Master's thesis

2022

Sara Elise Wøllo

Master's thesis

NTNU
Norwegian University of
Science and Technology
Faculty of Information Technology and Electrical
Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Sara Elise Wøllo

Correcting for under-reporting of violence against women in Italy using INLA

June 2022

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

Correcting for under-reporting of violence against women in Italy using INLA

Sara Elise Wøllo

Applied Physics and Mathematics
Submission date: June 2022
Supervisor: Sara Martino

Norwegian University of Science and Technology
Department of Mathematical Sciences

Abstract

The WHO estimate that 30% of all women globally have been subjected to physical and/or sexual violence from an intimate partner or sexual violence from a non-partner in their lifetime. In Italy, this estimate is 31.5%. Although the societal consequences are serious, few victims report the incident to the police. This severe under-reporting can lead to the prevalence of violence being under-estimated, which again will lead to preventative measures being too small or not being deployed at all.

Many statistical models have been developed to account for such under-reporting, including the Bayesian hierarchical Poisson-Logistic model. For a subclass of Bayesian hierarchical models fulfilling certain criteria Integrated Nested Laplace Approximations (INLA) have become an established method for inference. Until now INLA could not be applied to the Poisson-Logistic model. We investigate if the extension to the INLA methodology implemented in the `R`-library `inlabru` allow us to extend the scope of the INLA methodology onto the Poisson-Logistic model with severely under-reported count data. To ensure model identifiability we build upon recently developed theory and use an informative prior distribution on the rate of under-reporting. We prove the effectiveness of `inlabru` on the Poisson-Logistic model through a comprehensive simulation study, and then apply the model to the issue of violence against women on a regional level in Italy, where poor modelling results are returned. We believe that this is due to the quality of data, and that data aggregated up to a regional level is too generalised for this application. Because of this we suggest applying the model to the issue of violence against women in Italy on a provincial level as a further work.

Sammendrag

Verdens helseorganisasjon (WHO) anslår at 30% av alle kvinner på verdensbasis har vært usatt for partnervold eller seksuelle overgrep av noen som ikke er en samlivspartnerpartner. I Italia er dette tallet estimert til å være 31.5%. De samfunnsmessige konsekvensene er alvorlige, men det er det få av ofrene som rapporterer hendelsen til politiet. Denne alvorlige underrapporteringen kan føre til at utbredelsen undervurderes, noe som igjen gjør at preventative tiltake er for små til å ha en reell virkning.

Det har blitt utviklet mange statistiske modeller som tar høyde for underrapporterte data. En av disse er den bayesianske hierarkiske poisson-logistiske modellen. Integrerte nøstede Laplace approksimasjoner (INLA) en etablert metode for å gjennomføre inferens på en underklasse av bayesianske hierarkiske modeller. Før nå har det ikke vært mulig å bruke INLA på den poisson-logistiske modellen. Vi undersøker om utvidelsen implementert i `R`-biblioteket `inlabru` gjør det mulig å bruke INLA-metoden til å gjennomføre inferens på den poisson-logistiske modellen. Vi tar høyde for identifiseringsproblemer i den poisson-logistiske modellen ved å implementere en informativ prior-fordeling på underrapporteringsraten. Vi viser effektiviteten av `inlabru` på den poisson-logistiske modellen gjennom en større simuleringsstudie, for så å bruke metoden og modellen til å modellere utbredelsen av vold mot kvinner på det regionale nivået i Italia, hvor modellen yter dårlig. Vi tror dette er på grunn av datakvaliteten og at data aggregert opp til det regionale nivået i Italia blir for generalisert for denne problemstillingen. Vi foreslår derfor å bruke samme modell til å modellere utbredelsen av vold mot kvinner i Italia, men å bruke data aggregert opp til det provinsielle nivået i stedet. Dette overlates til videre arbeid.

Preface

This thesis is the result of my work in the course *TMA4900 Industrial Mathematics, Master's Thesis* at the Norwegian University of Science and Technology (NTNU). With this I conclude my studies of applied physics and mathematics at NTNU, specialising in industrial mathematics. I would like to extend a big thank you to my advisor Sara Martino, for very helpful meetings and discussions. I would also like to direct a thanks to Esten Nicolai Wøien for unwavering support and encouragement.

