POLITECNICO
MILANO 1863

POLITECNICO DI MILANO

SCHOOL OF INDUSTRIAL AND INFORMATION ENGINEERING
Master of Science in Mathematical Engineering

A Bayesian Poisson hurdle model applied to spatial data

Author
Jacopo RESCALDANI,
matricola: 905626

Supervisor
Prof. Federico BASSETTI

Academic Year 2021-2022

Ai miei genitori

Abstract

This thesis considers two aspects of count data that can raise in a real case scenario: the over-abundance of the zero value and the spatial reference. When there is an excess of zeros the more conventional probabilistic distributions, such as Poisson or Negative Binomial, are inefficient. The proposed alternative solution to the problem is the hurdle distribution, where all zeros are enclosed in a probability point mass at zero and a discrete distribution is truncated to account for only the strictly positive observations. Four Bayesian models that exploits Poisson hurdle distributions are introduced and tested with synthetic data.
Therefore we consider a dataset in which the burned hectares of a forest are assumed to have a Poisson hurdle distribution. Furthermore each observation is characterized by geological and meteorological measurements which act as covariates and by a spatial reference such that each observation belongs to a sub-region treated as a group. The most suitable model out of the four proposed, has been applied to the dataset. It has a regression structure in order to take into account the covariates and some group-specific random effects to explain the spatial dependence. After assigning some prior distributions to the model parameters we obtain the posterior distributions (using Markov Chain Monte Carlo methods to sample from them). Then a Bayesian inference on the parameters is performed to assert the relevance (or not) of the group structure among data and which covariates are the most significant. We have also compared the bayesian estimates of the parameters with empirical ones (computed through the MLE). At the end a prediction on new unseen data is made.

Keywords: Bayesian Approach, Markov Chain Monte Carlo Methods, Poisson Hurdle.

Sommario

Questa tesi considera due aspetti dei dati discreti che possono nascere in uno scenario reale: l'estrema abbondanza del valore zero e la dipendenza spaziale. Quando c'è un eccesso di zeri le convenzionali distribuzioni di probabilità, come Poisson o Binomiale Negativa, risultano inefficienti. La soluzione alternativa proposta è la distribuzione hurdle, in cui gli zeri sono rappresentati da un unico punto di massa di probabilità a zero e una ditribuzione discreta viene troncata per tener conto solo dei valori strettamenti positivi. Quattro modelli Bayesiani che sfruttano la distribuzione Poisson hurdle sono presentati e testati su dati artificiali.

Poi consideriamo un dataset dove gli ettari bruciati di una foresta sono ritenuti avere una distribuzione Poisson hurdle. Inoltre ogni osservazione è caratterizzata da misurazioni meteorologiche e geologiche che fungono da covariate e da un riferimento spaziale secondo cui ogni osservazione appartiene ad una sotto-regione, trattata come un gruppo. Il modello dei quattro proposti più adatto è applicato al dataset. Si tratta di quello che presenta una regressione per tener conto delle covariate e dei random effects di gruppo per spiegare il riferimento geografico. Dopo aver assegnato ai parametri del modello delle distribuzioni a priori abbiamo ottenuto le distribuzioni a posteriori (usando le catene di Markov Monte Carlo per campionare). Poi abbiamo fatto inferenza bayesiana sui parametri per asserire la significatività o meno della struttura a gruppi dei dati e quali covariate fossero le più significative. Abbiamo confrontato la stima bayesiana dei parametri con quella empirica (calcolata con MLE) e infine è stata fatta predizione su nuovi dati inosservati.

Parole chiave: Approccio bayesiano, Markov Chain Monte Carlo Methods, Poisson Hurdle.

Contents

Abstract iii

Sommario v

Introduction 1

1 Statistical models 3
- 1.1 Hurdle and zero-inflated models . . . 3
  - 1.1.1 Poisson hurdle base model . . . 5
  - 1.1.2 Poisson hurdle regression . . . 6
  - 1.1.3 Poisson hurdle with group structure . . . 8
  - 1.1.4 Poisson hurdle regression with group structure . . . 8

2 Bayesian statistics 11
- 2.1 Introduction to Bayesian approach . . . 11
- 2.2 Markov chains and Monte Carlo methods . . . 13
  - 2.2.1 Metropolis-Hastings . . . 15

3 Test on synthetic data 17
- 3.1 Poisson Hurdle base model . . . 19
- 3.2 Poisson Hurdle Regression . . . 23
- 3.3 Poisson Hurdle with group structure . . . 26
- 3.4 Poisson Hurdle Regression with group structure . . . 30

4 Analysis of a real dataset 33
- 4.1 Dataset . . . 33
- 4.2 Descriptive statistics . . . 35
- 4.3 The Poisson hurdle model . . . 39
  - 4.3.1 Model setting . . . 39
  - 4.3.2 Prior distributions' choice . . . 42
  - 4.3.3 Bayesian posterior inference . . . 44
  - 4.3.4 Sensitivity analysis . . . 48
  - 4.3.5 Prediction . . . 49
  - 4.3.6 Computational costs . . . 53

Discussion and conclusions **55**

A R Functions **57**

A.1 Functions used for models in Sections 3.1, 3.2, 3.3 and 3.4 . . 57

Bibliography **59**

List of Figures

3.1 Traceplot of p parameter. 20
3.2 Traceplot of λ parameter. 20
3.3 Posterior distribution of p parameter 21
3.4 Posterior distribution of λ parameter. 21
3.5 Auto-correlation function for the chains of p and λ parameters. 21
3.6 Traceplot of $\beta^{(p)}$ parameter. 23
3.7 Traceplot of $\beta^{(\lambda)}$ parameter. 23
3.8 Posterior distribution of $\beta^{(p)}$ parameter. 24
3.9 Posterior distribution of $\beta^{(\lambda)}$ parameter. 24
3.10 Auto-correlation function for the chains of βs and λs parameters. 24
3.11 Groups cardinality (given by n). 26
3.12 Traceplots of p parameter's components. 27
3.13 Traceplots of λ parameter's components. 27
3.14 Posterior distribution of p parameters. 28
3.15 Posterior distribution of λ parameters. 28
3.16 Traceplot of $\beta^{(p)}$ parameter. 31
3.17 Traceplot of $\beta^{(\lambda)}$ parameter. 31
3.18 Traceplot of $\theta^{(p)}$ parameter. 31
3.19 Traceplot of $\theta^{(\lambda)}$ parameter. 31

4.1 Montesinho Natural park map. 34
4.2 Burned hectares (all observations to the left, 96% of the observations to the right). 36
4.3 Burned area distribution of the first 12 sub-regions of the map (corresponding to areas such that $X \in \{1, 2, 3\}$ and $Y \in \{2, 3, 4, 5\}$). It is the West territory of the park. 37
4.4 Burned area distribution of the 13-24th sub-regions of the map (corresponding to areas such that $X \in \{4, 5, 6\}$ and $Y \in \{3, 4, 5, 6\}$ plus region (7,3)). It is the middle territory of the park. 37
4.5 Burned area distribution of the 25-36th sub-regions of the map (corresponding to areas such that $X \in \{7, 8, 9\}$ and $Y \in \{3, 4, 5, 6, 7, 8, 9\}$). It is the East territory of the park. 38

4.6 Cardinality of observations of each region/group (denoted by n_k). 38
4.7 Boxplots of $FFMC$, DMC, DC and ISI quantities. 39
4.8 Boxplots of $Temperature$, $Relative\ humidity$, $Wind\ intensity$ and $Rain\ amount$ quantities. 39
4.9 Percentage of zeros (zero burned hectare) for each sub-region of the map. 40
4.10 Traceplots of $\beta^{(p)}$ parameters. 43
4.11 Traceplots of $\beta^{(\lambda)}$ parameters. 44
4.12 Posterior credible intervals for all parameters $\beta^{(p)}$, $\beta^{(\lambda)}$, $\theta^{(p)}$, $\theta^{(\lambda)}$, $\phi^{(p)}$ and $\phi^{(\lambda)}$; on the left the ones relative to p, on the right the ones relative to λ. 45
4.13 Groups effect $\theta^{(p)}$ relative to the fire ignition: regions in shades of red corresponds to positive $\theta_k^{(p)}$ (high probability of fire ignition), shades of blue regions corresponds to regions with negative $\theta_k^{(p)}$ (low proability), proportional to the magnitude of the coefficient. 47
4.14 Groups effect $\theta^{(\lambda)}$ relative to the fire spread: regions in shades of red corresponds to positive $\theta_k^{(\lambda)}$ (high fire spread), shades of blue regions corresponds to regions with negative $\theta_k^{(\lambda)}$ (low fire spread), proportional to the magnitude of the coefficient. 48
4.15 Empirical value of p parameter (computed through MLE) for each sub-region. Each single dot represents the empirical probability that a fire breaks out in the corresponding sub-region. 50
4.16 Bayesian estimate of p parameter for each sub-region. Each single dot represents the bayesian mean posterior probability that a fire breaks out in the corresponding sub-region. 50
4.17 Relative deviation RD_{p_k} between MLE and Bayesian estimations of p parameter for each sub-region $k \in \{1, \ldots, 35\}$. . . . 51
4.18 Empirical value of λ parameter (computed through MLE) for each sub-region. Each single dot represents the empirical magnitude of a broke out fire in the corresponding sub-region. . . 51
4.19 Bayesian estimate of λ parameter for each sub-region. Each single dot represents the bayesian posterior mean of the magnitude of a broke out fire in the corresponding sub-region. . . 52
4.20 Relative deviation RD_{λ_k} between MLE and Bayesian estimations of λ parameter for each sub-region $k \in \{1, \ldots, 35\}$. . . . 52
4.21 Prediction on new unseen data: the prediction is a single data (represented in red) obtained using as coefficients the parameters' posterior means $\mathbb{E}[\beta^{(\cdot)}|\mathcal{D}]$, $\mathbb{E}[\theta^{(\cdot)}|\mathcal{D}]$ and $\mathbb{E}[\phi^{(\cdot)}|\mathcal{D}]$, compared with the real observed value (the black dot). 54

4.22 Bayesian prediction of 50 new unseen data (whose outputs are known and represented in black dots): for each data 20 outcomes generated from the Bayesian predictive distribution are represented in red dots. 54

List of Tables

3.1 Parameter β and λ summary statistics about posterior distribution. 20
3.2 Required computational time for different sample size. 21
3.3 MSE of each parameter for different sample size. 22
3.4 Parameters summary statistics. 24
3.5 Required computational time for different sample size. 25
3.6 MSE of each parameter for different sample size. 25
3.7 Required computational time for different sample size. 28
3.8 MSE of each parameter for different sample size. 29
3.9 Posterior distributions' summary statistics. 31
3.10 Required computational time for different sample size. 32
3.11 MSE of each parameter for different sample size. 32

4.1 Summary of dataset attributes. 35
4.2 Dataset details. 35
4.3 Posterior distributions' summary statistics of $\beta^{(p)}$ and $\beta^{(\lambda)}$. . 46
4.4 Posterior distributions' summary statistics of $\phi^{(p)}$ and $\phi^{(\lambda)}$. . 46
4.5 Posterior distributions' summary statistics of $\theta^{(p)}$ and $\theta^{(\lambda)}$. . . 46
4.6 Required computational time for current dataset. 53

Introduction

Between all data types that exist in real world, count data play an important part. There are several ways to treat and model this kind of data but, inevitably, some discrete probability distributions like Poisson, Binomial and Negative Binomial are more suitable than others and strong of this fact, they are recurrent in the majority of studies which deal with modeling real problems with positive or null integer data.

A further aspect of these kind of data is that in some real-case circumstances they may have a spatial reference, provided that are observed within a metric space: they are immersed in a reference space that is usually a multi-dimensional space, not necessary a surface, even if in this elaborate we will consider only flat surfaces. A particular case is represented by areal data which is generated partitioning a domain in a finite number of sub-regions at which outcomes are aggregated.

When a non standard scenario shows up, for example when data contemplate an overabundance of zero values, the models that take advantage of the above mentioned distributions encounter some limitations, because of the limited shapes which the more conventional probability distributions offer. Fortunately some powerful countermeasures exist and they have been discussed in some scientific papers under the name of Hurdle and Zero-inflated models. For example, in [1] the choice of Ver Hoef and Jansen, in order to take in consideration the spatial dependence among data, has been to consider some random effects both in a hurdle and a zero-inflated regression model. In [2] Neelon, Ghosh and Loebs focused on the socio-economic problem related to the increasing demand of emergency department visits in Durham, North Carolina and they also adopted an hurdle regression model with random effects with the intent of discovering if a correlation between the access to the Emergency department (access or not) and the number of the visits would exists. The contribution of Ghosh, Gelfand, Zhu and Clark [3] is to address the limitations of these models when even the extreme overabundance of zero values (potentially more than 80%) represents an issue.

Taking this literature as a starting point we are going to deal with Bayesian models that involve the so called Hurdle distributions. In particular, in this elaborate the purpose is to investigate the potentiality of an hurdle model in a Bayesian framework and studying its application in cases where there is a group structure distributed over a space.

The thesis will be divided into 4 main chapters and final conclusions.
In particular, chapters are organized as follows. In Chapter 1, after submitting the Hurdle and the Zero-inflated concepts together with their probability notions, a bunch of models that exploits Poisson hurdle distributions are introduced, increasing the complexity one model after another: we start from a minimal Poisson Hurdle model and then we develop firstly a regression structure, using generalized linear models' theory, and subsequently a group structure among the observations. Chapter 2 briefly recap the theory of the Bayesian approach, along with the theory of Markov chains and the Markov chain Monte Carlo (MCMC) methods, which facilitates to sample from the Bayesian posterior distribution. In Chapter 3 we have tested with simulated data all the four models covered in the first chapter, before considering, in Chapter 4, a real-world dataset concerning the ignition and the spread of wildfires occurred in a western region of the Iberian Peninsula. The most specific model out of four has been applied to the dataset. After that, a Bayesian inference on model parameters and prediction complete the chapter.
In the end some conclusions on the work are made, included some ideas for future developments and improvements.

Chapter 1

Statistical models

When dealing with processes that yield an overabundance of zeros, hurdle and zero-inflated models are commonly used to address this problem because they allow more flexibility in modeling the probability of a zero outcome. The basic idea is to act on a well-known discrete distribution, such as a Poisson or Negative Binomial, increasing the probability of observing a zero value. We present now two classes of discrete models.

1.1 Hurdle and zero-inflated models

Suppose a Bernoulli random variable Z governs the binary outcome of whether a count random variable Y has a zero or a positive realization. The above statement rules both the zero-inflated model and the hurdle one.

In the case the Bernoulli realization is positive, if the conditional distribution of the positives is governed by a distribution whose support is any strictly positive count value (some possible choices are truncated-at-zero Poisson or Poisson plus one) is the case of hurdle models; otherwise, if the conditional distribution includes the zero value (such as Poisson or Negative Binomial (NB)), it is a Zero-inflated model and generally they are referred as ZI, with the addition of the suffix related to the chosen discrete distribution, like ZIP or ZINB for the previous cases.

Zero-inflated models have been theorized by Diane Lambert in 1992 [4] and the differences between them and hurdle models have been investigated so far; as said before they are both suitable to describe the high occurrence of zeros but their main difference is intrinsic and related to the source of the zero value: in a Poisson hurdle model the realization of value zero is uniquely possible if it is drawn from the Bernoulli trail because the Truncated Poisson has support only on strictly positive values. On the other hand, if we consider a Zero-inflated Poisson (ZIP), it is a mixture model where the zero value could come from both the Bernoulli trial and the Poisson distribution

since, in this case, the Poisson distribution has a support that includes zero. Just for completeness and to understand better what we have argued so far we formally introduce a ZIP model although it will not be used from now on because we will focus only on hurdle one.
We define a ZIP model starting from the conditional distribution of Y given the Bernoulli trial Z:

$$P(Y = y|Z = z) = \begin{cases} 0 & \text{if } z = 0 \\ \text{Pois}(y|\lambda) & \text{if } z = 1 \end{cases}$$

Given the above probability and knowing the distribution of $Z \sim Be(p)$, the marginal distribution of Y is:

$$P(Y = y) = \begin{cases} (1-p) + pe^{-\lambda} & \text{if } y = 0 \\ p \ \text{Pois}(y|\lambda) & \text{if } y = 1, 2, \ldots \end{cases}$$

where we can notice the two possible sources of the probability of observing a value of zero.

Let us introduce the other model class, the hurdle model, with the specific case of a Poisson hurdle one.
A random variable $Y \sim HPois(p, \lambda)$ has a Poisson hurdle distribution if the conditional distribution of Y given Z is:

$$P(Y = y|Z = z) = \begin{cases} 0 & \text{if } z = 0 \\ \text{tPois}(y|\lambda) & \text{if } z = 1 \end{cases}$$

where

$$\text{tPois}(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!(1 - e^{-\lambda})} = \frac{\lambda^y}{y!(e^{\lambda} - 1)} \qquad y = 1, 2, \ldots$$

is the truncated Poisson density and the marginal distribution of Y is:

$$P(Y = y) = \begin{cases} 1-p & \text{if } y = 0 \\ p \ \text{tPois}(y|\lambda) & \text{if } y = 1, 2, \ldots \end{cases}$$

The next subsections will be organized following a common path: firstly the $Z = \mathbb{I}\{Y > 0\}$ random variable is introduced: it has two possible distinct realizations that are 1 if $Y > 0$ and 0 if $Y = 0$. Then the conditional distribution $Y|Z$ is computed. After these two preliminary steps and using the property of conditional probability the likelihood is straightforward.
The likelihood function describes the joint probability of the observed data as function of the parameters of the chosen statistical model and if the random variables that are responsible of the given data (let us indicate them as Y_i in a general scenario) are assumed to be independent and identically distributed

the likelihood is given by:

$$L(y|\theta) = P(Y_1 = y_1, \cdots, Y_N = y_n|\theta) = \prod_{i\in\{1,\dots,N\}} P(Y_i = y_i|\theta) \tag{1.1}$$

where y is the vector of realization of $\{Y_i\}_{i\in\{1,\cdots,N\}}$ and θ is the vector of hidden parameters under the distribution of Y_i.
Let us consider a sample of independent Poisson hurdle random variables $\{Y_1, \cdots, Y_N\}$ and use it to formally introduce some models. The main focus is to exploit the model structure, get the likelihood function and take advantage of it, later on, for a Bayesian analysis.

1.1.1 Poisson hurdle base model

We start from the simplest model involving a Poisson hurdle scenario. Consider the sample of independent random variables given by $\{Y_1, \dots, Y_N\}$, where:

$$Y_i \overset{\text{iid}}{\sim} HPois(p, \lambda) \qquad i \in \{1, 2, ..., N\}$$

Its likelihood function (seen as a function of the parameters given the data) is:

$$\begin{aligned} L(p, \lambda|y) &= \prod_{i\in\Omega_0} (1-p) \prod_{i\in\Omega_1} p \frac{\lambda^{y_i} e^{-\lambda}}{y_i!(1-e^{-\lambda})} \\ &= (1-p)^{|\Omega_0|} \cdot p^{N-|\Omega_0|} \cdot \prod_{i\in\Omega_1} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!(1-e^{-\lambda})} \end{aligned} \tag{1.2}$$

where $y = (y_1, \dots, y_N)$ is the realization of $Y = (Y_1, \dots, Y_N)$, $\Omega_0 = \{i|y_i = 0\}$, $\Omega_1 = \{i|y_i > 0\}$ and $\Omega_0 \cup \Omega_1 = \{1, 2, ..., N\}$ (so $|\Omega_0| + |\Omega_1| = N$, where $|\cdot|$ denotes the cardinality of a set).

