the two datasets. A model with a binomial likelihood function for the mark would be applicable.

5.1.4 Monthly Model

Recall that a model with a temporal AR(1) component was fit to our 2017 dataset in section 4.2. We can see the fitted field for each month in Figure 4.13. Now, the parameter of the autoregressive model was not found to be significant. Perhaps this was not the best-suited model for the data. The US targeted the Islamic State terrorist group in June specifically [39]. It may be more applicable to split the data into two groups - before and after this occurred and fit models to each group independently to quantify the difference that US intervention made on the spatio-temporal distribution of terrorist attacks in Iraq.

5.1.5 Spatial Accuracy of the Data

The existence of overlapping data points has been discussed in section 4.1.1. However, the importance of collecting data that is as accurate as possible is yet to be seen. Essentially, the impact on the modelling of the data not being entirely accurate is unknown. According to the GTD, it is the location of the city that the terrorist event occurs in that is usually recorded rather than the location of the attack itself. Another database variable - geocoding specificity - which is noted alongside the information for each attack, which gives the geospatial resolution of the latitude/longitude fields, is also accessible [7]. This means that the recorded latitude/longitude has some unknown margin of error. Due to this analysis being done using the INLA framework, overlapping data points can be dealt with; however, in doing so, it is assumed that they are accurate to within a certain distance, which may not be the case. Since we are primarily interested in the trends of terrorism on a large scale (i.e. the entire country of Iraq), this is not much of a limitation for this project. However, if future work wanted to analyse terrorism trends on a smaller scale - perhaps only considering events within Baghdad - this would be needed to be considered. Similarly, methods other than INLA may have an issue with the non-overlapping assumption of the Poisson process.

5.2 Conclusion

This project aimed to analyse spatio-temporal terrorism trends in Iraq. Several log-Gaussian Cox models have successfully been fit to data from the Global Terrorism Database using INLA software in R on high-performance computing system NeSI. The spatial model fitting data from 2017 gave strong evidence that both population density and the distance to the nearest primary road are highly correlated with the intensity of terrorism. A model with an autoregressive temporal component fit to the data from 2007 to 2018 inclusive found a weak temporal year-on-year correlation. Finally, two models fit to bombing-type attacks, and all other attack types in 2017 found that bombing attacks are significantly more correlated with road distance and less correlated with population compared to other attack types.

Bibliography

[1] C. A. Hannah Ritchie, Joe Hasell and M. Roser, “Terrorism,” *Our World in Data*, 2013. https://ourworldindata.org/terrorism.

[2] G. LaFree and L. Dugan, “Research on terrorism and countering terrorism,” *Crime and Justice*, vol. 38, no. 1, p. 413–477, 2009. https://www.jstor.org/stable/10.1086/599201.

[3] Walden University, “Importance of national crime statistics.” https://www.waldenu.edu/online-bachelors-programs/bs-in-criminal-justice/resource/why-national-crime-statistics-are-important, Mar 2021.

[4] New Zealand Security Intelligence Service, “Counter terrorism.” https://www.nzsis.govt.nz/our-work/counter-terrorism/.

[5] H. M. Karami, “The political and social roots of terrorism in iraq (2003-2017),” *Researchers World : Journal of Arts, Science and Commerce*, vol. VIII, no. 3(1), p. 32–37, 2017.

[6] Institute for Economics & Peace, “Global terrorism index 2020: Measuring the impact of terrorism.” https://www.visionofhumanity.org/global-terrorism-index-2020-summary-and-key-findings/, 2020.

[7] Global Terrorism Database, “About the global terrorism database.” https://www.start.umd.edu/gtd/about/.

[8] H. Rue, S. Martino, and N. Chopin, “Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion).,” *Journal of the Royal Statistical Society B*, vol. 71, pp. 319–392, 2009.

[9] R Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013.

[10] P. Gao, D. Guo, K. Liao, J. J. Webb, and S. L. Cutter, “Early detection of terrorism outbreaks using prospective space–time scan statistics,” *The Professional Geographer*, vol. 65, no. 4, p. 676–691, 2013.

[11] S. L. Linton, J. M. Jennings, C. A. Latkin, M. B. Gomez, and S. H. Mehta, "Application of space-time scan statistics to describe geographic and temporal clustering of visible drug activity," *Journal of Urban Health*, vol. 91, no. 5, p. 940–956, 2014.

[12] L. K. Siebeneck, R. M. Medina, I. Yamada, and G. F. Hepner, "Spatial and temporal analyses of terrorist incidents in iraq, 2004–2006," *Studies in Conflict & Terrorism*, vol. 32, no. 7, p. 591–610, 2009.

[13] M. Townsley, S. D. Johnson, and J. H. Ratcliffe, "Space time dynamics of insurgent activity in iraq," *Security Journal*, vol. 21, no. 3, p. 139–146, 2008.

[14] N. J. Clark and P. M. Dixon, "Modeling and estimation for self-exciting spatio-temporal models of terrorist activity," *The Annals of Applied Statistics*, vol. 12, no. 1, 2018.

[15] A. Python, J. B. Illian, C. M. Jones-Todd, and M. Blangiardo, "A bayesian approach to modelling subnational spatial dynamics of worldwide non-state terrorism, 2010–2016," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 182, no. 1, p. 323–344, 2018.