Sara Elise Wøllo
June 2022
Trondheim

Table of Contents

Chapter 1

Introduction

The Declaration on the Elimination of Violence against Women, adopted by the United Nations General Assembly (UNGA) in 1993 defines violence against women as “any act of gender-based violence that results in, or is likely to result in, physical, sexual, or mental harm or suffering to women, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or in private life” (United Nations 1993). The issue of violence against women is not something new, but it is an issue that has come more into focus recently. This is in part due to efforts from the United Nations General Assembly, as well as the World Health Organisation (WHO). The WHO published its first report on the topic in 2005 (WHO 2005), followed in 2013 by another report containing global and regional prevalence estimates of two forms of violence against women, “intimate partner violence” and “non-partner sexual violence” (WHO 2013). Honour killings, genital mutilation and trafficking of women are also forms of violence against women, but data on these topics are scarce, and therefore not included. The 2013 WHO report acknowledges that emotional abuse and controlling behaviour by a partner is also a form of violence against women, but as there is currently no consensus on when unkind behaviour crosses into emotional abuse, this is not included in the WHO report. The latest WHO estimates (WHO 2021) suggest that globally 30% (with 95% confidence interval $(25\%, 34\%)$) of all women aged 15 or older has been subjected to physical and/or sexual violence from either a current or former partner, or sexual violence from someone other than a current or former partner. In high income countries, as defined by The World Bank (The World Bank 2022), this estimate is 28% (with 95% confidence interval $(23\%, 34\%)$).

The main source of data on violence against women globally, as identified by the WHO, are specialised surveys on violence against women and modules on violence against women within larger national health surveys (WHO 2021). The WHO also address that violence against women, and especially sexual violence are topics that are stigmatised and incidents are often not disclosed of in these surveys. This results in under-reporting and prevalence estimates that are too low. This is a significant obstacle to implementing effective preventative measures, but it does give us more accurate estimates than before.

The societal consequences of violence against women are many and serious. The most serious of which is homicide. Globally, it is estimated that 39% of female homicides were committed by an intimate partner (Stöckl et al. 2013), although this estimate is likely to be under-estimated, as the victim-offender relationship is not always known (WHO 2013). If we look at high-income countries, this estimate rise to 41% (Stöckl et al. 2013). It is also estimated that 42% of all women subjected to intimate partner violence were physically injured by their partner. This in itself shows that intimate partner violence is an immense health burden to women experiencing it. Victims of intimate partner violence also report a higher rate of several other serious health issues. This includes them being 16% more likely to give birth to a low-weight baby, more than 50% more likely to contract HIV and other sexually transmitted diseases and twice as likely to have an abortion (WHO 2013). Experience with intimate partner violence is also connected to a higher rate of depression and attempted suicide, with depression rates being almost twice compared to women not having experienced violence in their relationships, and suicide attempts being more than four times as likely (WHO 2013). When it comes to the health-effects of non-partner sexual violence, less is known. There has been less research on this topic, and less data is available. Non-partner sexual violence has however been linked to increased risk of anxiety and depression (WHO 2013, FRA - European Union Agency for Fundamental Rights 2015) and alcohol use disorders (WHO 2013). In the 2014 EU-wide survey on violence against women, 42% of victims of non-partner sexual violence stated that they had obtained one or more injuries due to the incidence (FRA - European Union Agency for Fundamental Rights 2015). This clearly shows that also this form of violence towards women has detrimental societal consequences.

It is difficult to identify all factors contributing to violence against women, but some risk factors are similar across different sites. Abramsky et al. 2011 has analysed data from the WHO multi-country study on women's health and domestic violence (WHO 2005), and found that secondary education, high social economic status and formal marriage decreased the risk of intimate partner violence. Factors such as alcohol use, young age, attitudes supportive of domestic violence, having outside partners, experiencing childhood abuse or having grown up in a household that experienced intimate partner violence increases the risk of experiencing intimate partner violence. The same is true for women that has experienced or is with a partner that has perpetrated other forms of violence outside the home in adulthood (Abramsky et al. 2011) and for women that have a higher education and/or income level than their partner (Vugt et al. 2022). Less is known about the factors increasing the risk of non-partner sexual violence.