Introducing $Z_i = \mathbb{I}\{Y_i > 0\}$, by the independence of $\{Y_i\}$ for $i \in \{1, \cdots, N\}$, the likelihood of $(Z_1, \dots, Z_N)$ is the following:

$$P(Z_1 = z_1, \dots, Z_N = z_N|p) = \prod_{i=1}^{N} p^{z_i}(1-p)^{1-z_i} = p^{\sum_{i=1}^N z_i} \cdot (1-p)^{N-\sum_{i=1}^N z_i} \tag{1.3}$$

Compute now the joint conditional likelihood of $Y|Z$, where $Z = (Z_1, \dots, Z_N)$ and $z = (z_1, \dots, z_N)$ is its realization:

$$P(Y_1 = y_1, \dots, Y_N = y_N|z, \lambda) = \prod_{i=1}^{N} \Big(\frac{\lambda^{y_i} e^{-\lambda}}{y_i!(1-e^{-\lambda})}\Big)^{z_i} \mathbb{I}(y_i = 0)^{1-z_i} \tag{1.4}$$

In this way, multiplying (1.3) and (1.4) we get exactly the previous result (1.2) for the likelihood of $\{Y_1, \dots, Y_N\}$, written this time as a function of Z and considering that $|\Omega_0| = N - \sum_{i=1}^{N} z_i$:

$$\begin{aligned} L(p,\lambda|y,z) &= \prod_{i \in 1,\dots,N} (1-p)^{1-z_i} \left(p \frac{\lambda^{y_i} e^{-\lambda}}{y_i!(1-e^{-\lambda})} \right)^{z_i} = \\ &= (1-p)^{N-\sum_{i=1}^{N} z_i} \cdot p^{\sum_{i=1}^{N} z_i} \cdot \prod_{i \in \Omega_1} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!(1-e^{-\lambda})} \end{aligned} \tag{1.5}$$

And by the exponential property:

$$L(p,\lambda|y,z) = (1-p)^{N-\sum_{i=1}^{N} z_i} \cdot p^{\sum_{i=1}^{N} z_i} \cdot \frac{\lambda^{\sum_{i=1}^{N} y_i} e^{-|\Omega_1|\lambda}}{(1-e^{-\lambda})^{|\Omega_1|}} \prod_{i \in \Omega_1} \frac{1}{y_i!}$$

1.1.2 Poisson hurdle regression

The next model of our list (and some others further on) is based on the Generalized Linear Model (GLM) concept. The term *generalized* linear model refers to a class of models spread for the first time by McCullagh and Nelder in 1982 [5]. In these types of models, the response variable Y has a distribution belonging to the exponential family and its mean μ is assumed to be dependent on some covariates $x \in \mathbb{R}^J$ through some predictors $\beta \in \mathbb{R}^J$ using a (often nonlinear) function f such that $\mu = f(x^T\beta)$. The fact that the function is "nonlinear" is misleading because according to McCullagh and Nelder the covariates affect the distribution of Y only through the linear combination $x^T\beta$.

In a GLM framework we adopt a link function for both the parameters p and λ and we refer to the predictor as the usual GLM notation $\eta = x^T\beta$. The possible choices for the link function of p are:

(a) log-log link function $\Longleftrightarrow$ $ln(ln(p)) = \eta$;

(b) log link function $\Longleftrightarrow$ $ln(p) = \eta$

(c) logit link function $\Longleftrightarrow$ $ln(\frac{p}{1-p}) = \eta$, where $\frac{p}{1-p}$ is the called odds ratio;

(d) probit link function $\Longleftrightarrow$ $\Phi^{-1}(p) = \eta$, where $\Phi^{-1}(\cdot)$ is the cumulative distribution function of the standard normal distribution;

Instead for λ usually consider:

(a) log link function $\Longleftrightarrow$ $ln(\lambda) = \eta$;

We have introduced the GLM concept because we want to extend the model of the previous section, introducing a regression on its parameters p and λ.

We could have some covariates related to the p parameter and some others related to the λ parameter, which we collect them in the following matrices $X^{(p)} \in \mathbb{R}^{N\times J}$ and $X^{(\lambda)} \in \mathbb{R}^{N\times J}$. In particular each row $x_i^{(p)}$ and $x_i^{(\lambda)}$ of the matrices refers to a single observation.
For example choosing (c) for p and (a) for λ as desired link functions and substituting them in (1.5), the model becomes:

$$\begin{array}{ll} Y_i|x_i^{(p)}, x_i^{(\lambda)} \overset{\text{iid}}{\sim} HPois(p_i, \lambda_i) & i \in \{1, 2, \dots, N\} \\ log(\frac{p_i}{1-p_i}) = x_i^{(p)}\beta^{(p)} & i \in \{1, 2, \dots, N\} \\ log(\lambda_i) = x_i^{(\lambda)}\beta^{(\lambda)} & i \in \{1, 2, \dots, N\} \end{array}$$

The likelihood is a function of the new parameters $\beta^{(p)}$ and $\beta^{(\lambda)}$:

$$\begin{aligned} &L(\beta^{(\lambda)}, \beta^{(p)}|y, z, X^{(p)}, X^{(\lambda)}) = \\ &= \prod_{i\in 1,\dots,N} (1 - p_i(x_i^{(p)}\beta^{(p)}))^{1-z_i} \left(p_i(x_i^{(p)}\beta^{(p)})) \frac{\lambda_i(x_i^{(\lambda)}\beta^{(\lambda)})^{y_i} e^{-\lambda_i(x_i^{(\lambda)}\beta^{(\lambda)}))}}{y_i!(1 - e^{-\lambda_i(x_i^{(\lambda)}\beta^{(\lambda)})})} \right)^{z_i} \end{aligned} \tag{1.6}$$

Instead of considering the likelihood itself we apply the logarithmic function to (1.6) because we want to show an important peculiarity of hurdle models. The so called log-likelihood is the following:

$$\begin{aligned} &ln(L(\beta^{(\lambda)}, \beta^{(p)}|y, z, X^{(p)}, X^{(\lambda)})) = \\ &= ln\left(\prod_{i\in 1,\dots,N} (1 - p_i(x_i^{(p)}\beta^{(p)}))^{1-z_i} \left(p_i(x_i^{(p)}\beta^{(p)})) \frac{\lambda_i(x_i^{(\lambda)}\beta^{\lambda})^{y_i} e^{-\lambda_i(x_i^{(\lambda)}\beta^{\lambda})}}{y_i!(1 - e^{-\lambda_i(x_i^{(\lambda)}\beta^{\lambda})})} \right)^{z_i} \right) = \\ &= \sum_{i=1}^{N} (1 - z_i) \cdot ln\{1 - p_i(x_i^{(p)}\beta^{(p)})\} + \sum_{i=1}^{N} z_i \cdot ln\{p_i(x_i^{(p)}\beta^{(p)})\} + \\ &+ \sum_{i=1}^{N} z_i \cdot ln\left\{ \left(\frac{\lambda_i(x_i^{(\lambda)}\beta^{(\lambda)})^{y_i} e^{-\lambda_i(x_i^{(\lambda)}\beta^{(\lambda)}))}}{y_i!(1 - e^{-\lambda_i(x_i^{(\lambda)}\beta^{(\lambda)}))}} \right) \right\} = \\ &= \sum_{i=1}^{N} (1 - z_i) \cdot ln\{1 - p_i(x_i^{(p)}\beta^{(p)})\} + \sum_{i=1}^{N} z_i \cdot ln\{p_i(x_i^{(p)}\beta^{(p)})\} + \\ &+ \sum_{i:z_i>0} \Big(y_i \cdot ln(\lambda_i(x_i^{(\lambda)}\beta^{(\lambda)})) - \lambda_i(x_i^{(\lambda)}\beta^{(\lambda)}) - ln(y_i!) - ln(1 - e^{-\lambda_i(x_i^{(\lambda)}\beta^{(\lambda)})}) \Big) \\ &= \Lambda_1(\beta^{(p)}|z, X^{(p)}) + \Lambda_2(\beta^{(\lambda)}|y, z, X^{(\lambda)}) \end{aligned} \tag{1.7}$$

Observe that the log-likelihood function is separable with respect to the parameters $\beta^{(p)}$ and $\beta^{(\lambda)}$. Due to this fact the log-likelihood function consists in two contributions: a term related to the binary outcome trial, that is a function of the only $\beta^{(p)}$ parameter, and the other term responsible of the zero-truncated trial which incorporates the parameter $\beta^{(\lambda)}$.

1.1.3 Poisson hurdle with group structure

Let us consider now a more specific case in which data are grouped in such a way we have K groups of n_k observations. Basically we have K groups and each group has proper parameters p_k and λ_k that distinguish the group. Assuming the above structure the model is the following:

$$Y_{i,k} \overset{\text{iid}}{\sim} HPois(p_k, \lambda_k), \qquad i \in \{1, 2, ..., n_k\},\ k \in \{1, 2, ..., K\}$$

where i denotes the single observation of the group, k denotes the specific group, $n = (n_1, \ldots, n_K)^T$ is the vector which collects the cardinality of each group, $p = (p_1, \ldots, p_K)^T$ collects the p_k parameter of each group and the same for λ_k with $\lambda = (\lambda_1, \ldots, \lambda_K)^T$.
Adopting the usual notation for the $Z_{i,k} = \mathbb{I}\{Y_{i,k} > 0\}$ variable the likelihood function of Z is:

$$\begin{aligned} &P(Z_{1,1} = z_{1,1}, \ldots, Z_{n_K,K} = z_{n_K,K}|p) = \\ &\prod_{k=1}^{K}\prod_{i=1}^{n_k} p_k^{z_{i,k}}(1-p_k)^{1-z_{i,k}} = \prod_{k=1}^{K} p_k^{\sum_{i=1}^{n_k} z_{i,k}} \cdot (1-p_k)^{n_k - \sum_{i=1}^{n_k} z_{i,k}} \end{aligned} \tag{1.8}$$

Consider now the joint contribution of the response variables Y given the Bernoulli trials Z, $Y|Z$:

$$\begin{aligned} &P(Y_{1,1} = y_{1,1}, \ldots, Y_{n_K,K} = y_{n_K,K}|z, \lambda) = \\ &\prod_{k=1}^{K}\prod_{i=1}^{n_k}\Big(\frac{\lambda^{y_{i,k}}e^{-\lambda_k}}{y_{i,k}!(1-e^{-\lambda_k})}\Big)^{z_{i,k}}\mathbb{I}(y_{i,k}=0)^{1-z_{i,k}} \end{aligned} \tag{1.9}$$

Multiplying (1.8) and (1.9) we get the likelihood function:

$$L(p, \lambda|y, z) = \prod_{k=1}^{K}\prod_{i=1}^{n_k}(1-p_k)^{1-z_{i,k}}\left(p_k\frac{\lambda^{y_{i,k}}e^{-\lambda_k}}{y_{i,k}!(1-e^{-\lambda_k})}\right)^{z_{i,k}} \tag{1.10}$$

1.1.4 Poisson hurdle regression with group structure

The following model has the same group structure described in Section 1.1.3 and the same notation of Section 1.1.2 for the quantities involved in the regression part, with the addition of some covariates linked to the group membership and collected in the matrices $\hat{X}^{(p)} \in \mathbb{R}^{K\times J}$ and $\hat{X}^{(\lambda)} \in \mathbb{R}^{K\times J}$, where each row $\hat{x}_k^{(p)}$ and $\hat{x}_k^{(\lambda)}$ of the matrices refers to a single observation.

A regression on both parameters is performed.
Given $\{Y_{1,1},\dots,Y_{n_1,1},\dots,Y_{1,K},\dots,Y_{n_k,K}\}$ the GLM model is the following:

$$\begin{array}{ll}
Y_{i,k}|x^{(p)}_{i,k},x^{(\lambda)}_{i,k} \overset{\text{iid}}{\sim} HPois(p_{i,k},\lambda_{i,k}) & i\in\{1,2,\dots,n_k\}, k\in\{1,2,\dots,K\}\\
log(\frac{p_{i,k}}{1-p_{i,k}}) = x^{(p)}_{i,k}\beta^{(p)}+\hat{x}^{(p)}_{k}\theta^{(p)} & i\in\{1,2,\dots,n_k\}, k\in\{1,2,\dots,K\}\\
log(\lambda_{i,k}) = x^{(\lambda)}_{i,k}\beta^{(\lambda)}+\hat{x}^{(\lambda)}_{k}\theta^{(\lambda)} & i\in\{1,2,\dots,n_k\}, k\in\{1,2,\dots,K\}
\end{array}$$

Observe that in the regression structure there are two different contributions: the first one which includes the $\beta^{(\cdot)}$ parameter is observation-wise, while the second one, with $\theta^{(\cdot)}$ parameter, is group-wise. Hence the row $x^{(\cdot)}_{i,k}$ of $X^{(\cdot)}$ refers to the observation i of group k, while the row $\hat{x}^{(\cdot)}_{k}$ of $\hat{X}^{(\cdot)}$ refers to the observation of group k.
As done before it is useful to get the marginal likelihood of Z, where $Z_{i,k}=\mathbb{I}\{Y_{i,k}>0\}$:

$$\begin{aligned}
&P(Z_{1,1}=z_{1,1},\dots,Z_{n_K,K}=z_{n_K,K}|\beta^{(p)},\theta^{(p)},X^{(p)},\hat{X}^{(p)}) =\\
&=\prod_{k=1}^{K}\prod_{i=1}^{n_k}p_{i,k}(x^{(p)}_{i,k}\beta^{(p)}+\hat{x}^{(p)}_{k}\theta^{(p)})^{z_{i,k}}(1-p_{i,k}(x^{(p)}_{i,k}\beta^{(p)}+\hat{x}^{(p)}_{k}\theta^{(p)}))^{1-z_{i,k}}=\\
&=: L_1(\beta^{(p)},\theta^{(p)})|z,X^{(p)},\hat{X}^{(p)})
\end{aligned} \tag{1.11}$$

Then we consider each $Y_{i,k}|Z_{i,k}$ contribution $\forall i\in\{1,2,\dots,n_k\},\forall k\in\{1,2,\dots,K\}$ and compute:

$$\begin{aligned}
&P(Y_{1,1}=y_{1,1},\dots,Y_{n_K,K}=y_{n_K,K}|z,\beta^{(\lambda)},\theta^{(\lambda)},X^{(\lambda)},\hat{X}^{(\lambda)})\\
&=\prod_{k=1}^{K}\prod_{i=1}^{n_k}\Big(\frac{\lambda_{i,k}(x^{(\lambda)}_{i,k}\beta^{(\lambda)}+\hat{x}^{(\lambda)}_{k}\theta^{(\lambda)})^{y_{i,k}}}{y_{i,k}!(1-e^{-\lambda_{i,k}(x^{(\lambda)}_{i,k}\beta^{(\lambda)}+\hat{x}^{(\lambda)}_{k}\theta^{(\lambda)})})}e^{-\lambda_{i,k}(x^{(\lambda)}_{i,k}\beta^{(\lambda)}+\hat{x}^{(\lambda)}_{k}\theta^{(\lambda)})}\Big)^{z_{i,k}}\cdot\\
&\cdot\mathbb{I}(y_{i,k}=0)^{1-z_{i,k}}=: L_2(\beta^{(\lambda)},\theta^{(\lambda)}|y,z,X^{(\lambda)},\hat{X}^{(\lambda)})
\end{aligned} \tag{1.12}$$

The full likelihood is the product of (1.11) and (1.12):

$$\begin{aligned}
&L(\beta^{(p)},\beta^{(\lambda)},\theta^{(p)},\theta^{(\lambda)}|y,z,X^{(p)},\hat{X}^{(p)},X^{(\lambda)},\hat{X}^{(\lambda)})=\\
&L_1(\beta^{(p)},\theta^{(p)}|z,X^{(p)},\hat{X}^{(p)})\times L_2(\beta^{(\lambda)},\theta^{(\lambda)}|y,z,X^{(\lambda)},\hat{X}^{(\lambda)})=\\
&=\prod_{k=1}^{K}\prod_{i=1}^{n_k}(1-p_{i,k}(x^{(p)}_{i,k}\beta^{(p)}+\hat{x}^{(p)}_{k}\theta^{(p)}))^{1-z_{i,k}}\cdot p_{i,k}(x^{(p)}_{i,k}\beta^{(p)}+\hat{x}^{(p)}_{k}\theta^{(p)})^{z_{i,k}}\cdot\\
&\cdot\Big(\frac{\lambda_{i,k}(x^{(\lambda)}_{i,k}\beta^{(\lambda)}+\hat{x}^{(\lambda)}_{k}\theta^{(\lambda)})^{y_{i,k}}}{y_{i,k}!(1-e^{-\lambda_{i,k}(x^{(\lambda)}_{i,k}\beta^{(\lambda)}+\hat{x}^{(\lambda)}_{k}\theta^{(\lambda)})})}e^{-\lambda_{i,k}(x^{(\lambda)}_{i,k}\beta^{(\lambda)}+\hat{x}^{(\lambda)}_{k}\theta^{(\lambda)})}\Big)^{z_{i,k}}
\end{aligned} \tag{1.13}$$

Chapter 2

Bayesian statistics

2.1 Introduction to Bayesian approach

There are two approaches to statistics. On one side the frequentist one and on the other side the Bayesian one. If we are in a parametric environment, the former one assumes that the underlying numerical characteristics of the population, the parameters, are unknown but fixed, with the direct consequence that probability statements about their distributions can not be made. Vice versa the Bayesian assumption consists in the randomization of the parameters in fact the philosophical and intuitive explanation is that since they are unknown, they must be considered as random variables with a proper distribution and with the possibility to make inference about it. Considering the random nature of the parameters, a *prior* distribution is the probabilistic translation of the possible prior believe one could have about the parameters before observing the data.

The prior assumption is just a guess in fact the Bayes' theorem allows to update the prior distribution through data, getting in this way, the *posterior* distribution. In this way the prior believe on parameters could be modified and partially (or totally) changed according to the observed data.

As long as the frequentist approach, the first step in a Bayesian framework is to choose a probability model for the data. This process requires understanding and deciding on a probability distribution for the data. Suppose we have N data given by $\mathcal{D} = \{Y_1, \cdots, Y_N\}$ where $Y_i \in \mathbb{R}^D$ and their density is given by $f(y_1, \ldots, y_N|\theta)$. The quantity $\theta = (\theta_1, \cdots, \theta_K)^T \in \Theta \subseteq \mathbb{R}^K$ is the vector of unknown parameters assumed random.

Once the data model is chosen, it is mandatory the assertion of a prior distribution $\pi(\theta)$ on the unknown parameters. The choice of the prior distribution could be of different types: for example an approach could be choosing an *informative* prior distribution and with this strategy the statistician exploits his knowledge about the problem to construct a prior distribution that reflects his beliefs about the unknown parameters. Another approach

is choosing a *non-informative* prior, which represents ignorance about the model parameters. Choosing a non-informative prior distribution could be a smart option when no prior knowledge about the parameters exists before observing the data.
Another strategy to choose the prior distribution is to adopt a certain prior such that the posterior ends up being in the same distribution family as the prior. If this holds it is called *conjugate* prior. The advantage of this option is that if your conjugate prior distribution has a closed-form form expression, the posterior distribution's summary statistics such as maximum, mean, variance and so on are easy to compute analytically. However, proceeding in this way, we can hardly put the knowledge about the real problem in the prior distribution.
Retrieve now the Bayes' theorem [6], which will help us for the next considerations. Given a random vector (X, Y), let $f_{X,Y}(x, y)$, $f_X(x)$, $f_Y(y)$ be the density functions of (X, Y), X, Y respectively, and let $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ be the density function of X conditioned on Y and Y conditioned on X. The Bayes theorem reads as:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) \cdot f_X(x)}{f_Y(y)} \tag{2.1}$$

Once data has been observed, the likelihood function $\mathcal{L}(\mathcal{D}|\theta)$, which describes the distribution of data, is straightforward. Then, using the Bayes' theorem and substituting the likelihood and the prior distribution, we are able to compute the posterior distribution:

$$\pi(\theta|\mathcal{D}) = \frac{\mathcal{L}(\mathcal{D}|\theta)\pi(\theta)}{\int_\Theta \mathcal{L}(\mathcal{D}|\theta)\pi(\theta)\, d\theta} \tag{2.2}$$

where $\pi(\theta|\mathcal{D})$ is the posterior distribution of θ, $\mathcal{L}(\mathcal{D}|\theta)$ is the likelihood of the data, $\pi(\theta)$ is the prior distribution on parameter θ and Θ is the domain of θ. Note that the denominator does not depend on θ so the posterior is proportional the the numerator.
The main interest is to handle the posterior distribution but often it is not easy since it is not guaranteed to be in a close-form expression, making it difficult sampling from it. For this purpose there exist some techniques and our focus is on the Markov Chains Monte Carlo methods (MCMC). Basically they are a family of algorithms that uses Monte Carlo methods to construct a Markov Chain from which we get the posterior distribution.
Once the posterior is available we would still like to have a single estimate $\hat{\theta}$, as well as an interval describing our uncertainty about θ. The posterior mean ($\mathbb{E}[\theta|\mathcal{D}]$) is commonly used for $\hat{\theta}$, while a $100(1-\alpha)\%$ Bayesian credible interval for θ (let us call it IC_θ), which is an interval such that the posterior probability $P(\theta \in IC_\theta|\mathcal{D}) = 1 - \alpha$, is used to point out the uncertainty of θ. One common choice for IC_θ is simply the interval $[\theta_{\frac{\alpha}{2}}, \theta_{\frac{1-\alpha}{2}}]$, where $\theta_{\frac{\alpha}{2}}$ and

$\theta_{\frac{1-\alpha}{2}}$ are the $\frac{\alpha}{2}$ and $\frac{1-\alpha}{2}$ quantiles of the posterior distribution $\pi(\theta|\mathcal{D})$.
In the end we may be interested in making prediction of new (or future) observations and in the Bayesian framework predicting means sampling from the posterior predictive distribution.
The posterior predictive distribution for a new data point Y^{new} after having observed data $\mathcal{D}$ is given by:

$$f(Y^{new}|\mathcal{D}) = \int_{\Theta} f(Y^{new}|\theta, \mathcal{D})\pi(\theta|\mathcal{D})\, d\theta = \int_{\Theta} f(y|\theta)\pi(\theta|\mathcal{D})\, d\theta$$

(since Y^{new} is independent of the sample data $\mathcal{D}$ and $f(y|\theta)$ is the density of the Poisson hurdle random variable).
For any further information about Bayesian theory and inference see [7] and [8].