[16] S. Shirota and A. E. Gelfand, "Space and circular time log gaussian cox processes with application to crime event data," *The Annals of Applied Statistics*, vol. 11, no. 2, 2017.

[17] G. Mohler, "Modeling and estimation of multi-source clustering in crime and security data," *The Annals of Applied Statistics*, vol. 7, no. 3, 2013.

[18] United States Institute of Peace, "Iraq timeline: Since the 2003 war." `https://www.usip.org/iraq-timeline-2003-war`, 2020.

[19] A. Baddeley, E. Rubak, and R. Turner, *Spatial point patterns: Methodology and applications with R*. CRC Press, 2016.

[20] A. Baddeley and R. Turner, "spatstat: An R package for analyzing spatial point patterns," *Journal of Statistical Software*, vol. 12, no. 6, pp. 1–42, 2005.

[21] Stanford Department of Statistics, "Random fields." `https://statweb.stanford.edu/~jtaylo/courses/stats352/notes/random_fields.pdf`.

[22] E. T. Krainski, V. Gómez-Rubio, H. Bakka, A. Lenzi, D. Castro-Camilo, D. Simpson, F. Lindgren, and H. Rue, *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman & Hall/CRC Press, 2019.

[23] "The dmatern model." `https://inla.r-inla-download.org/r-inla.org/doc/latent/dmatern.pdf`.

[24] D. Simpson, J. B. Illian, F. Lindgren, S. H. Sørbye, and H. Rue, "Going off grid: Computationally efficient inference for log-gaussian cox processes," *Biometrika*, vol. 103,

no. 1, p. 49–70, 2016.

[25] University of Washington, "Inla for spatial statistics - grouped models." https://faculty.washington.edu/jonno/SISMIDmaterial/8-Groupedmodels.pdf.

[26] B. M. Taylor and P. J. Diggle, "Inla or mcmc? a tutorial and comparative evaluation for spatial prediction in log-gaussian cox processes," *Journal of Statistical Computation and Simulation*, vol. 84, no. 10, p. 2266–2284, 2013.

[27] J. B. Illian, S. Martino, S. H. Sørbye, J. B. Gallego-Fernández, M. Zunzunegui, M. P. Esquivias, and J. M. Travis, "Fitting complex ecological point process models with integrated nested laplace approximation," *Methods in Ecology and Evolution*, vol. 4, no. 4, p. 305–315, 2013.

[28] University of Washington, "Inla for spatial statistics - log gaussian cox processes." https://faculty.washington.edu/jonno/SISMIDmaterial/4-LGCPs.pdf.

[29] S. Martino and A. Riebler, "Integrated nested laplace approximations (inla)," *Wiley StatsRef: Statistics Reference Online*, p. 1–19, 2020.

[30] Yuan, Yuan, Bachl, F. E., Lindgren, Finn, Borchers, D. L., Illian, J. B., Buckland, S. T., Rue, Håvard, Gerrodette, and Tim, "Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales," *Ann. Appl. Stat.*, vol. 11, pp. 2270–2297, 12 2017.

[31] New Zealand eScience Infrastructure, "About us." https://www.nesi.org.nz/about-us.

[32] A. Python, A. Bender, A. K. Nandi, P. A. Hancock, R. Arambepola, J. Brandsch, and T. C. Lucas, "Predicting non-state terrorism worldwide," *Science Advances*, vol. 7, no. 31, 2021.

[33] Wikipedia, "List of places in iraq." https://en.wikipedia.org/wiki/List_of_places_in_Iraq, Jul 2021.

[34] Embassy of the Republic of Iraq Public Relations Office, "Geography." http://www.iraqiembassy.us/page/geography.

[35] World Population Review, "Iraq population 2021." https://worldpopulationreview.com/countries/iraq-population.

[36] Humanitarian Data Exchange, "Iraq - population density." https://data.humdata.org/dataset/worldpop-population-density-for-iraq.

[37] N. Tierney, "visdat: Visualising whole data frames," *JOSS*, vol. 2, no. 16, p. 355, 2017.

[38] A. South, *rnaturalearth: World Map Data from Natural Earth*, 2017. R package

version 0.1.0, https://CRAN.R-project.org/package=rnaturalearth.

[39] Wilson Center, “Timeline: The rise, spread, and fall of the islamic state.” https://www.wilsoncenter.org/article/timeline-the-rise-spread-and-fall-the-islamic-state.

[40] Stars and Stripes, “‘sticky bombs,’ like those used in iraq, now appearing in afghanistan.” https://www.stripes.com/theaters/middle_east/sticky-bombs-like-those-used-in-iraq-now-appearing-in-afghanistan-1.183623, 2012.

[41] Swarajyamag, “Explained: Why ‘sticky bombs’ could become a new headache for security forces in kashmir.” https://swarajyamag.com/defence/explained-why-sticky-bombs-could-become-a-new-headache-for-security-forces-in-kashmir.

[42] R. Johnston, J. Vetrone, and J. Warner, “Sticky bomb detection with other implications for vehicle security,” *Journal of Physical Security*, vol. 4, 01 2010.