2.2 Markov chains and Monte Carlo methods

Let us consider, for our purpose, a discrete-time Markov chain. Briefly a discrete-time Markov chain is a sequence of random variables where the current value is probabilistically dependent just on the value of the previous variable.
Formally, given a probability space $(\Omega, \mathcal{F}, P)$ and a space state set Θ with its $\mathcal{B}$orel σ-algebra, the succession of random variables $\{\theta_n\}_{n\geq 0}$ on $(\Omega, \mathcal{F}, P)$ with values in Θ is a discrete-time Markov chain if the following property (called *Markov property*) holds:

$$\begin{aligned} &P(\theta_{n+1} \in (\cdot)\ \ |\theta_n = \bar{\theta}_n, \theta_{n-1} = \bar{\theta}_{n-1}, \ldots, \theta_1 = \bar{\theta}_1) = \\ &P(\theta_{n+1} \in (\cdot)\ \ |\theta_n = \bar{\theta}_n) \quad \forall n \in \mathbb{N}, \quad \forall\ (\bar{\theta}_1, \ldots, \bar{\theta}_{n-1}, \bar{\theta}_n, \bar{\theta}_{n+1})^T \end{aligned} \tag{2.3}$$

As already explained, the realization of the next variable θ_{n+1} is only dependent upon the last variable in the chain (θ_n).
MCMC methods expect to construct a Markov Chain with support on the parameter space Θ, whose invariant distribution (the distribution of the states of the chain after a sufficiently long time that the distribution do not change any longer) coincides with the posterior distribution $\pi(\theta|\mathcal{D})$; in this way the chain could be considered as a sample obtained from the posterior distribution.
In application studies it is important to check that we are sampling from the correct posterior distribution otherwise the results, included any posterior inference, are not reliable at all. We have guarantee to get reliable results if and only if the chain reaches its invariant distribution (see [9] for a rigorous proof) and for the occurrence are used some graphical techniques that are going to be explained in details in the next chapter.

As extensively said, within the Bayesian framework the main interesting object is the posterior probability distribution of parameters but not always it is straightforward to compute it: in many cases the designed model has extremely complex structure and Monte Carlo methods are useful as a computational device with the ultimate goal of characterizing these complex distributions and performing statistical inference at the end.
In general Monte Carlo methods help us in several different situations; a simple scenario is when quantities like integrals must be computed.
For example define:

$$I = \int_{\Theta} g(\theta)\pi(\theta|\mathcal{D})\,d\theta = \mathbb{E}_{\pi(\theta|\mathcal{D})}(g(\theta)) < \infty = \mathbb{E}_{\pi(\theta|\mathcal{D})}[g(\theta)] < \infty$$

Having an analytical solution of these integrals is usually unfeasible when $\pi(\theta|\mathcal{D})$ is complex or even not in close form expression, but we can obtain an approximate solution. For example here a Monte Carlo method simply consists of getting an independent and identically distributed (i.i.d.) sample from the parameter vector to be inferred and use it to approximate the desired integral by mean of an unweighted sum. The N draws $\{\theta^{(n)}\}_{n\in\{1,\cdots,N\}}$ can be obtained either by sampling directly from the target distribution (i.e., the posterior $\pi(\theta|\mathcal{D})$), as shown in Algorithm 1, or by replicating the physical procedure where the desired parameters are involved.

ALGORITHM 1

- Draw an i.i.d. sample $\{\theta^{(n)}\}$ from $\pi(\theta|\mathcal{D})$, for $n \in \{1, \cdots, N\}$
- Compute $\hat{I}_N = \frac{1}{N}\sum_{n=1}^{N} g(\theta^{(n)})$

$\hat{I}_N$ is the Monte Carlo estimate of I and it is an unbiased approximation. Moreover, by the strong law of large numbers, $\hat{I}_N \to I$ almost surely (a.s.) as $N \to \infty$. Furthermore, if $g(\theta)$ is square integrable w.r.t. $\pi(\theta|\mathcal{D})$, then we can use the central limit theorem (TCL) [10] to state that:

$$\frac{\hat{I}_N - I}{\sqrt{V_N}} \xrightarrow{\mathrm{d}} N(0,1) \qquad N \to \infty,$$

where $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution, and

$$V_N = \frac{1}{N}\,\mathbb{E}_{\pi}((g(\theta) - I)^2) = \frac{1}{N}\int_{\Theta}(g(\theta) - I)^2\pi(\theta|\mathcal{D})d\theta.$$

Unfortunately, Algorithm 1 cannot be always applied, because we cannot always draw samples directly from $\pi(\theta|\mathcal{D})$. However, in these cases, if we can perform point-wise evaluations of the quantity $\tilde{\pi}(\theta|\mathcal{D}) = \mathcal{L}(\mathcal{D}|\theta)\pi(\theta)$ (that is the numerator of (2.1), which is proportional to the posterior $\pi(\theta|\mathcal{D})$), we can apply other types of Monte Carlo algorithms, like Markov chain Monte Carlo (MCMC).

2.2.1 Metropolis-Hastings

One of the most popular MCMC algorithm is the Metropolis-Hastings method [11]. In particular, as a general idea, some values are sampled iteratively and then accepted or rejected according to a certain probability. If they are accepted, not even they contribute to the resulting Markov chain itself but they are also responsible for the sampled value of the next value of the chain; otherwise, if rejected, they are discarded.

Specifically, the algorithm let $\pi(x)$ be the probability mass (or density) function of the distribution from which we wish to extract a sample of draws. We call it target distribution. Denote by $q(x^{new}|x)$ a family of conditional distributions of arbitrary choice, from which it is easy to generate draws (let us call it proposal distribution). It is required that x, x^{new} have the same dimension.

The Metropolis-Hastings algorithm starts from any value x_0 belonging to the support of the target distribution. Then, the values $x_1, \dots x_T$ are generated. In particular, a generic value x_t with $t \in \{1, \dots, T\}$, is generated as follows:

1. Draw x^{new} from the proposal distribution with density $q(x^{new}|x_{t-1})$;
2. Set
$$\alpha_t = min\Big(\frac{\pi(x^{new})}{\pi(x_{t-1})}\frac{q(x_{t-1}|x^{new})}{q(x^{new}|x_{t-1})}, 1\Big) \tag{2.4}$$
3. Draw u_t from a uniform distribution on $[0, 1]$;
4. Check the condition: if $u_t \leq \alpha_t$, set $x_t = x^{new}$; otherwise, set $x_t = x_{t-1}$. Since u_t is uniform the probability of accepting the proposal x^{new} as the new draw x_t is equal to α_t.

Let us point out some elements of the algorithm (but also typical of other MCMC processes):

- **Burn-in**: a random point is chosen to be the first sample from the chain. It may take some time for moving far away from this initial starting point. If the target distribution has a sparser density in that values of the support, the estimates produced from the MCMC will be biased. To mitigate this, an initial portion of the Markov chain is discarded so that the effect of initial values on inference is minimized. This is referred to as the *burn-in* period;
- **Efficiency**: A probability density, or proposal distribution, is assigned to suggest a candidate for the next sample value, given the previous sample value. A typical choice is to let the proposal distribution be such that points closer to the previous sample point are more likely to be visited next. Whatever form (Gaussian or otherwise) the proposal

distribution takes on, the goal is for this function to adequately and efficiently explore the sample space where the target distribution has the greatest density. If the target distribution is very broad and the proposal distribution is too "narrow" it may take quite a while for the walk to find its way around the whole target distribution, and the MCMC will not be very efficient;

- **Acceptance ratio**: an acceptance ratio is used to decide whether to accept or reject the next proposed sample. Observe that this ratio is proportional to the density of the target distribution. If the proposal distribution is too broad, the acceptance ratio may hardly allow the chain to move from the current spot and the the chain may be trapped in a localized area of the target distribution.

The probabilistic programming language for statistical modeling that we are going to use is the open-source software STAN [12]. It takes advantage of the *rstan* package which allows it to interface with the already mentioned *R* (and also with other different programs like *Python*, *MATLAB*, *Stata*, and *Mathematica*).
Briefly to define a model using the STAN language it is sufficient to specify the data (as input for STAN), the parameters to be estimated (or even the transformed parameters which are optional variables used as transformation of the model parameters), the model (which includes definition of priors for each parameter and the likelihood for the data) and the generated quantities (some quantities that are not part of the model but can be computed from the parameters). Note that STAN's MCMC techniques are based on Hamiltonian Monte Carlo (HMC), a more efficient and robust sampler than Metropolis-Hastings for models with complex posteriors.

Chapter 3

Test on synthetic data

In this Chapter we go back to models widely described in Chapter 1 to adopt and use them under the Bayesian assumption. Data is generated with some user-defined functions (see Appendix A) that sample from a Poisson hurdle model with fixed hidden parameters (each model has its own function according to its structure). With hidden we refer to the fixed input parameters given to the user-defined functions in order to generate the data sample.

For each model we care about varying the number of available data, in order to simulate different sizes of the hypothetical dataframe and the starting points from which we start sampling to get the posterior distribution. In addition to this, different hyper-parameter values, resulting in different prior distributions, have been tried. In other words a sensitivity analysis, in order to check the robustness of the models with respect to the hyper-parameters, has been performed. Then we have used some diagnostic tools to monitor the reliability of the results, especially addressing the chains of the parameters' posterior distribution. Traceplots are helpful to check if the chains are mixed (in this way the chain explores all possible values of the posterior distribution), without trends, cycles or seasonalities and ACF (auto-correlation function) plot provides a rough estimate about how much information we are loosing through thinning (the conservative procedure which keep just a value in a buffer of values provided by the sampler). In the end we compute the Mean Square Error of each parameter (MSE), using the hidden parameter values θ and the values produced by the chain of the model $\hat{\theta}_i$, according to the formula:

$$MSE_\theta = \frac{1}{T} \sum_{i=1}^{T} (\hat{\theta}_i - \theta)^2$$

where T is the size of the chain.

The goal is to analyse and compare the performances of each model using generated data.

For each model we have decided to run two independent chains to have a more reliable parameters estimate. Then we varied the fixed starting points of the chains to check that the choice does not influence the posterior results. The two chains are run for a total maximum of 10000 iterations with a warm-up of 1000 (so these values are not considered in the chain because they are the initial points of the Markov chain and for this reason more subject to the arbitrary chosen starting points) and a thinning of 20 which means that every 20 states of the Markov chain the state value is kept and considered for the final Markov chain.

In the end, in order to simulate different real cases scenarios we have considered different sample sizes of data, from a minimum a 10 to a maximum of 1000 and even 10000 in some cases.

Note that in all the plots and tables of the current Section (where not explicitly stated) a standard data sample size of 100 has been considered and every time a different sample size is taken, although the plots have not been reported, they are observed and analyzed.

3.1 Poisson Hurdle base model

Going back to the first model explained in details in Chapter 1, Section 1.1.1, and following the above explained procedure, we start with the choice of the hidden parameters and their prior distributions.
The current model consists in just two parametrs, p and λ; suppose that the hidden parameters are:

- p=0.5
- λ=4

The model that has been implemented in STAN is the following:

$$\begin{gathered} Y_i|p,\lambda \overset{\text{iid}}{\sim} HPois(p,\lambda) \qquad i \in \{1,2,...,N\} \\ p \sim Beta(2,5) \\ \lambda \sim Gamma(2,0.5) \end{gathered}$$

Let us motivade the choice of the prior distributions recalling the likelihood of the data:

$$\begin{aligned} L(p,\lambda|y,z) &= (1-p)^{N-\sum_{i=1}^N z_i} \cdot p^{\sum_{i=1}^N z_i} \cdot \prod_{i\in\Omega_1} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!(1-e^{-\lambda})} = \\ &= (1-p)^{N-\sum_{i=1}^N z_i} \cdot p^{\sum_{i=1}^N z_i} \cdot \prod_{i\in\Omega_1} \frac{1}{y_i!} \cdot \frac{\lambda^{\sum_{i=1}^N y_i} e^{-|\Omega_1|\lambda}}{(1-e^{-\lambda})^{|\Omega_1|}} \end{aligned} \tag{3.1}$$

Using the Bayes' theorem the joint posterior distribution of p and λ $\pi(p,\lambda|y,z) \propto \pi(p,\lambda,y,z) = L(y,z|p,\lambda)\pi(p)\pi(\lambda)$. Looking at the likelihood in (3.1), just to the term containing the parameter p, it is straightforward to recognize the kernel of a Beta distribution. Hence it is possible to choose a conjugate Beta prior Beta(a,b) for p, i.e. $\pi(p) \sim Beta(a,b)$ (due to the separability in p and λ of the likelihood); we get:

$$\pi(p|\cdot) \propto (1-p)^{N-\sum_{i=1}^N z_i+b-1} \cdot p^{\sum_{i=1}^N z_i+a-1}$$

where $\pi(p|\cdot)$ denotes a distribution where just the parameter p is considered not fixed. Hence the posterior distribution of p is in close form since it is a Beta$(\sum_{i=1}^N z_i + a, N - \sum_{i=1}^N z_i + b)$. The parameters of the prior distribution are updated adding the number of positive realizations $(\sum_{i=1}^N z_i)$ to the first parameter and the number of zero values $(N - \sum_{i=1}^N z_i)$ to the second parameter.

Looking at (3.1) for the choice of the prior of parameter λ, in the same way, it is possible to recognize the kernel of a Gamma distribution but in this case, after choosing a Gamma prior $\Gamma(\alpha,\beta)$ for λ, i.e. $\pi(\lambda) \sim Gamma(\alpha,\beta)$, the posterior is not in close form, because of the term $\frac{1}{(1-e^{-\lambda})^{|\Omega_1|}}$. Indeed it is given by:

$$\pi(\lambda|\cdot) \propto \frac{\lambda^{\sum_{i=1}^{N} y_i+\alpha-1} e^{-(|\Omega_1|+\beta)\lambda}}{(1-e^{-\lambda})^{|\Omega_1|}}$$

where in $\pi(\lambda|\cdot)$ we consider just λ as not fixed.
Even in a simple model like this we observe that the posterior distribution is not guaranteed to assume a close form. In these circumstances it is necessary to apply MCMC methods to sample from the posterior distribution.
With the STAN software we are able to sample from the posterior distribution. Hence, after running the model in STAN using a data sample of size 100, we can extract some descriptive statistics (reported in Table 3.1) and make inference about the posterior parameters:

Parameter	Mean	SD	Q0.025	Q0.25	Q0.5	Q0.75	Q0.975
p	0.4984	0.01121	0.4778	0.4909	0.4985	0.5058	0.5205
λ	3.9666	0.06542	3.8470	3.9197	3.9649	4.0093	4.1033

Table 3.1: Parameter β and λ summary statistics about posterior distribution.

Results' reliability is strengthened by acceptable traceplots for each parameter (see Figures 3.1 and 3.2), in fact the chains are mixed and without any seasonality, trend or cycle. The auto-correlation function ACF (Figure 3.5), which show the correlation ρ between a state of the chain and any previous state at a certain lag, does not show any peak and is extremely low in fact $|\rho| < 0.1$.
In Figure 3.3 and 3.4 the approximation of the posterior distributions are shown.

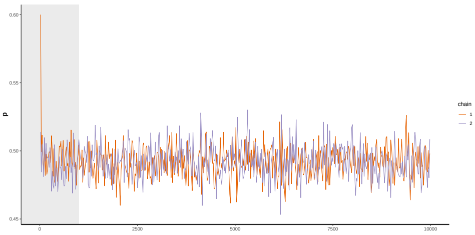

Figure 3.1: Traceplot of p parameter.

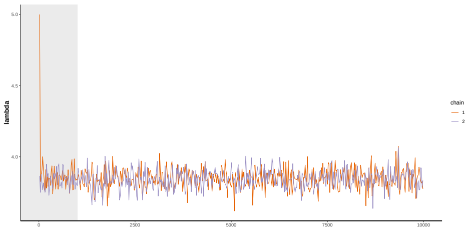

Figure 3.2: Traceplot of λ parameter.

For the sake of completeness we are interested in the computational workload which we investigate looking at the time spent on running the chains of model. The results are reported in Table 3.2.

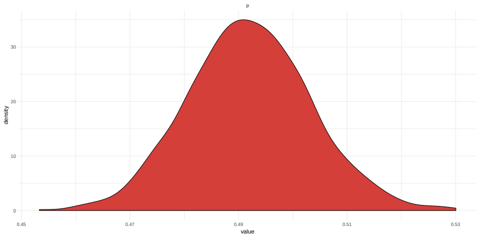

Figure 3.3: Posterior distribution of p parameter

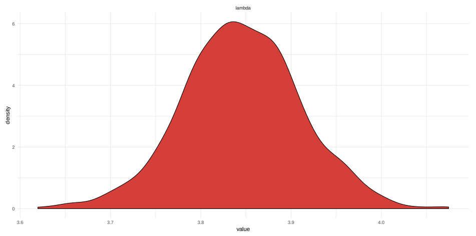

Figure 3.4: Posterior distribution of λ parameter.

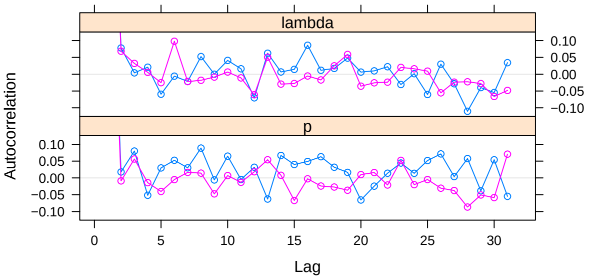

Figure 3.5: Auto-correlation function for the chains of p and λ parameters.

Number of data	Elapsed time Chain 1 [sec]	Elapsed time Chain 2 [sec]
1000	35.461	31.894
100	3.355	8.503
10	0.706	0.511

Table 3.2: Required computational time for different sample size.

The computational time required by the STAN implemented model is extremely low since just around 30 seconds are sufficient to run a chain of 10000 iterations (450 sample size per chain because of the warm-up of 1000 and thinning of 20) on a dataset of size 1000 observations.

The Mean Square Error, computed varying the data sample size, is reported in Table 3.3. We observe that increasing the sample size up to 1000 obser-

Number of data	MSE of p	MSE of λ
1000	0.00012	0.04397
100	0.00116	0.045714
10	0.03289	0.92078

Table 3.3: MSE of each parameter for different sample size.

vations makes the MSE decreasing. In fact an important property of the Bayesian assumption is that since the posterior distribution maximizes the chance of observing the given data according to our prior beliefs and data itself, when the sample size is large the posterior is largely affected and driven by the data and less by the prior distribution. The MSE behaviour that characterizes Table 3.3 is a typical trend that we are going to appreciate also for the next (more complex) models.

3.2 Poisson Hurdle Regression

In this Section we update the previous model introducing some independent variables in order to perform a regression model. We generated the response $Y_i \; \forall i \in \{1, \cdots, N\}$, through a user-defined function implemented in R (see appendix A), such that the synthetic data sample behaves as the model in Section 1.1.2. The choice of the model parameters is indifferent for our purpose, provided that it does not affect the result's stability. For example consider the following hidden parameters of dimension 3:

- $\beta^{(p)} = (0.2, 0.9, 0.2)$
- $\beta^{(\lambda)} = (1, 0.2, 0.7)$

For the current model it is also necessary to generate random covariates values $\{x_i\}_{i=1,\cdots,N}$ through R.
Given all these elements we are able to implement in STAN the following model:

$$
\begin{aligned}
Y_i | p_i, \lambda_i &\overset{\text{iid}}{\sim} HPois(p_i, \lambda_i) && i \in \{1, 2, ..., N\} \\
log(\tfrac{p_i}{1-p_i}) &= x_i^{(p)} \beta^{(p)} && i \in \{1, 2, ..., N\} \\
log(\lambda_i) &= x_i^{(\lambda)} \beta^{(\lambda)} && i \in \{1, 2, ..., N\} \\
\beta_j^{(p)} &\sim N(0, 9) && j \in \{1, 2, 3\} \\
\beta_j^{(\lambda)} &\sim N(0, 9) && j \in \{1, 2, 3\}
\end{aligned}
$$

Let us motivate the prior choices saying that they are all weakly informative. The single elements $\beta_j^{(\cdot)}$ for $j \in \{1, 2, 3\}$ are uncorrelated with mean 0 and a variance of 9, such that it is sufficiently large and little informative.

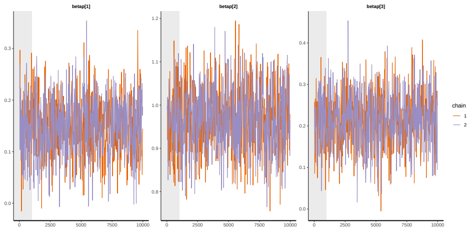

Figure 3.6: Traceplot of $\beta^{(p)}$ parameter.

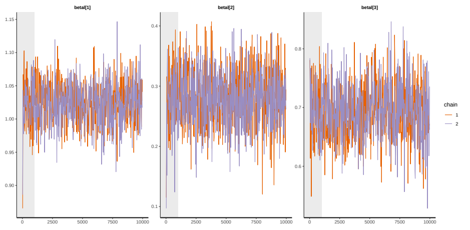

Figure 3.7: Traceplot of $\beta^{(\lambda)}$ parameter.

The chains are mixed and do not show cycles, seasonalities or trends.
Figures 3.8 and 3.9 show the posterior distribution of $\beta^{(p)}$ and $\beta^{(\lambda)}$ parameters and Table 3.4 collects their means and standard deviations.
The AFC plot in Figure 3.10 excludes any correlation between states of the chain at different lags since the maximum correlation in absolute value is

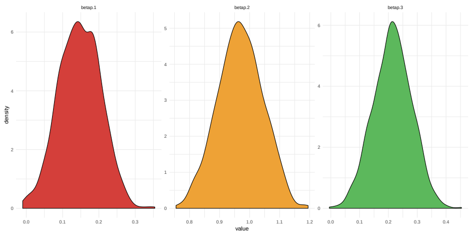

Figure 3.8: Posterior distribution of $\beta^{(p)}$ parameter.

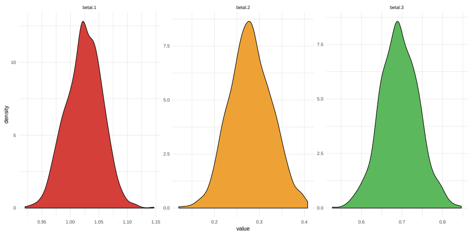

Figure 3.9: Posterior distribution of $\beta^{(\lambda)}$ parameter.

Parameter	Mean	SD	Parameter	Mean	SD
$\beta_1^{(p)}$	0.1472	0.05746	$\beta_1^{(\lambda)}$	1.0241	0.03155
$\beta_2^{(p)}$	0.9691	0.07466	$\beta_2^{(\lambda)}$	0.2803	0.04664
$\beta_3^{(p)}$	0.2139	0.06652	$\beta_3^{(\lambda)}$	0.6939	0.04705

Table 3.4: Parameters summary statistics.

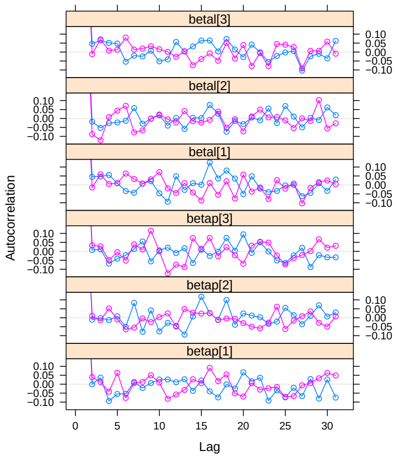

Figure 3.10: Auto-correlation function for the chains of βs and λs parameters.

smaller than 0.1.

Computationally speaking the elapsed time for running the chains are showed in Table 3.5.

The computational cost synthesized by the occurred time to run the chains of the model is larger than the previous model (around 4-5 times larger) but still characterized by an acceptable order of magnitude. This is an expected result since we have introduced the regression part.

Number of data	Elapsed time Chain 1 [sec]	Elapsed time Chain 2 [sec]
1000	145.554	130.102
100	14.699	16.622
10	5.692	5.509

Table 3.5: Required computational time for different sample size.

Number of data	MSE of $\beta_1^{(p)}$	MSE of $\beta_2^{(p)}$	MSE of $\beta_3^{(p)}$
1000	0.00357	0.00639	0.00585
100	0.12912	0.08826	0.05151
10	9.75035	10.35407	6.17989
Number of data	MSE of $\beta_1^{(\lambda)}$	MSE of $\beta_2^{(\lambda)}$	MSE of $\beta_3^{(\lambda)}$
1000	0.00098	0.01359	0.00542
100	0.13266	0.32287	0.02766
10	0.24860	0.70593	0.17076

Table 3.6: MSE of each parameter for different sample size.

The MSEs reported in Table 3.6 decrease as data sample size increases for all the parameters, because the posterior is more and more driven by the likelihood.

3.3 Poisson Hurdle with group structure

Consider now the model in Chapter 1, Section 1.1.3 and its notation. As widely described the corresponding section, a group structure is introduced among observations: each group has its own parameters that could vary from one group to another. This condition is a natural assumption since there is no constrain in real cases to assume the opposite, as long as the the fact that the groups could have different cardinalities.
Consider the case where the hidden parameters are the following:

- n=(130,70,95,105,145,55,80,120,65,135)
- p=(0.1,0.4,0.3,0.7,0.6,0.8,0.2,0.5,0.9,0.25)
- λ = (3,5,5,7,2,8,4,6,9,5)

The vector n contains the cardinality of each group (using the same notation as in Section 1.1.3 the single component k coincides with n_k for every $k \in \{1, 2, \dots, K\}$) for a total number of groups of $K = 10$. The p and λ parameters are vectors that preserve the order observed in n in such a way that the first elements p_1 and λ_1 are the parameters referred to the first group n_1, the second ones to the second group and so on. The groups cardinality, given by vector n, is shown in Figure 3.11 and have been chosen in order to have heterogeneous groups (the smallest one has size 65, the largest 145).

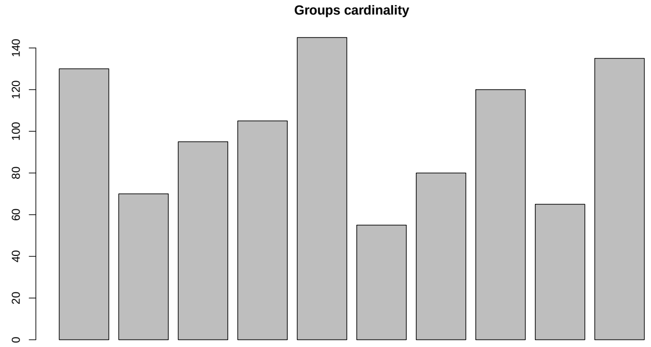

Figure 3.11: Groups cardinality (given by n).

The model implemented in STAN is:

$$\begin{aligned}
Y_{i,k}|p_k, \lambda_k &\overset{\text{iid}}{\sim} HPois(p_k, \lambda_k) && i \in \{1, 2, ..., n_k\},\;\; k \in \{1, 2, ..., 10\} \\
p_j &\sim Beta(2, 5) && j \in \{1, 2, ..., 10\} \\
\lambda_j &\sim Gamma(2, 0.5) && j \in \{1, 2, ..., 10\}
\end{aligned}$$

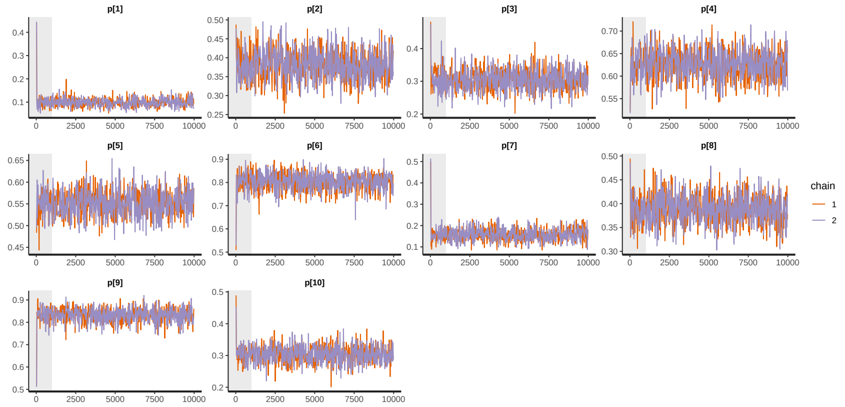

Figure 3.12: Traceplots of p parameter's components.

We have choosen a prior distribution for each parameter and we have assumed that the components of p and λ are uncorrelated.

The traceplots are mixed, without seasonalities, trends and cycles, as shown in Figures 3.12 and 3.13.

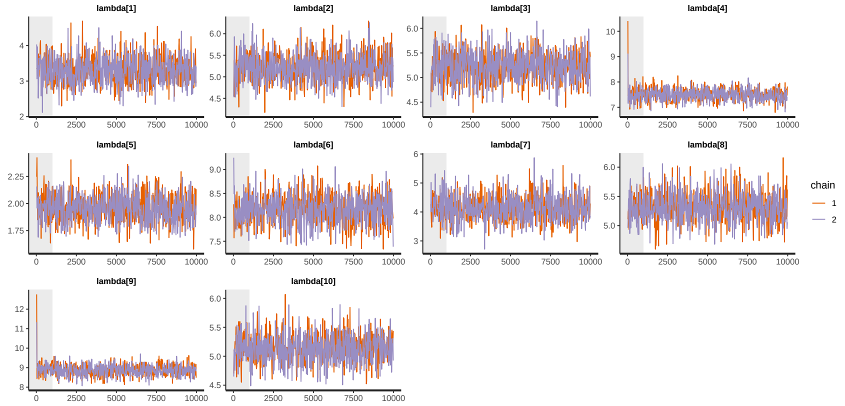

Figure 3.13: Traceplots of λ parameter's components.

From Figures 3.14 it is possible to observe that when the "true" parameter is high, for example $p_4 = 0.7$, the mean of the posterior distribution is "pulled" towards low values because the posterior distribution (when the data sample size is not so large) is driven by the prior Beta(2,5) distribution which has a mean of $\frac{2}{7}$, quite lower than 0.7. Instead in Figure 3.15 are reported the λ parameters.

The AFC plot is not reported but for each parameter the condition of low autocorrelation between the states of the chains is satisfied.

In Table 3.7 it is possible to see the required time for running a single chain

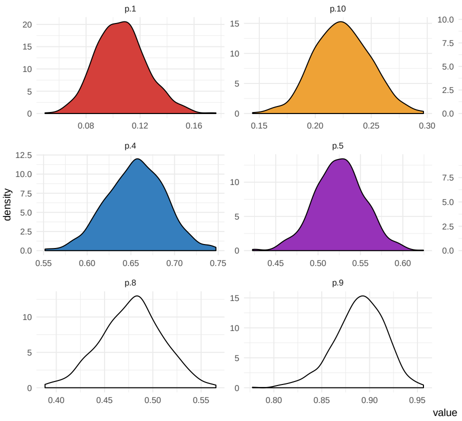

Figure 3.14: Posterior distribution of p parameters.

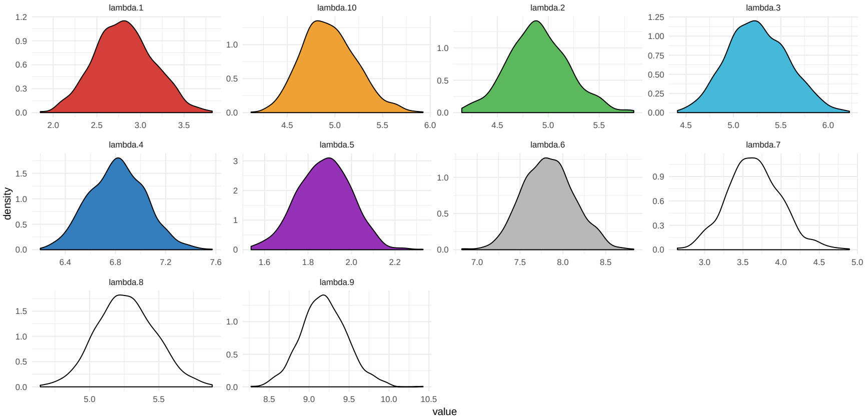

Figure 3.15: Posterior distribution of λ parameters.

Number of data	Number of groups	Elapsed time Ch.1 [sec]	Elapsed time Ch.2 [sec]
10000	10	1098.58	1211.48
1000	10	100.229	101.281
100	10	8.533	8.55

Table 3.7: Required computational time for different sample size.

of the model when the number of groups is kept fixed at 10. Increasing by an order of magnitude the data sample size and passing from 10^3 to 10^4, the time required to run one chain increases by an order of magnitude too.
With the inclusion of 10000 sample size is possible to appreciate that the parameters' MSE decreases because the data sample size for each group increases.

Number of data	MSE of β_1	MSE of β_2	MSE of β_3	MSE of β_4
10000	4.749150e-05	1.651637e-04	1.866497e-04	9.654945e-05
1000	0.00044	0.00159	0.00157	0.00286
100	0.01244	0.01604	0.01440	0.01865
Number of data	MSE of β_5	MSE of β_6	MSE of β_7	MSE of β_8
10000	1.651999e-04	1.583563e-04	4.292488e-04	1.454318e-04
1000	0.00152	0.00145	0.01027	0.00107
100	0.09735	0.16929	0.00761	0.06313
Number of data	MSE of β_9	MSE of β_{10}	MSE of λ_1	MSE of λ_2
10000	9.979544e-05	6.279057e-04	2.627937e-02	9.526981e-03
1000	0.00121	0.00639	0.15395	0.84043
100	0.02958	0.01228	1.89052	0.58129
Number of data	MSE of λ_3	MSE of λ_4	MSE of λ_5	MSE of λ_6
10000	7.975336e-03	6.520310e-02	1.279396e-02	3.326599e-02
1000	0.58221	0.05630	0.04143	0.10790
100	15.51225	1.87481	0.21697	2.80244
Number of data	MSE of λ_7	MSE of λ_8	MSE of λ_9	MSE of λ_{10}
10000	7.902756e-02	3.690021e-02	9.836123e-03	2.638838e-02
1000	0.2188	0.17999	0.14807	0.86464
100	4.24819	1.10936	0.61259	3.03410

Table 3.8: MSE of each parameter for different sample size.

3.4 Poisson Hurdle Regression with group structure

We continue with the model described in Section 1.1.4 of Chapter 1 which mixes the regression formulation and groups structure.
As done before the general procedure forces to generate some data through a user-defined function (see Appendix A) and to adopt a prior distribution for each parameter. Consider the case where the hidden parameters are the following:

- $n = (250, 350, 125, 275)$;
- $\beta^{(p)} = (1, 0.05, 0.03)$;
- $\beta^{(\lambda)} = (1, 0.06, 0.07)$;
- $\theta^{(p)} = (0.1, 0.3, 0.5, 0.7)$;
- $\theta^{(\lambda)} = (0.8, 0.6, 0.4, 0.2)$;

Using the same notation of the previous section we have $K = 4$ total number of groups of different cardinality. $\beta^{(\cdot)}$ parameter has $J = 3$ components but a different choice is possible; vice versa since there are four groups $\theta^{(\cdot)}$ parameter has 4 components, one for each different group.
The implemented STAN model is:

$$\begin{array}{ll} Y_{i,k}|p_{i,k}, \lambda_{i,k} \sim HPois(p_{i,k}, \lambda_{i,k}) & i \in \{1, 2, ..., n_k\},\ k \in \{1, 2, 3, 4\} \\ log(\frac{p_{i,k}}{1-p_{i,k}}) = x^{(p)}_{i,k}\beta^{(p)} + \hat{x}^{(p)}_k\theta^{(p)} & i \in \{1, 2, ..., n_k\},\ k \in \{1, 2, 3, 4\} \\ log(\lambda_{i,k}) = x^{(\lambda)}_{i,k}\beta^{(\lambda)} + \hat{x}^{(\lambda)}_k\theta^{(\lambda)} & i \in \{1, 2, ..., n_k\},\ k \in \{1, 2, 3, 4\} \\ \beta^{(p)}_j \sim N(0, 9) & j \in \{1, 2, 3\} \\ \beta^{(\lambda)}_j \sim N(0, 9) & j \in \{1, 2, 3\} \\ \theta^{(p)}_k \sim N(0, 9) & k \in \{1, 2, 3, 4\} \\ \theta^{(\lambda)}_k \sim N(0, 9) & k \in \{1, 2, 3, 4\} \end{array}$$

As prior distribution for β parameter we choose a normal distribution centered in 0. The variance of $\beta^{(\cdot)}$ and $\theta^{(\cdot)}$ are kept large (both 9), to guarantee a diffuse prior distribution. In the current example we decided to take as $\hat{x}^{(p)}_k$ and $\hat{x}^{(\lambda)}_k$ a vector whose component are all zeros with the exception of the position relative to the observation's group that takes value 1, i.e. (0,0,1,0) for an observation belonging to group 3.

Diagnostics starts observing the traceplots of all the model parameters, reported in Figures 3.16, 3.17, 3.18 and 3.19, which are all mixed, without any cycle, trend and seasonality.

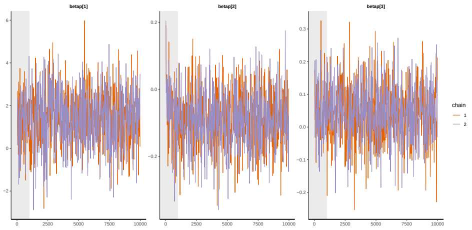

Figure 3.16: Traceplot of $\beta^{(p)}$ parameter.

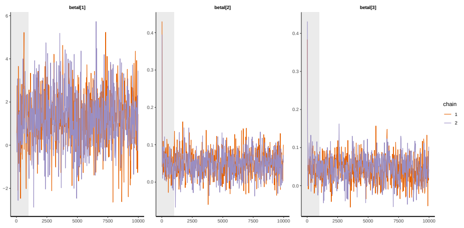

Figure 3.17: Traceplot of $\beta^{(\lambda)}$ parameter.

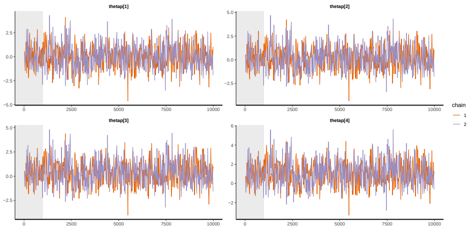

Figure 3.18: Traceplot of $\theta^{(p)}$ parameter.

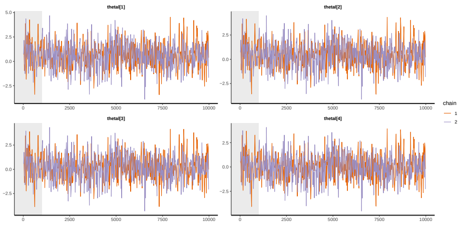

Figure 3.19: Traceplot of $\theta^{(\lambda)}$ parameter.

The summary statistics about the parameters' posterior distributions are reported in Table 3.9.

Parameter	Mean	SD	Parameter	Mean	SD
$\beta_1^{(p)}$	1.09652	0.098332	$\theta_1^{(p)}$	0.11661	0.990282
$\beta_2^{(p)}$	0.04147	0.059114	$\theta_2^{(p)}$	0.24932	0.981004
$\beta_3^{(p)}$	0.01944	0.052958	$\theta_3^{(p)}$	0.47296	1.001548
$\beta_1^{(\lambda)}$	1.21032	0.098447	$\theta_4^{(p)}$	0.61525	1.005025
$\beta_2^{(\lambda)}$	0.06108	0.009317	$\theta_1^{(\lambda)}$	0.68121	0.982752
$\beta_3^{(\lambda)}$	0.07213	0.009162	$\theta_2^{(\lambda)}$	0.52177	0.987242
			$\theta_3^{(\lambda)}$	0.33365	0.974984
			$\theta_4^{(\lambda)}$	0.13154	0.962914

Table 3.9: Posterior distributions' summary statistics.

From a qualitative and quick observation, the posterior distributions' estimates of $\beta^{(\cdot)}$ and $\theta^{(\cdot)}$ parameters, in terms of their means, recover the "true" parameters.

The auto-correlation functions, which are not reported due to the large number of parameters, do not show any significant correlation between the states of the chain up to a lag of 30 (all the parameters have an auto-correlation

$|\rho| < 0.1$).
The required computational time is reported in Table 3.10.

Number of data	Number of groups	Elapsed time Ch.1 [sec]	Elapsed time Ch.2 [sec]
1000	4	10218	17316.9
100	4	380.6	450.1

Table 3.10: Required computational time for different sample size.

The order of magnitude, for a data sample size of 1000 observations, is around 10^4 seconds (10 times slower than model is Subsection 3.3 and 100 times slower than model in 3.2).
Finally, the MSE is reported in Table 3.11.

Number of data	MSE of $\beta_1^{(p)}$	MSE of $\beta_2^{(p)}$	MSE of $\beta_3^{(p)}$	
1000	0.09541	0.00353	0.00429	
100	1.75568	0.80552	0.16465	
Number of data	MSE of $\beta_1^{(\lambda)}$	MSE of $\beta_2^{(\lambda)}$	MSE of $\beta_3^{(\lambda)}$	
1000	0.10822	0.00073	0.00042	
100	1.51221	0.04644	0.06307	
Number of data	MSE of $\theta_1^{(p)}$	MSE of $\theta_2^{(p)}$	MSE of $\theta_3^{(p)}$	MSE of $\theta_4^{(p)}$
1000	0.19640	0.18258	0.18892	0.12458
100	1.84794	2.0868	1.86553	2.82616
Number of data	MSE of $\theta_1^{(\lambda)}$	MSE of $\theta_2^{(\lambda)}$	MSE of $\theta_3^{(\lambda)}$	MSE of $\theta_4^{(\lambda)}$
1000	0.15317	0.11076	0.14468	0.10599
100	2.09258	2.0224	2.05052	1.9646

Table 3.11: MSE of each parameter for different sample size.

Chapter 4

Analysis of a real dataset

4.1 Dataset

The dataset has been proposed for the first time by Cortez and Morais [13] and deals with the well known and one of the major environmental issue of our epoch: forest fires (also called wildfires). Considering the negative consequences that wildfires can bring, from an ecological havoc itself to the economic waste of resources to regenerate the area, it is challenging to investigate if there exist some connections among areas that may lead us to infer the high or low risk of wildfire of specific regions. In addition to wildfires data itself, the dataset provides also meteorological and geological data that will be included in the analysis because if we exclude the human direct responsibility they are the most intuitive and sensible marker to take in consideration.

Every year millions of forest hectares (where 1 $ha = 10^4\ m^2$) are destroyed all around the world. Portugal, due to position and weather conditions, is highly affected by forest fires. From 1980 to 2005, over 2.7 million hectares of forest area have been destroyed. In particular the 2003 and 2005 fire seasons were especially dramatic, affecting 4.6% and 3.1% of the territory. The area of our interest is the Montesinho natural park, in the Northeast region of Portugal (Figure 4.1).

The data used in this elaborate has been collected in a time interval of 4 years, from January 2000 to December 2003. It was built using two sources and before entering into the detail of the dataset, we need to explain what the two sources consist of.

The first source is the forest Fire Weather Index (FWI) which is the Canadian system used for rating fires danger [14]. In particular it is an indicator such that high values suggest severe burning conditions. The index consists in six components: Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI) and FWI [15]. They jointly contribute to provide an index of the fire

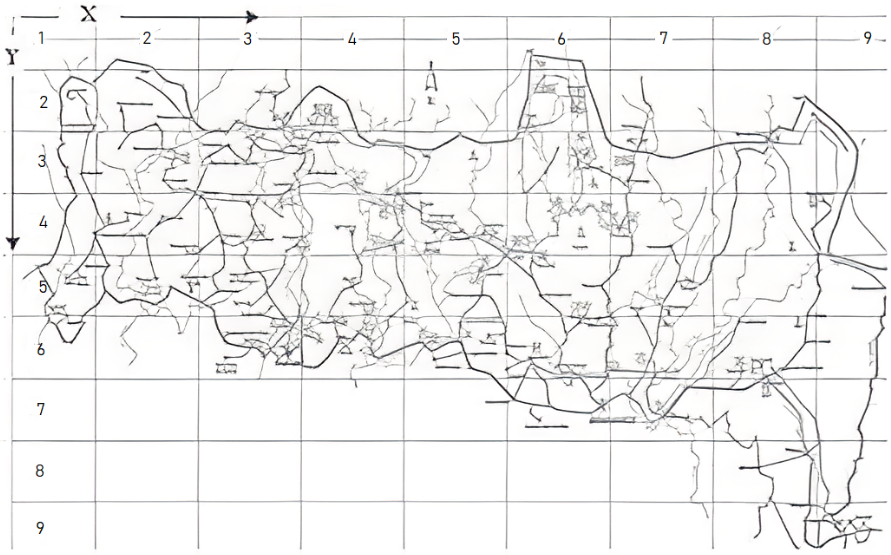

Figure 4.1: Montesinho Natural park map.

intensity. However we are not interested in the mechanism to produce the index because we will consider the components one by one. Consider the first four components: according to the FWI theory FFMC denotes the numeric rating of the moisture content surface litter (an indicator of the flammability of the surrounding fine fuel), while the DMC and DC represent the numeric rating of moisture content of shallow and deep organic layers (this code gives an indication of fuel consumption in moderate layers). In the end the ISI is a score that correlates with expected rate of fire velocity spread (it is the influence of wind speed on fine surface). Moreover these indexes incorporate a memory (in terms of a time lag) of past weather conditions: 16 hours for FFMC, 12 days for DMC and 52 days for DC.
Another type of data come from a different source: it has been collected by the Bragança Polytechnic Institute with automatic meteorological stations that are often available in real time. The data contains weather observations (temperature, relative humidity, wind speed and rain). Unlike the FWI that includes time lags, in this case each value stands for an instant of time, recorded by the station sensors when the fire was detected (the only exception is the rain variable, which denotes the accumulated precipitation within the previous 30 minutes).
Other information about the wildfire like where and when it has been detected, as long as the amount of burned hectares, completes the dataset. Intuitively every time a forest fire breaks out, the date and the spatial location (within a 9×9 grid) are transcribed.
In Table 4.1 a summary of the quantities involved in the dataset is shown.

X	x-axis coordinate (from 1 to 9)
Y	y-axis coordinate (from 1 to 9)
month	Month of the year (January to December)
day	day of the week (Monday to Sunday)
FFMC	Fine Fuel Moisture Code
DMC	Duff Moisture Code
DC	Drought Code
ISI	Initial Spread Index
temp	Outside temperature (in °C)
RH	Outside relative humidity (in %)
wind	Outside wind speed (in km/h)
rain	Outside rain (in mm/m^2)
area	Total burned area (in ha)

Table 4.1: Summary of dataset attributes.

4.2 Descriptive statistics

The dataset consists in 4 categorical (X, Y, *Day* and $Month$) and 9 numerical ($FFMC$, DMC, DC, ISI, *Temperature*, *Relative humidity*, $Wind$ *intensity*, *Rain amount* and *Burned area*) attributes. The first six observations are reported in Table 4.2.

X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0.0	0
7	4	oct	tue	90.6	35.4	669.1	6.7	18.0	33	0.9	0.0	0
7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0.0	0
8	6	mar	fri	91.7	33.3	77.5	9	8.3	97	4	0.2	0
8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0.0	0
8	6	aug	sun	92.3	85.3	488	14.7	22.2	29	5.4	0.0	0

Table 4.2: Dataset details.

The most interesting variable, represented by the hectares of burned area, is expressed in ha, (1 ha= 10^4 m^2). Since the unit of measure is expressed in hectares the majority of the data are concentrated near zero. In particular the percentage of exactly zero values is 47,8% while considering the values that do not reach the measure of 1 ha the percentage raises to 53%. With this percentage the assumption of overabundance of zero is satisfied; to have a close look of the zeros and the distribution of positive values of burned hectares see Figure 4.2.

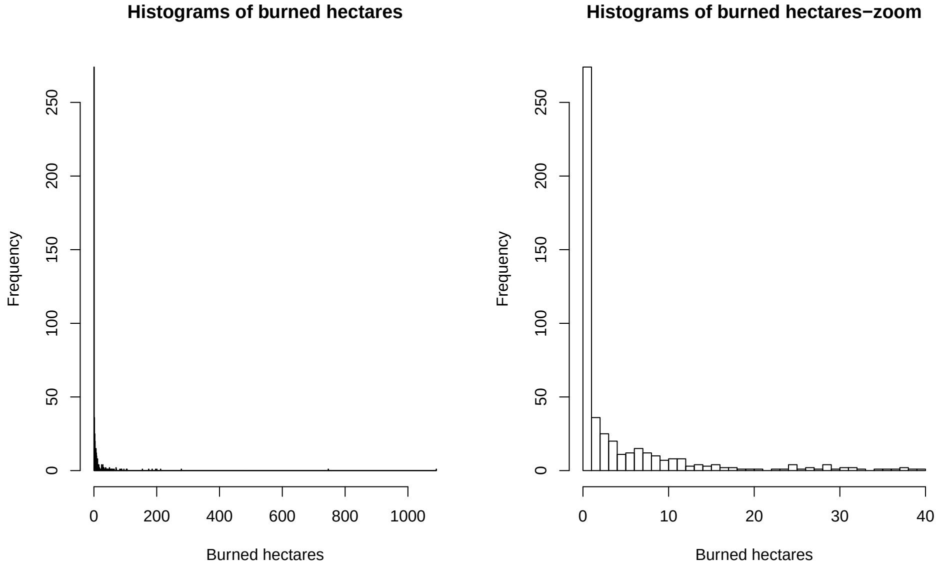

Figure 4.2: Burned hectares (all observations to the left, 96% of the observations to the right).

As said before the whole national park can be divided in sub-regions of equal extension, introducing a coordinate system and a grid 9×9 (see Figure 4.1). Hence data can be grouped according to their location over the entire area and it is possible to have a look at the distribution of burned hectares in each region. Note that since data are collected over four years each sub-region encloses measurements detected over that temporal span and not necessarily in the same year. Made this clarification we can have a look at the distribution of the amount of burned territories caused by the wildfire, for each area of the map (Figures 4.3, 4.4 and 4.5). In these figures and in some other analysis later on, the sub-regions of the map might be indicated with progressive numbers from 1 to 36; we use this convenient numeration to avoid to report the geographical coordinates X and Y (also because there are empty sub-regions where no data has been recorded). This numeration proceeds scanning the map in Figure 4.1 from North to South and from West to East (like a scanning by column from left to right).
Another useful and interesting plot is Figure 4.6, where the number of measurements per sub-region is represented. We notice that the number of measurements has a large variability denoting that exist areas more subjected to possible wildfires than others. Furthermore some regions have been characterised by just one broken wildfire in four years and other areas where many more measurements (even over 50) have been taken. In detail, for the

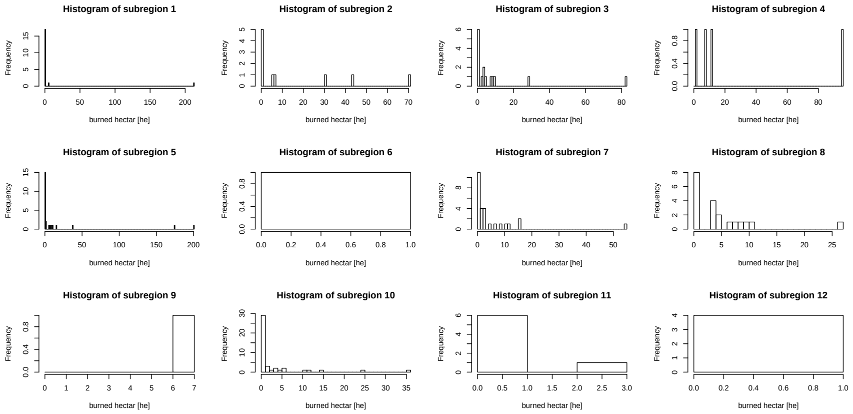

Figure 4.3: Burned area distribution of the first 12 sub-regions of the map (corresponding to areas such that $X \in \{1, 2, 3\}$ and $Y \in \{2, 3, 4, 5\}$). It is the West territory of the park.

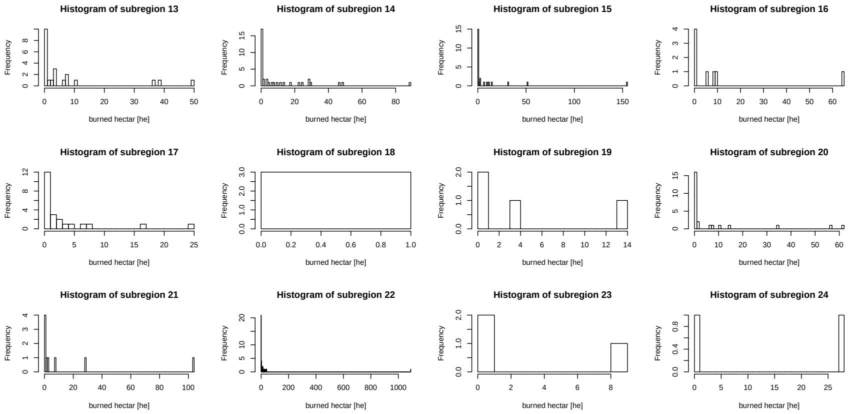

Figure 4.4: Burned area distribution of the 13-24th sub-regions of the map (corresponding to areas such that $X \in \{4, 5, 6\}$ and $Y \in \{3, 4, 5, 6\}$ plus region (7,3)). It is the middle territory of the park.

analysis, consider the fact that considering all the 517 data (subdivided in 36 groups), there are: 5 groups of just 1 measurement, 3 groups of 2 measurements, 3 groups of 3 measurements and 5 groups of 4 measurement.
Before applying any model to the dataset we show some boxplots of the quantities described so far, which will be used as covariates. They are organized by sub-regions and are reported in Figure 4.7 and 4.8.
The quantities that have the smallest variability on data are $FFMC$ and $rain$, where we can state that it is extremely rare to have some millimeters of rain in the 30 minutes before the outbreak of a fire.
All the quantities have more or less a quite symmetrical distribution with

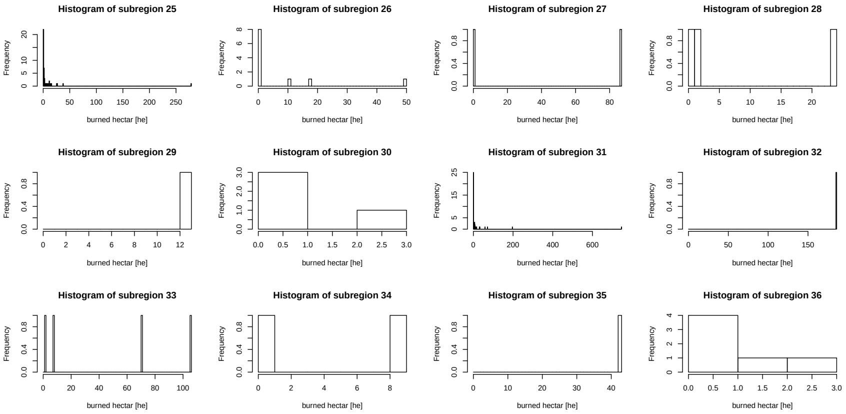

Figure 4.5: Burned area distribution of the 25-36th sub-regions of the map (corresponding to areas such that $X \in \{7, 8, 9\}$ and $Y \in \{3, 4, 5, 6, 7, 8, 9\}$). It is the East territory of the park.

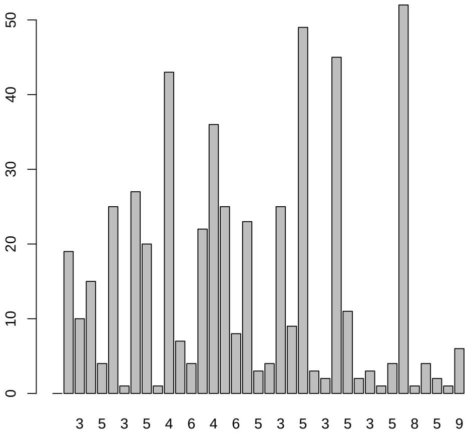

Figure 4.6: Cardinality of observations of each region/group (denoted by n_k).

the exception of $FFMC$ which have more extreme values toward low values. In general looking at all the quantities involved we can assert from a qualitative point of vie that there is no clear distinction between some groups and others.

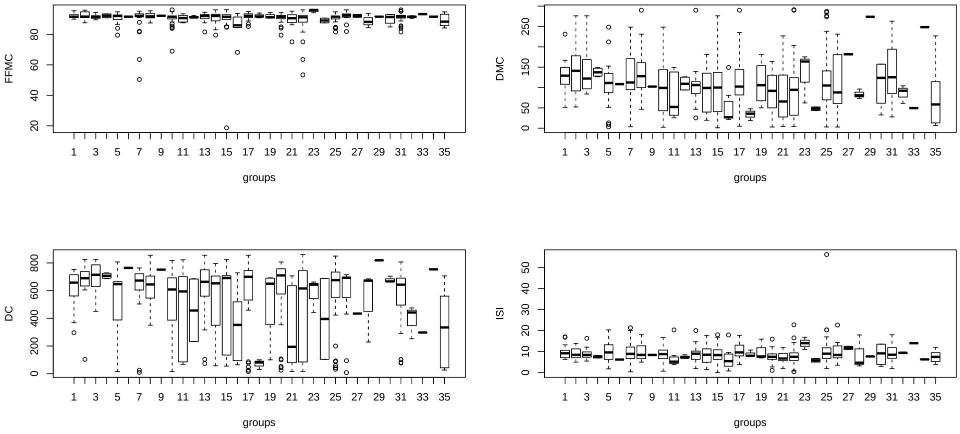

Figure 4.7: Boxplots of $FFMC$, DMC, DC and ISI quantities.

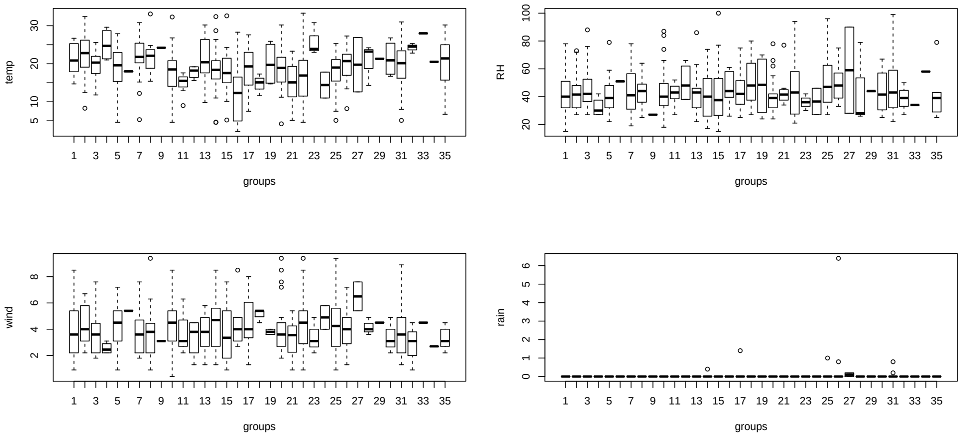

Figure 4.8: Boxplots of $Temperature$, $Relative\ humidity$, $Wind\ intensity$ and $Rain\ amount$ quantities.

4.3 The Poisson hurdle model

4.3.1 Model setting

The above described dataset is suitable for the implementation of a Poisson hurdle generalized linear model with group structure and therefore we have chosen to adopt the model in Section 1.1.4 for our analysis. In particular consider the burned area attribute of the dataset, whose unit of measure is *hectares*, as the dependent variable Y; in order to make this data appropriate to our model, Y is converted from real numbers to natural numbers with an operation of rounding down to the nearest integer number. Then, since data is already grouped in sub-regions according to its spatial coordinates (look at attributes X and Y in Table 4.2 and recover Figure 4.1 to see the display of the groups partition), we have aggregated each instance of the dataset according to this criterion. We end up having 517 instances divided in 36 groups.

The dataset is then reduced from 517 number of data to 506 (11 observations are not considered for the analysis) due to numerical stability issue and groups/sub-regions becomes 35 since the region $(X = 8, Y = 8)$ bottom right in Figure 4.1 disappears, in the sense that there is no longer any observation in this area.
To have a look at the wildfires distribution over the entire park observe Figure 4.9. Each point represent the percentage of zero burned hectares in each

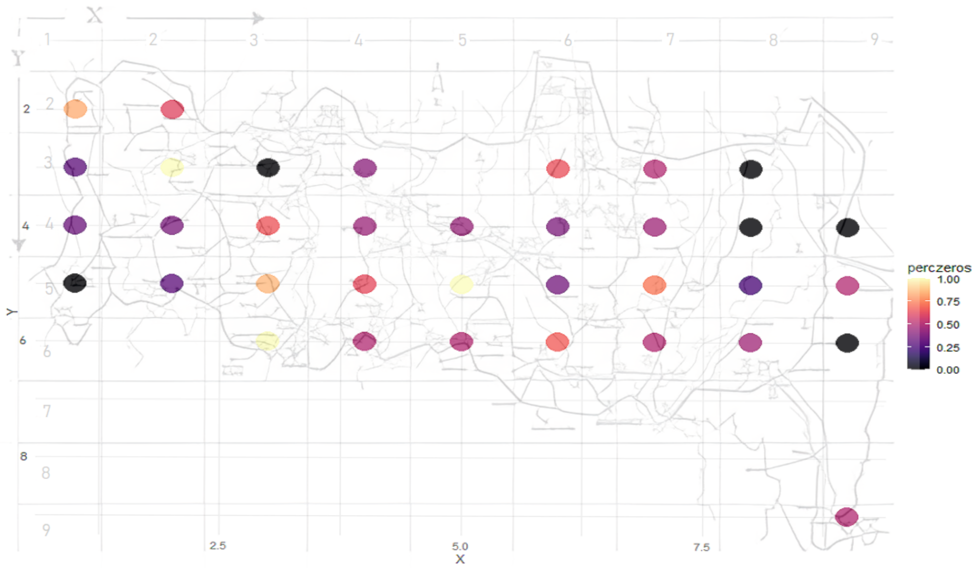

Figure 4.9: Percentage of zeros (zero burned hectare) for each sub-region of the map.

area of the map represented in Figure 4.1. In the details the percentages of zeros in each sub-region are respectively 0.83, 0.3, 0.33, 0, 0.61, 1, 0.37, 0.3, 0, 0.65, 0.86, 1, 0.41, 0.44, 0.62, 0.5, 0.43, 1, 0.5, 0.64, 0.37, 0.35, 0.67, 0.5, 0.45, 0.73, 0.5, 0, 0, 0.25, 0.46, 0, 0.5, 0, 0.5 (considering the progressive numeration). The maximum value is 1 and denotes a sub-region full of zeros, where no wildfires broke out (like it happens for sub-regions 6, 12 and 18 represented by light dots). Looking at the figure in a quick way, the extreme left and right regions of map, corresponding to the extreme West and East regions of the park seem to be the areas most affected by wildfires. Note that this plot gives indication of the zones most affected by the fires in terms of number of broken wildfires, but do not give any indication about the intensity of the aforementioned fires: in principle there could be areas where the weather mitigates the spread of severe fires and other areas where even a contained number of fires encounter suitable climatic conditions for growth.

We have chosen to keep all the geothermal and meteorological attributes described in the previous section as covariates (*FFMC, DMC, DC, ISI, Temperature, Relative Humidity, Wind and Rain*), with the addition of the temporal covariate $Month$, while Day is not considered (notice that the categorical variable $Month$ depends on the single observation of the dataset).
Made these assumptions, we have implemented in STAN the following model with the following settings:

$$
\begin{array}{ll}
Y_{i,k}|p_{i,k},\lambda_{i,k} \sim HPois(p_{i,k},\lambda_{i,k}) & i \in \{1,2,\dots,n_k\}, k \in \{1,2,\dots,35\} \\
\log\Big(\frac{p_{i,k}}{1-p_{i,k}}\Big) = x^{(p)}_{i,k}\beta^{(p)}+\tilde{x}_{i,k}\phi^{(p)}+\hat{x}^{(p)}_{k}\theta^{(p)} & i \in \{1,2,\dots,n_k\}, k \in \{1,2,\dots,35\} \\
\log(\lambda_{i,k}) = x^{(\lambda)}_{i,k}\beta^{(\lambda)}+\tilde{x}_{i,k}\phi^{(\lambda)}+\hat{x}^{(\lambda)}_{k}\theta^{(\lambda)} & i \in \{1,2,\dots,n_k\}, k \in \{1,2,\dots,35\} \\
\beta^{(p)}_{j} \sim N(0,9) & j \in \{1,2,\dots,9\} \\
\beta^{(\lambda)}_{j} \sim N(0,9) & j \in \{1,2,\dots,9\} \\
\phi^{(p)}_{h} \sim N(0,9) & h \in \{1,2,\dots,12\} \\
\phi^{(\lambda)}_{h} \sim N(0,9) & h \in \{1,2,\dots,12\} \\
\theta^{(p)}_{k} \sim N(0,9) & k \in \{1,2,\dots,35\} \\
\theta^{(\lambda)}_{k} \sim N(0,9) & k \in \{1,2,\dots,35\}(4.1)
\end{array}
$$

The response represents the hectares of burnt forest and is distributed as a Poisson hurdle with parameters p and λ. The covariates used in the regression of the two parameters are the same because even if we can guess that some covariates are intuitively linked to p parameter and others to λ (in fact given the Poisson hurdle structure, p rules just the probability that a fire breaks out and λ rules the intensity of the broken fire in terms of burned hectares), we have chosen a more conservative approach, so $x^{(p)}_{i,k} \equiv x^{(\lambda)}_{i,k}$, $\tilde{x}^{(p)}_{i,k} \equiv \tilde{x}^{(\lambda)}_{i,k}$ and $\hat{x}^{(p)}_{k} \equiv \hat{x}^{(\lambda)}_{k}$ $\forall i \in \{1,2,...,n_k\}$, $\forall k \in \{1,2,...,35\}$. The group covariates $\hat{x}^{(p)}_{k}$ and $\hat{x}^{(\lambda)}_{k}$ are vectors $[0,\dots,0,1,0,\dots,0]^T$, where a 1 is put in correspondence of the group $k \in \{1,\dots,35\}$ in which the observation i belongs. As long as $\hat{x}^{(\cdot)}_{k}$, $\tilde{x}^{(p)}_{i,k}$ and $\tilde{x}^{(\lambda)}_{i,k}$ are vectors $[0,\dots,0,1,0,\dots,0]^T$ where a 1 is put in correspondence of the month (keeping the conventional sorting, so a 1 in position 1 indicates *January* and a 1 in position 12 refers to *December*). Briefly $x^{(\cdot)}_{i,k} \in \mathbb{R}^9$ (8 covariates plus the intercept), $\hat{x}^{(\cdot)}_{k} \in \{0,1\}^{35}$ and $\tilde{x}^{(\cdot)}_{i,k} \in \{0,1\}^{12}$. Then each single observation is collected row-wise in the matrices $X^{(p)} \equiv X^{(\lambda)} \in \mathbb{R}^{506\times 9}$, $\hat{X}^{(p)} \equiv \hat{X}^{(\lambda)} \in \{0,1\}^{506\times 35}$ and $\tilde{X}^{(p)} \equiv \tilde{X}^{(\lambda)} \in \{0,1\}^{506\times 12}$.

4.3.2 Prior distributions' choice

The prior distributions are Gaussian vectors, with zero mean and diagonal covariance matrix with a homogeneous variance for each single component of 9; they are all not so informative priors because of the pretty large variance. We recover the likelihood of the data from equation (1.13) of Section 1.1.4 (note that in (1.13) does not appear the $\phi^{(\cdot)}$ parameter because we had not included categorical variables, which here, instead, are considered), using ρ to refer to the collection of all the parameters, i.e. $\rho = (\beta^{(p)}, \beta^{(\lambda)}, \theta^{(p)}, \theta^{(\lambda)}, \phi^{(p)}, \phi^{(\lambda)})$ and using $\mathcal{D} = \{y, z, X^{(p)}, X^{(\lambda)}, \hat{X}^{(p)}, \hat{X}^{(\lambda)}, \tilde{X}^{(p)}, \tilde{X}^{(\lambda)}\}$ for the collection of the data. The likelihood becomes:

$$
\begin{aligned}
&L(\rho|\mathcal{D}) = L(\beta^{(p)}, \beta^{(\lambda)}, \theta^{(p)}, \theta^{(\lambda)}, \phi^{(p)}, \phi^{(\lambda)}|y, z, X^{(p)}, X^{(\lambda)}, \hat{X}^{(p)}, \hat{X}^{(\lambda)}, \tilde{X}^{(p)}, \tilde{X}^{(\lambda)}) = \\
&= \prod_{k=1}^{K}\prod_{i=1}^{n_k}\Big(1 - p_{i,k}(x_{i,k}^{(p)}\beta^{(p)} + \tilde{x}_{i,k}^{(p)}\phi^{(p)} + \hat{x}_{k}^{(p)}\theta^{(p)})\Big)^{1-z_{i,k}} \times \\
&\Big(p_{i,k}(x_{i,k}^{(p)}\beta^{(p)} + \tilde{x}_{i,k}^{(p)}\phi^{(p)} + \hat{x}_{k}^{(p)}\theta^{(p)})\Big)^{z_{i,k}} \times \\
&\Big(\frac{\lambda_{i,k}(x_{i,k}^{(\lambda)}\beta^{(\lambda)} + \tilde{x}_{i,k}^{(\lambda)}\phi^{(\lambda)} + \hat{x}_{k}^{(\lambda)}\theta^{(\lambda)})^{y_{i,k}}}{y_{i,k}!(1 - e^{-\lambda_{i,k}(x_{i,k}^{(\lambda)}\beta^{(\lambda)}+\tilde{x}_{i,k}^{(\lambda)}\phi^{(\lambda)}+\hat{x}_{k}^{(\lambda)}\theta^{(\lambda)})})} e^{-\lambda_{i,k}(x_{i,k}^{(\lambda)}\beta^{(\lambda)}+\tilde{x}_{i,k}^{(\lambda)}\phi^{(\lambda)}+\hat{x}_{k}^{(\lambda)}\theta^{(\lambda)})}\Big)^{z_{i,k}}
\end{aligned}
$$

In the formulation above the parameters $\beta^{(p)}, \beta^{(\lambda)}, \theta^{(p)}$, $\theta^{(\lambda)}$, $\phi^{(p)}$ and $\phi^{(\lambda)}$ are expressed as a function of $p(\cdot)$ and $\lambda(\cdot)$. However the implemented model prescribes to choose as link function for p and λ the $logit(p)$ and the $log(\lambda)$, so the likelihood becomes:

$$
\begin{aligned}
&L(\rho|\mathcal{D}) = L(\beta^{(p)}, \beta^{(\lambda)}, \theta^{(p)}, \theta^{(\lambda)}, \phi^{(p)}, \phi^{(\lambda)}|y, z, X^{(p)}, X^{(\lambda)}, \hat{X}^{(p)}, \hat{X}^{(\lambda)}, \tilde{X}^{(p)}, \tilde{X}^{(\lambda)}) = \\
&= \prod_{k=1}^{K}\prod_{i=1}^{n_k}\Big(1 - \frac{1}{1 + e^{-(x_{i,k}^{(p)}\beta^{(p)}+\tilde{x}_{i,k}^{(p)}\phi^{(p)}+\hat{x}_{k}^{(p)}\theta^{(p)})}}\Big)^{1-z_{i,k}} \times \\
&\Big(\frac{1}{1 + e^{-(x_{i,k}^{(p)}\beta^{(p)}+\tilde{x}_{i,k}^{(p)}\phi^{(p)}+\hat{x}_{k}^{(p)}\theta^{(p)})}}\Big)^{z_{i,k}} \times \\
&\Big(\frac{e^{(x_{i,k}^{(\lambda)}\beta^{(\lambda)}+\tilde{x}_{i,k}^{(\lambda)}\phi^{(\lambda)}+\hat{x}_{k}^{(\lambda)}\theta^{(\lambda)})^{y_{i,k}}}}{y_{i,k}!(1 - e^{-e^{(x_{i,k}^{(\lambda)}\beta^{(\lambda)}+\tilde{x}_{i,k}^{(\lambda)}\phi^{(\lambda)}+\hat{x}_{k}^{(\lambda)}\theta^{(\lambda)})}})} e^{-e^{(x_{i,k}^{(\lambda)}\beta^{(\lambda)}+\tilde{x}_{i,k}^{(\lambda)}\phi^{(\lambda)}+\hat{x}_{k}^{(\lambda)}\theta^{(\lambda)})}}\Big)^{z_{i,k}}
\end{aligned}
\tag{4.2}
$$

To recover the posterior distribution it is sufficient to apply the Bayes' Theorem and we get:

$$\pi(\rho|\mathcal{D}) =$$

$$= \frac{L(\mathcal{D}|\rho) \times \pi(\beta^{(p)}) \times \pi(\beta^{(\lambda)}) \times \pi(\theta^{(p)}) \times \pi(\theta^{(\lambda)}) \times \pi(\phi^{(p)}) \times \pi(\phi^{(\lambda)})}{\int_{supp(\rho)} L(\mathcal{D}|\rho) \times \pi(\beta^{(p)}) \times \pi(\beta^{(\lambda)}) \times \pi(\theta^{(p)}) \times \pi(\theta^{(\lambda)}) \times \pi(\phi^{(p)}) \times \pi(\phi^{(\lambda)}) d\rho} \tag{4.3}$$

The posterior distribution in (4.3) is not in close-form expression (due to the non conventional expression of the likelihood (4.2)) but we can get it using STAN.

The sampler gives us the following traceplots of $\beta^{(p)}$ and $\beta^{(\lambda)}$, which are reported below in Figures 4.10 and 4.11.

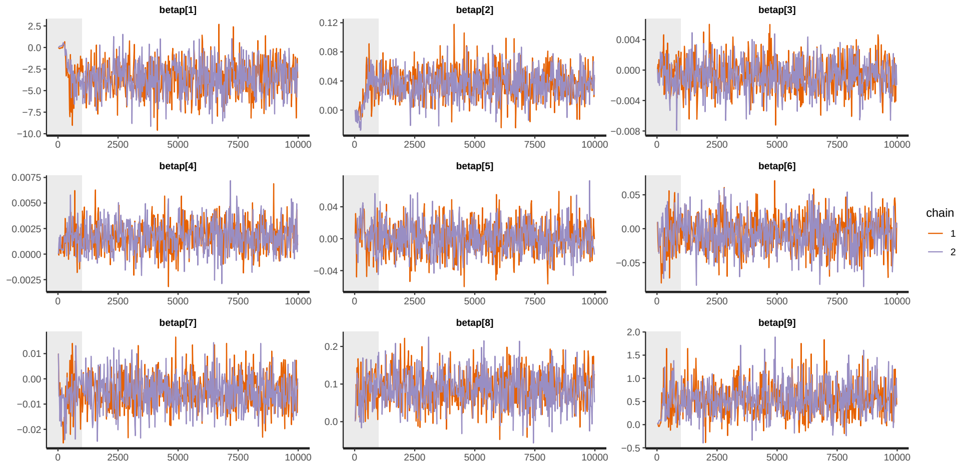

Figure 4.10: Traceplots of $\beta^{(p)}$ parameters.

We have run two different chains of size 20000 iterations, included a burn-in size of 1000, which forces the first 1000 values to be discarded, and we have decided to keep just one value every 20. Both the chains are mixed, converging to the invariant distribution, without seasonality, trends or cycles. The chains relative to $\theta^{(p)}$, $\theta^{(\lambda)}$, $\phi^{(p)}$ and $\phi^{(\lambda)}$ parameters are not reported here but, as long as the ones in Figures 4.10 and 4.11, do not show any alarming behaviour.
We end our diagnostic check looking at the AFC plot, which is not reported here due to high number of parameters, but no correlation between states of the chain at different lags has been found for all the parameters of the model.

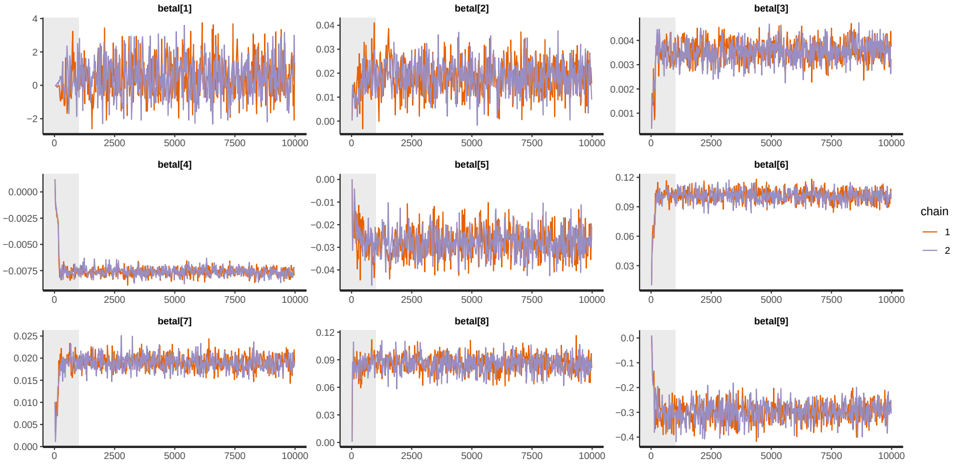

Figure 4.11: Traceplots of $\beta^{(\lambda)}$ parameters.

4.3.3 Bayesian posterior inference

Once the model has passed successfully the diagnostic check, we are ready to focus on the posterior distributions of the parameters provided by the sampler, which give us the opportunity to make inference about them and extract any interesting quantity.
The current model has several parameters (even 112 if we consider each component of $\beta^{(p)}$, $\beta^{(\lambda)}$, $\phi^{(p)}$, $\phi^{(\lambda)}$, $\theta^{(p)}$ and $\theta^{(\lambda)}$ individually, where 70 of them are related to the group effects). From the posterior distributions of the parameters, which is obtainable from the chain, we can select which parameters are most significant and which ones are negligible (at a certain level of credibility), in order to keep just the ones which have a direct influence on the response and to reduce the model complexity. Therefore, the 95% credible intervals are constructed and shown in Figure 4.12.
The figure highlights that not all the predictors of the regression are significant: when the zero value is contained in its 95% credible interval we assume it is zero at that fixed level of credibility, vice versa if zero value is not contained the predictor is taken different from zero. For the parameter $\beta^{(p)}$ the only significant component is $\beta_2^{(p)} \equiv \beta_{FFMC}^{(p)}$, relative to covariate FFMC (which is linked to the moisture content surface litter), while for parameter $\beta^{(\lambda)}$ almost all covariates (each one with its positive or negative sign) are significant at 95% level, with the exception of the intercept $\beta_1^{(p)}$. We can infer at a level of credibility of 95% that a fire ignition is mostly determined by the surface litter condition (and less by the condition of deeper layers) and its spread in the surrounding areas by a mixture of factor depending on soil and weather. For a briefly recap of all the quantities and abbreviation recover Table 4.1.

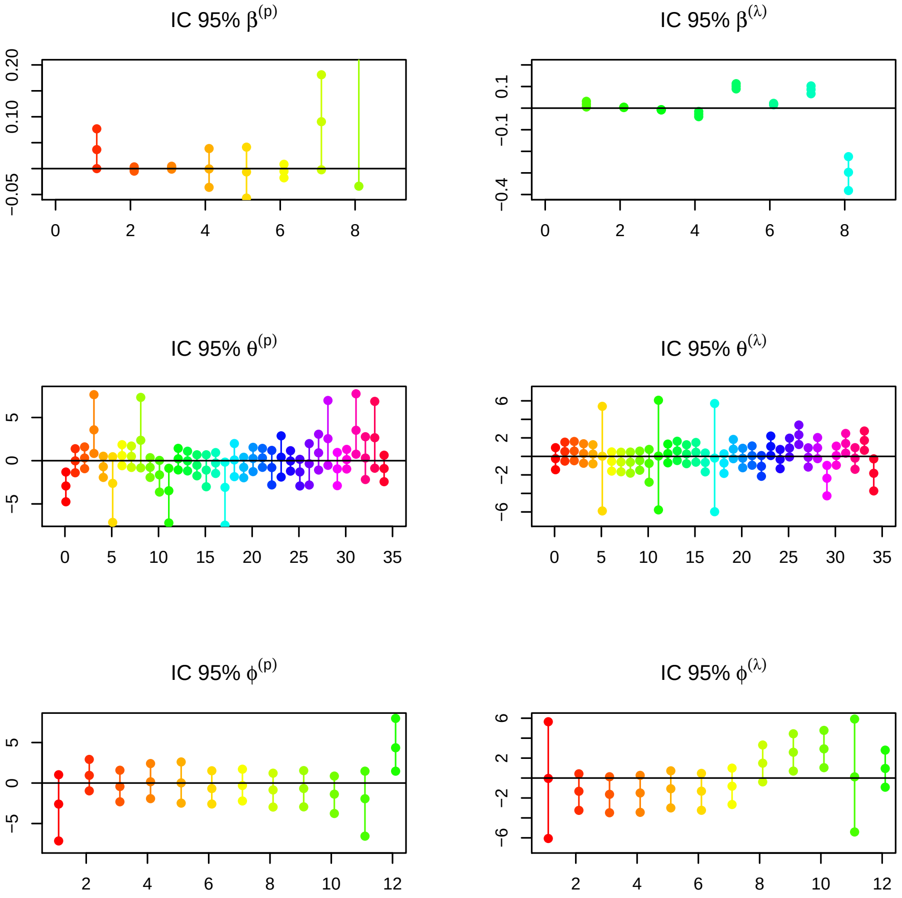

Figure 4.12: Posterior credible intervals for all parameters $\beta^{(p)}$, $\beta^{(\lambda)}$, $\theta^{(p)}$, $\theta^{(\lambda)}$, $\phi^{(p)}$ and $\phi^{(\lambda)}$; on the left the ones relative to p, on the right the ones relative to λ.

Some summary statistics (mean and standard deviation) about the significant parameters' posterior distributions are reported below in Table 4.3, Table 4.4 and Table 4.5. The FFMC coefficient ($\beta_2^{(p)}$) has a positive sign meaning that a small moisture content surface litter rating increases the probability of a fire outbreak (a high value of the FFMC code means high dryness conditions). With regards to $\beta^{(\lambda)}$ the features that increases the destructive power of the fire are high values of FFMC, DMC (related to moisture content of shallow organic layer rating), temperature, relative humidity and wind; the features that decreases it are DC (related to moisture content of deep organic layer), ISI (which is reasonable since, although the wind speed in general helps to spread the fire flames, it may turn off them at a surface level, especially when the fire has just been started) and rain.

Parameter	Mean	SD	Parameter	Mean	SD
$\beta^{(p)}_{FFMC}$	0.0365208	0.0202695	$\beta^{(\lambda)}_{FFMC}$	0.0184882	0.0068474
			$\beta^{(\lambda)}_{DMC}$	0.0035499	0.0004414
			$\beta^{(\lambda)}_{DC}$	-0.0076169	0.0003937
			$\beta^{(\lambda)}_{ISI}$	-0.0279919	0.0060162
			$\beta^{(\lambda)}_{temp}$	0.1012044	0.0061654
			$\beta^{(\lambda)}_{hum}$	0.0190489	0.0016437
			$\beta^{(\lambda)}_{wind}$	0.0859631	0.0092593
			$\beta^{(\lambda)}_{rain}$	-0.2983274	0.0398054

Table 4.3: Posterior distributions' summary statistics of $\beta^{(p)}$ and $\beta^{(\lambda)}$.

Looking at $\phi^{(\cdot)}$ coefficients we notice that in December the probability of a wildfire is increased, while in September and October the probability of observing a huge wildfire reaches its peak, after an increasing trend starting from July. The December results are quite counterintuitive but they are highly affected by the fact that the amount of recorded data about burnt hectares in this month is little (just 9), although they are all positive values.

Parameter	Mean	SD	Parameter	Mean	SD
$\phi^{(p)}_{Dec}$	4.5177281	1.7208704	$\phi^{(\lambda)}_{Sep}$	2.5760325	0.9175409
			$\phi^{(\lambda)}_{Oct}$	2.9132045	0.9237138

Table 4.4: Posterior distributions' summary statistics of $\phi^{(p)}$ and $\phi^{(\lambda)}$.

Parameter	Mean	SD	Parameter	Mean	SD
$\theta^{(p)}_{1}$	-2.9145979	0.8899205	$\theta^{(\lambda)}_{24}$	1.0898245	0.5431789
$\theta^{(p)}_{4}$	3.7392289	1.7430845	$\theta^{(\lambda)}_{27}$	2.3174036	0.5375751
$\theta^{(p)}_{12}$	-3.6203240	1.6963759	$\theta^{(\lambda)}_{30}$	-2.4375235	0.8506232
$\theta^{(p)}_{18}$	-3.3144143	1.8926843	$\theta^{(\lambda)}_{32}$	1.3967926	0.5379530
			$\theta^{(\lambda)}_{34}$	1.7050257	0.5337973
			$\theta^{(\lambda)}_{35}$	-1.8988314	0.8479513

Table 4.5: Posterior distributions' summary statistics of $\theta^{(p)}$ and $\theta^{(\lambda)}$.

About the $\theta^{(p)}_k$ and $\theta^{(\lambda)}_k$ coefficients look at Figures 4.13 and 4.14 respectively. $\theta^{(p)}$ parameter in Figure 4.13 is responsible for the fire ignition, while

in Figure 4.14 $\theta^{(\lambda)}$ is responsible for the fire propagation.
Each sub-region is shown in a coloured scale from red to blue, in such a way that regions in shades of red corresponds to positive $\theta_k^{(\cdot)}$, proportional to their magnitude, and the same with shades of blue regions which corresponds to regions with negative $\theta_k^{(\cdot)}$. Note that the most marked regions corresponds to the most significant coefficients reported in Table 4.5 (like $\theta_1^{(p)}$ at coordinates $(1,2)$, $\theta_4^{(p)}$ at $(1,5)$, $\theta_{12}^{(p)}$ at $(3,6)$, $\theta_{18}^{(p)}$ at $(5,5)$ for the p parameter and $\theta_{24}^{(p)}$ at coordinates $(7,3)$, $\theta_{27}^{(p)}$ at $(7,6)$, $\theta_{30}^{(p)}$ at $(8,5)$, $\theta_{32}^{(p)}$ at $(9,4)$, $\theta_{34}^{(p)}$ at $(9,6)$, $\theta_{35}^{(p)}$ at $(9,9)$ for the λ parameter), since they are the largest in magnitude and do not contain the zero value in the 95% posterior credibility interval. From these figures seems that both the extreme left and right areas of the park are more subjective to the outbreak of a fire, vice versa the middle area is a quite safe zone; the North-East area, instead, is more dangerous in terms of magnitude of the fire, while in the West area, which was more subjective to the ignition, the flames of the fires have much more difficulty to spread.

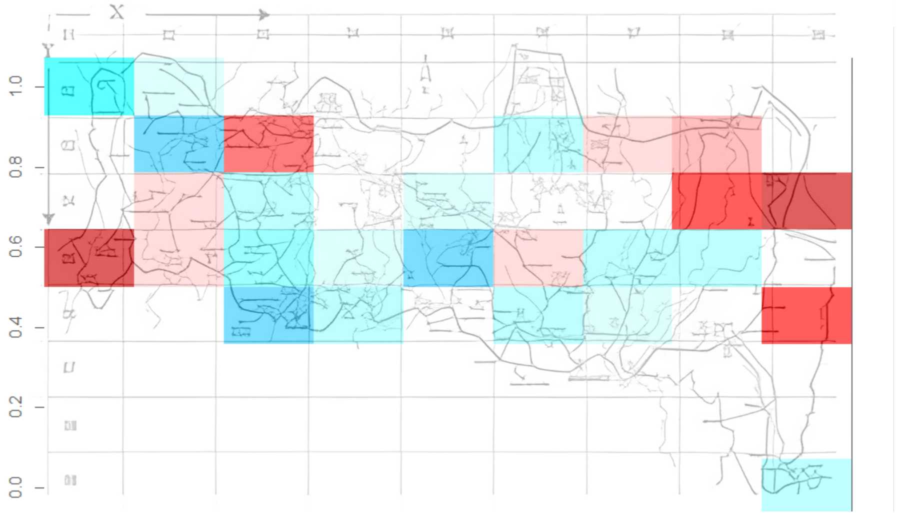

Figure 4.13: Groups effect $\theta^{(p)}$ relative to the fire ignition: regions in shades of red corresponds to positive $\theta_k^{(p)}$ (high probability of fire ignition), shades of blue regions corresponds to regions with negative $\theta_k^{(p)}$ (low proability), proportional to the magnitude of the coefficient.

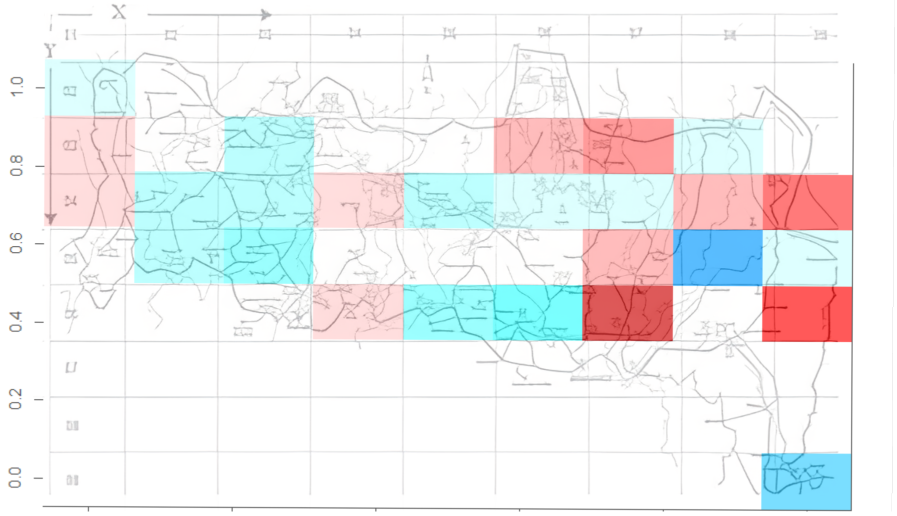

Figure 4.14: Groups effect $\theta^{(\lambda)}$ relative to the fire spread: regions in shades of red corresponds to positive $\theta_k^{(\lambda)}$ (high fire spread), shades of blue regions corresponds to regions with negative $\theta_k^{(\lambda)}$ (low fire spread), proportional to the magnitude of the coefficient.

4.3.4 Sensitivity analysis

Prior sensitivity examination plays an important role in applied Bayesian analysis. Especially in the situation where there is not much information available to use for selecting the suitable prior distribution. Hence, handling complex Bayesian models without any prior robustness may be problematic and could have an undesired influence on the posterior inference. In order to guarantee reliable and robust results, it is essential to check how sensitive the resulting posteriors are for each prior input.

In Bayesian statistics literature the general sensitivity concept is treated by Geisser (1992) [16], Clarke and Gustafson (1998) [17] and Millar and Stewart (2007) [18]. The technique we are considering consists of repetitive fits of the model with modified prior hyperparameters. If the posterior distributions do not differ much, robustness is claimed. The main drawback of this approach is that it requires several re-fits of the model, which may be extremely time consuming. Our strategy will be focusing on a bunch of prior distributions (simply varying the hyperparameter's values as said above) to limit the number of model re-fits.

In particular two scenarios will be considered:

1. Standard model (the one in equation (4.1));

2. Different parameters' values in equation (4.1) (in particular we adopted a normal distribution $\mathcal{N}(0{,}100)$ for all single components of the param-

eters $\beta^{(\cdot)}$, $\theta^{(\cdot)}$ and $\phi^{(\cdot)}$, instead of $\mathcal{N}(0,9)$, in order to have more diffuse priors);

Changing the hyperparameters the main parameters relative to the significant $\beta^{(\cdot)}$, $\theta^{(\cdot)}$ and $\phi^{(\cdot)}$ do not change neither in sign nor in the magnitude of the parameters. In conclusion, we can state that substantive different results from one model to the other do not show up.

4.3.5 Prediction

In [1] the authors were interested in comparing the posterior parameters estimations per location between a ZIP and a hurdle model; here we have a slightly different purpose which is to check how accurate a predicted value is, compared to an empirical value (for the moment not in terms of predicted values Y themselves but looking the the Poisson hurdle parameters p and λ). In order to have an unbiased estimate of both the empirical parameters p_k and λ_k of each sub-region, we use the MLE estimators for the Bernoulli trial and the Truncated Poisson distribution (for the latter an estimate of $\hat{\lambda}_{k_{MLE}}$ is provided by Moore [19]), therefore $\hat{p}_{k_{MLE}} = \frac{\sum_{i=1}^{n_k} z_i}{n_k}$ and $\hat{\lambda}_{k_{MLE}} = \frac{\sum\limits_{i\in\{1,\dots,n_k\}:y_i>0} y_i}{\sum\limits_{i\in\{1,\dots,n_k\}:y_i>0} \mathbb{I}(y_i<\hat{k})}$, where $\hat{k} = \max\limits_{i\in\{1,\dots,n_k\}:y_i>0}(y_1,\dots,y_{n_k})$. Observe that the estimators are computed without considering the covariates and looking just at the response Y.

The Bayesian estimate of λ_k and p_k (which we denote as $p_{k_{Bayes}}$ and $\lambda_{k_{Bayes}}$) instead, are computed through the estimated posterior parameters $\beta^{(p)}$, $\beta^{(\lambda)}$, $\theta^{(p)}$, $\theta^{(\lambda)}$, $\phi^{(p)}$ and $\phi^{(\lambda)}$ (setting the value of them at their posterior mean) and reconstructing the Bayesian posterior estimates of $p_{k_{Bayes}}$ and $\lambda_{k_{Bayes}}, \forall k \in \{1,\dots,35\}$ through the inv-logit function and the exponential one (basically replacing in (4.1) the covariates for each observation and then computing the mean by groups).

In the end, the relative deviations (between the MLE estimations and the Bayesian ones) $RD_{p_k} = \frac{|\hat{p}_{k_{MLE}} - p_{k_{Bayes}}|}{\hat{p}_{k_{MLE}}}$ and $RD_{\lambda_k} = \frac{|\hat{\lambda}_{k_{MLE}} - \lambda_{k_{Bayes}}|}{\hat{\lambda}_{k_{MLE}}}$ are computed for each sub-region and showed respectively in Figures 4.17 and 4.20.

In Figures 4.15, 4.16, 4.18 and 4.19 we report the empirical (through MLE) and the (Bayesian) posterior model parameters p and λ for each sub-region of the park; each dot represents the parameter of the relative sub-region and it is proportional to its magnitude. We observe, looking at the size of the dots in a qualitative way, that the model parameters are able to recover the empirical parameters. For a quantitative checking look at Figures 4.17 and 4.20.

The dashed line in Figure 4.17 is at a constant value of 0.1 and all the

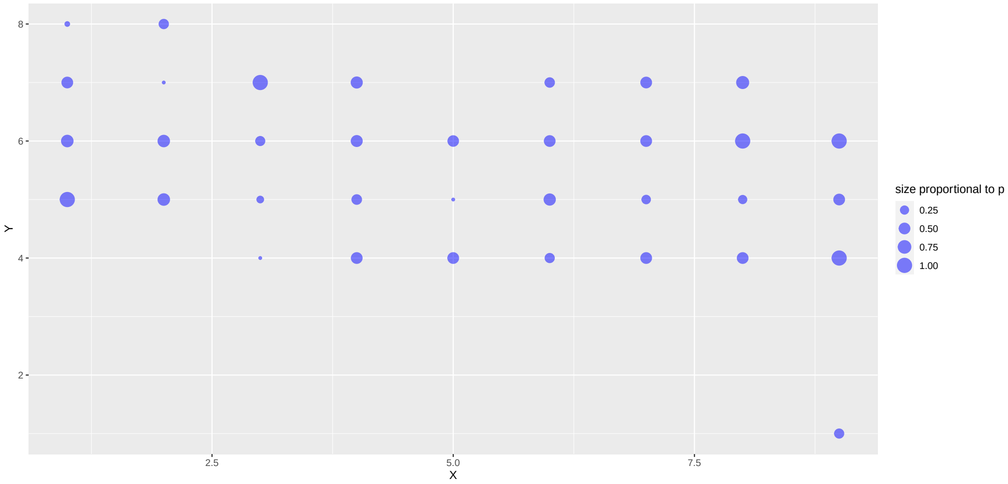

Figure 4.15: Empirical value of p parameter (computed through MLE) for each sub-region. Each single dot represents the empirical probability that a fire breaks out in the corresponding sub-region.

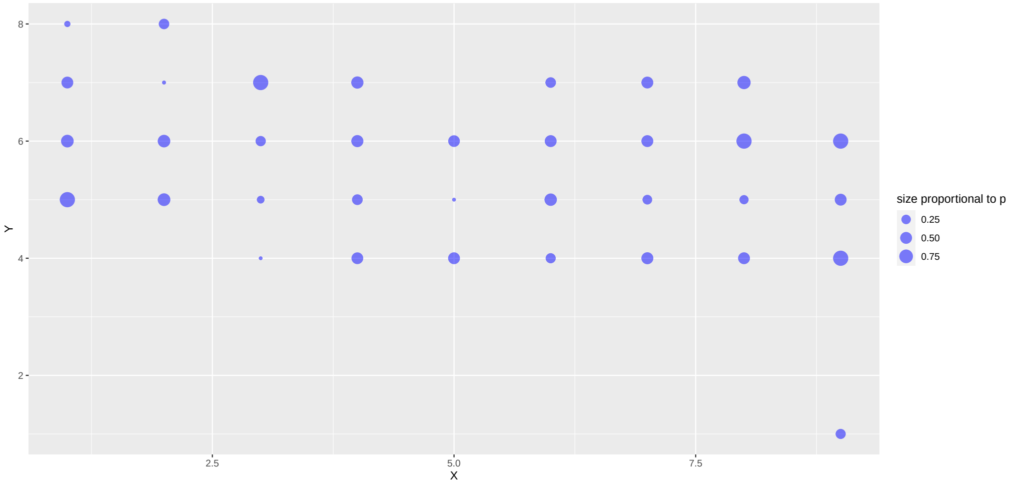

Figure 4.16: Bayesian estimate of p parameter for each sub-region. Each single dot represents the bayesian mean posterior probability that a fire breaks out in the corresponding sub-region.

relative errors are below this value. In particular the maximum value is 0.09785145, meaning that approximating the empirical p with the posterior p we are making less than 10% error with respect to the magnitude of the parameter itself.

Also for λ parameter the dashed line in Figure 4.20 is plotted at 0.1. How-

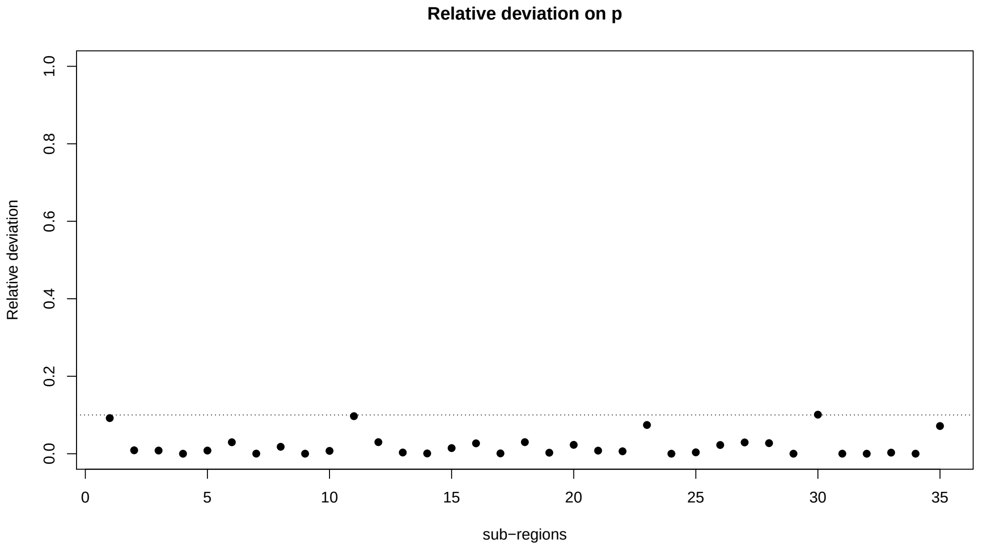

Figure 4.17: Relative deviation RD_{p_k} between MLE and Bayesian estimations of p parameter for each sub-region $k \in \{1, \dots, 35\}$.

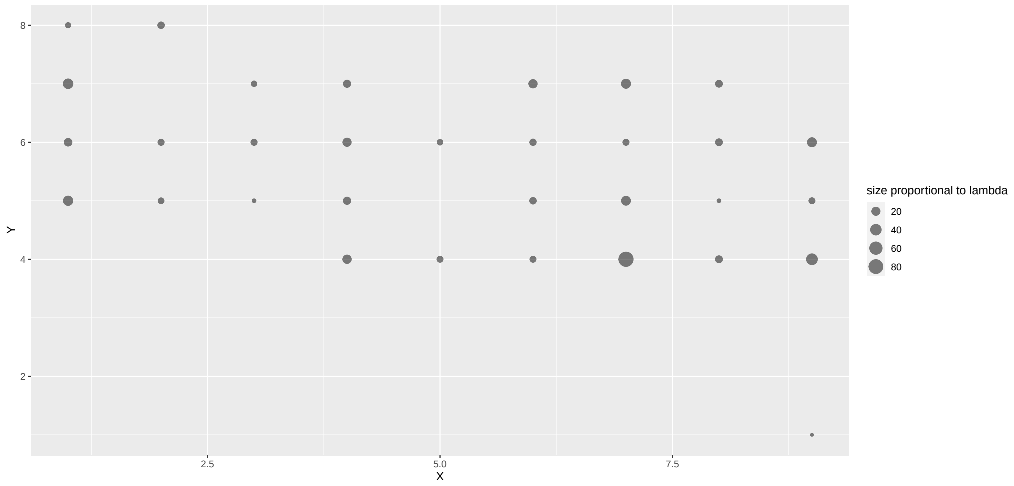

Figure 4.18: Empirical value of λ parameter (computed through MLE) for each sub-region. Each single dot represents the empirical magnitude of a broke out fire in the corresponding sub-region.

ever, this time, the greatest relative error is 0.740901401, which is much bigger than the maximum relative error on p, suggesting a greater uncertainty when predicting the burned area that follows the fire ignition, whose responsible is parameter λ.

Note that for sub-regions 6, 12 and 18 we do not have the relative error

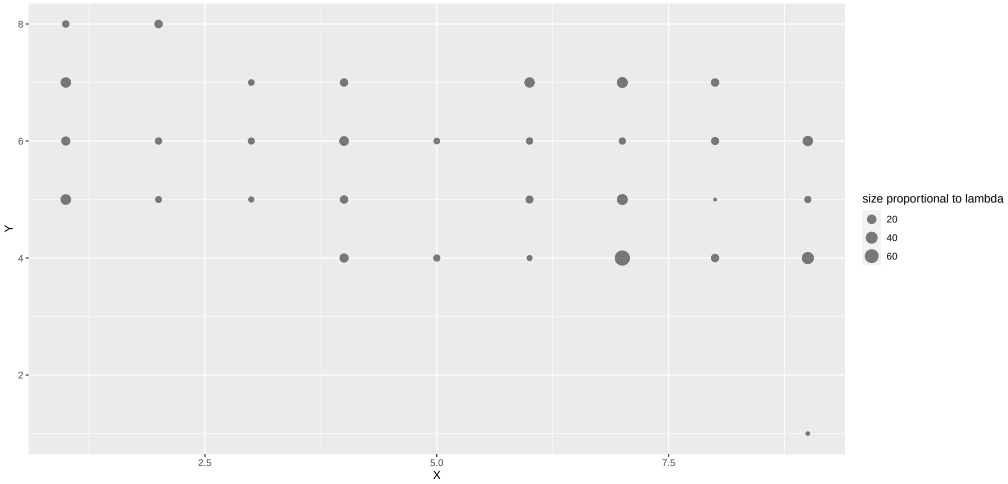

Figure 4.19: Bayesian estimate of λ parameter for each sub-region. Each single dot represents the bayesian posterior mean of the magnitude of a broke out fire in the corresponding sub-region.

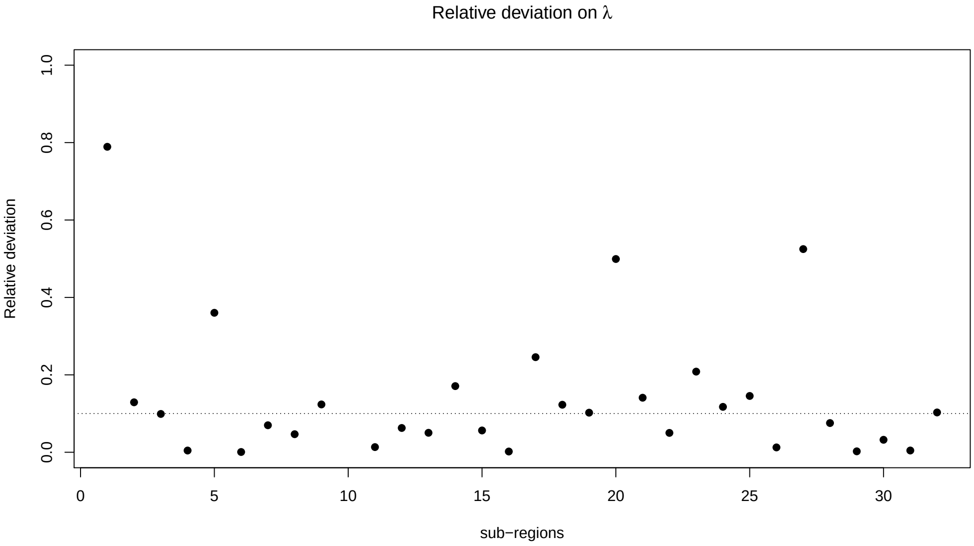

Figure 4.20: Relative deviation RD_{λ_k} between MLE and Bayesian estimations of λ parameter for each sub-region $k \in \{1, \ldots, 35\}$.

estimation because they are regions full of zeros (no fire recorded) and it is impossible to extract $\hat{\lambda}_{6_{MLE}}$, $\hat{\lambda}_{12_{MLE}}$ and $\hat{\lambda}_{18_{MLE}}$.

A Bayesian analysis leads naturally to making predictions about future observations from the random process that generated the data. A Bayesian prediction is the outcome value simulated from the posterior predictive distribution, which is the distribution of unobserved (or future) data given the observed data.

For our occurrence we have randomly selected a sample of 50 data among the total 506, which have been used as test test, just for making prediction. They corresponds roughly to 10% of the available data. Their covariates are used to simulate predictive values while the corresponding true outcome is used later, in comparison with predictions.

In Figure 4.21 we have simulated a single data (represented in red) using as coefficients the parameters' posterior means $\mathbb{E}[\beta^{(\cdot)}|\mathcal{D}]$, $\mathbb{E}[\theta^{(\cdot)}|\mathcal{D}]$ and $\mathbb{E}[\phi^{(\cdot)}|\mathcal{D}]$ and we have compared it with the real observed value (the black dot).

Instead, in Figure 4.22 we have simulated 20 outcomes from the Bayesian predictive distribution $P(Y^{new}|y)$ for each of the 50 testing data. In red dots are shown the predicted outcomes while the black dots are the corresponding true values. Differently from the previous case (Figure 4.21), where the values of the model posterior parameters have been kept fixed at their mean, in this case we are considering also their uncertainty, because we are sampling from the predictive distribution which uses the posterior distribution of the model parameters. In other words this is a Bayesian prediction which differs from a frequentist prediction because it consists in simulated outcomes and thus stochastic quantities.

In general the model seems to predict with more accuracy the situation where no fires occurs, while when the wildfire breaks out, it often underestimates its magnitude.

4.3.6 Computational costs

Table 4.6 shows the computational costs in terms of time spent to run the chains of the model. The elapsed time for each chain is in tune with the times in Section 3.4 (Table 3.10), where also the number of groups is significantly different from the test on synthetic data to the current real dataset.

Number of data	Number of groups	Elapsed time Ch.1 [sec]	Elapsed time Ch.2 [sec]
506	35	10600.9	10608.2

Table 4.6: Required computational time for current dataset.

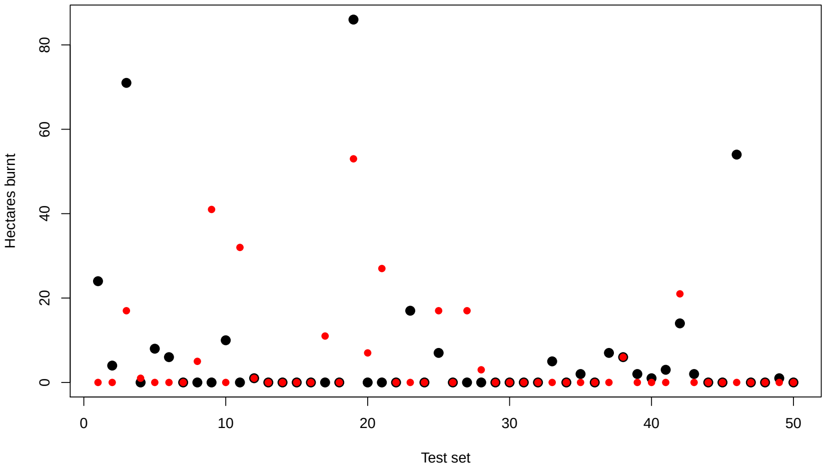

Figure 4.21: Prediction on new unseen data: the prediction is a single data (represented in red) obtained using as coefficients the parameters' posterior means $\mathbb{E}[\beta^{(\cdot)}|\mathcal{D}]$, $\mathbb{E}[\theta^{(\cdot)}|\mathcal{D}]$ and $\mathbb{E}[\phi^{(\cdot)}|\mathcal{D}]$, compared with the real observed value (the black dot).

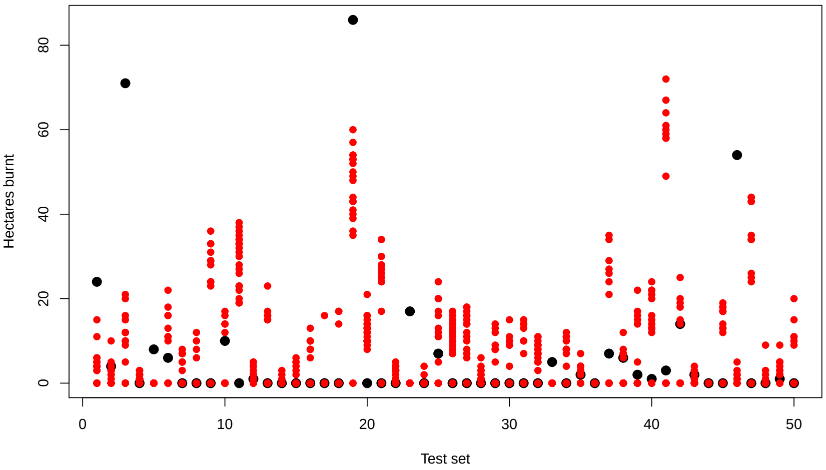

Figure 4.22: Bayesian prediction of 50 new unseen data (whose outputs are known and represented in black dots): for each data 20 outcomes generated from the Bayesian predictive distribution are represented in red dots.

Discussion and conclusions

The Poisson hurdle models proposed in Chapter 1 and tested in Chapter 3 of this elaborate can cover a variety of real case situations, from the simplest one without neither covariates nor group-specific structure to the most complex in which both the formulations could be present. Moreover, all the models have some attractive features, first of all the ability to address potential zero inflation relative to the ordinary Poisson distribution. In addition to this they consist of binary and truncated Poisson components, in fact this peculiarity of hurdle models guarantees to split the occurrence of an event and the magnitude of the occurred event. This has the powerful consequence that separate analyses could be carried out (a Poisson distribution, by its own intrinsic structure, would make it more complicated). This aspect has been widely covered in Chapter 4, which proposed a Poisson hurdle model for exploring geographic incidence of wildfires, by making inference on two distinct parameters. Another important aspect tackled in the elaborate has been the data spatial dependence which was solved using group-specific random effects that highlighted the presence of areas with different characteristics and strengthened our idea to adopt a model that accounts the existence of heterogeneous areas; here a quite surprising result has shown up in fact we discovered that there exist regions where a high incidence of fires does not correspond to an equal destructive power and viceversa. This fact can translate into a more careful management of the available resources, reinforcing the prevention of fire ignition in some areas and the prevention of fire spread in others. Regarding the significance of the other model parameters we can infer that do exist fire prone regions, in the terms described above and that the summer months bring the risk of wildfires but the peak of severe fire incidence occurs in the early autumn months.

Continuing to work the case study we obtained the posterior distributions of all the model parameters. Most of them contained the zero value in the 95% credibility interval. By far the most significant warning light for the fire ignition is concerning the moisture content surface litter, while its spread is affected by a mixture of meteorological and geothermal factors.

The chosen bayesian model is also able to replicate the parameters extracted in an empirical way from data, through the MLE, since the posterior parameters' estimates slightly differ from the empirical ones, though the Poisson

part of the model shows greater deviations and uncertainty.
The same uncertainty relative to the fire propagation (once it broke up) was found in prediction, where the model captured the no fires cases with good results and it did underestimate the fire case.

In the statistical community GLM regression and groups structure are recurrent concepts which, in our case, have been blended with the Poisson hurdle distribution. A further step forward could be done proposing and exploiting a new alternative model, which is different from all the ones seen in this elaborate but it recovers the Poisson hurdle structure and is suitable to handle both the over-abundance of zeros and the spatial reference. The model has the following structure:

$$\begin{aligned} &Y_{i,k} \sim U_k V_{i,k} N_{i,k} && i \in \{1,2,...,n_k\},\ \ k \in \{1,2,...,K\} \\ &U_k \sim Be(q) && k \in \{1,2,...,K\} \\ &V_{i,k} \sim Be(p_k) && i \in \{1,2,...,n_k\},\ \ k \in \{1,2,...,K\} \\ &N_{i,k} \sim tPois(\lambda_k) && i \in \{1,2,...,n_k\},\ \ k \in \{1,2,...,K\} \end{aligned}$$

It assumes that data is organized in K sub-regions (or macro-regions) and each group has a fixed number n_k of observations. Furthermore there are two steps to handle to have a positive data: the first step is given by drawing from a Bernoulli random variable U_k, which is responsible of the whole macro-region k, in the sense that according to its outcome we can have a region full of zero values ($U_k = 0$) or not ($U_k = 1$). In addition to this, it is straightforward to point out that conditionally on $U_k = 0$ the responses $\{Y_{i,k}\}_{i=1,...,n_k}$ in the same group are dependent because they are all zero, while conditionally on $U_k = 1$ the responses in the same group are independent. The second stage, which involves within groups variables $V_{i,k}$ and $N_{i,k}$, is a familiar Poisson hurdle structure whose definition is introduced at the beginning of Chapter 1, for which the same parameters has been used.
By its structure, this model is more suitable for dynamics where the are entire groups of data consisting of zero values.

Appendix A

R Functions

Here, in this Appendix, we report the functions implemented in software *R* to sample from a Poisson hurdle distribution.

A.1 Functions used for models in Sections 3.1, 3.2, 3.3 and 3.4

Poisson Hurdle base model (3.1)

```
sample_hurdle=function(N,p,lambda){
    z=rbern(N,p)
    y=rtpois(N, lambda, a = 0, b = Inf)
    return(z*y)
}
```

Poisson Hurdle Regression (3.2)

```
sample_hurdle_glm=function(N,beta_p,beta_l,X_p,X_l){
    p=1/(1 + exp(-(X_p %*% beta_p)))
    lambda=exp(X_l %*% beta_l)
    z=rbern(N,p)
    y=rtpois(N, lambda, a = 0, b = Inf)
    return(z*y)
}
```

Poisson Hurdle with group structure (3.3)

```
sample_hurdle_group=function(vectorng,vectorp,vectorlambda){
    y=c()
    for(i in 1:length(vectorp)){
        y=c(y,sample_hurdle(vectorng[i],vectorp[i],vectorlambda[i]))
    }
    return(y)
}
```

Poisson Hurdle Regression with group structure (3.4)

```
sample_hurdle_group_glm=function(vectorng,beta_p,beta_l,
theta_p,theta_l,X_p,X_l){
    y=c()
    vectorp=c()
    vectorlambda=c()
    k=1
    for(i in 1:length(vectorng)){
        for(j in 1:vectorng[i]){
            vectorp=c(vectorp,1/(1+exp(-(beta_p %*% X_p[k,]+theta_p[i]))))
            vectorlambda=c(vectorlambda,exp(beta_l %*% X_l[k,]+theta_l[i]))
            y=c(y,sample_hurdle(1,vectorp[k],vectorlambda[k]))
            k=k+1
        }
    }
    return(y)
}
```

Bibliography

[1] Jay M. Ver Hoef and John K. Jansen, *Space–time zero-inflated count models of Harbor seals*, John Wiley & Sons (2007).

[2] Brian Neelon, Pulak Ghosh and Patrick F. Loebs, *A spatial Poisson hurdle model for exploring geographic variation in emergency department visits*, Royal Statistical Society (2012).

[3] Souparno Ghosh, Alan E. Gelfand, Kai Zhu and James S. Clark, *The k-ZIG: Flexible Modeling for Zero-Inflated Counts*, International Biometric Society, Biometrics, Vol. 68, pp. 878-885 (2012).

[4] Diane Lambert, *Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing*, Taylor & Francis, Technometrics Volume 34 - Issue 1 (1992).

[5] P. McCullagh and John A. Nelder, *Generalized Linear Models*, Second edition, Chapman & Hall/CRC Monographs on Statistics and Applied Probability, Issue 37 (1989).

[6] J. Jacod, P. Protter, *Probability essentials*, Springer Science & Business Media (2004).

[7] Jackman S., *Bayesian analysis for the social sciences*, Wiley, New York (2009).

[8] Christian P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Springer (2007).

[9] C. P. Robert, G. Casella, *Monte Carlo Statistical Methods*, Springer, New York, NY (USA) (2004).

[10] J. Geweke, *Bayesian inference in econometric models using Monte Carlo integration*, Econometrica J. Econometric Soc., pp 1317–1339 (1989).

[11] W. K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika 57(1), pp 97–109 (1970).

[12] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, A. Riddell, *Stan: A probabilistic programming language*, J. Stat. Softw. 76(1) (2017).

[13] Paulo Cortez and Anibal Morais, *A Data Mining Approach to Predict Forest Fires using Meteorological Data*, Department of Information Systems/R&D Algoritmi Centre, University of Minho, 4800-058 Guimaraes, Portugal (2007).

[14] B.D. Lawson and O.B. Armitage , *Weather Guide for the Canadian Forest Fire Danger Rating System*, Natural Resources Canada, Canadian Forest Service, Northern Forestry Centre, Edmonton, AB, (2008).

[15] J.A. Turner and B.D. Lawson, *Weather in the Canadian Forest Fire Danger Rating System. A user guide to national standards and practices*, Environment Canada, Pacific Forest Research Centre, Victoria, BC. BC-X-177 (1978).

[16] Seymour Geisser, *Bayesian Perturbation Diagnostics and Robustness*, Bayesian Analysis in Statistics and Econometrics pp 289–301 (1992).

[17] Bertrand Clarke and Paul Gustafson, *On the overall sensitivity of the posterior distribution to its inputs*, Elsevier B.V., Volume 71, Issues 1–2, pp 137-150 (1998).

[18] Russell B. Millar and Wayne S. Stewart, *Assessment of locally influential observations in Bayesian models*, International Society for Bayesian Analysis 2(2), pp 365-383 (2007).

[19] P. G. Moore, *The transformation of a truncated poisson distribution*, Taylor & Francis, Scandinavian Actuarial Journal Vol 1956(1), pp 19-25 (1952).