Although the WHO estimates that 30% of all adult women have experienced sexual or physical violence in their lifetime, far fewer women report this to the police. The reasons for this are complex and varies across the world. Different countries also have different laws and regulations, that can make it harder for women to come forward with a complaint. How extensive the scope of this under-reporting is can be difficult to determine, but results from an EU-wide survey suggests that among their 28 member states, only 14% of women subjected to intimate partner violence, and 13% of women subjected to non-partner violence reported their most serious incident to the police (FRA - European Union Agency for Fundamental Rights 2015). The same EU-wide survey reports that in a number of cases, the police came to know about the incident in some other way, and it is therefore estimated that 20% of all incidents of intimate partner violence and 19% of all incidents of non-partner violence were eventually reported to the police (FRA - European Union Agency for Fundamental Rights 2015). There can be many reasons why a woman choose to not report an incidence of violence to the police. Sanz-Barbero et al. 2018 looked at intimate partner violence data from Spain, and found that the most frequent reason for women to not report the incidence was them not giving importance to the situation, fear and lack of trust in the reporting process, or because the violent relationship ended. About 25% of the women that reported the incidence to the police later withdrew the complaint. The reasons given for this was that they thought the assailant would change, that the relationship ended, or due to fear and threats (Sanz-Barbero et al. 2018).

In this work, we investigate a spatial hierarchical model used on under-reported data, and apply this to the topic of violence against women on a regional level in Italy. Numbers from 2014 published by the Italian National Institute of Statistics (ISTAT) estimates that 31.5% of Italian women aged $16-70$ have suffered sexual or physical violence in their lifetime (Violenceagainstwomen.Stat 2022c). ISTAT also suggests that 4.5% of all Italian women in the same age group have experienced violence within the last 12 months (Violenceagainstwomen.Stat 2022c). This shows that violence against women is a societal issue also in Italy. Under-reporting of these events of violence is prevalent also in Italy, and a survey conducted by ISTAT in 2014 called “Survey on women’s safety” estimates that only 12.2% of all violence by an intimate partner and only 6% of violence conducted by a non-partner is reported to the Italian police (Department for Equal Opportunities 2014). This survey also looks at risk factors associated with violence against women in Italy. It found that inter-generational transmission of violence was a risk factor. It found that men growing up in homes where violence was present in the relationship between their parents, has a greater chance of becoming perpetrators of violence. Women growing up in similar circumstances also have a greater chance of becoming victims of violence. Women growing up in homes where they were abused and beaten has the greatest chance of experiencing violence also as an adult. 64% of all Italian women growing up in homes where they were beaten by a mother or father experiences physical or sexual violence as adults. This is twice as many as the national average in Italy at 31.5% (Department for Equal Opportunities 2014).

This thesis builds on a previous work (Wøllo 2022), and uses a similar Bayesian hierarchical model to perform inference. In Chapter 2 we introduce different methods for handling and modelling under-reported data, and present the Poisson-Logistic model that will be used throughout this work. The chapter also introduces the problem of identifiability in the Poisson-Logistic model, and methods to solve this problem. Chapter 3 presents Latent Gaussian Models and methods for performing inference on such models. The chapter also introduces `inlabru`, an extension to the INLA methodology that will be used to perform inference on the Poisson-Logistic model in this work. Chapter 4 presents a thorough simulation study to prove the usefulness of the `inlabru` approach on the chosen model. The chapter specifies a structured spatial component in the model, an discusses how to specify appropriate prior distributions for the model parameters. Chapter 4 also discusses the problem of a disconnected spatial structure, and how a model with few spatial regions influence the posterior estimates of the model. The chapter also discuss any concerns to be aware of when applying this model. Chapter 5 introduce a real-data application to the Poisson-Logistic model in the form of an application of the model to the incidence rate of violence against women on the regional level in Italy. We identify indicators that can be used to model the prevalence of violence against women, as well as an indicator used to explain variability in the extent of under-reporting across the regions of Italy. A Poisson-Logistic model is then specified. Inference on the model is performed using `inlabru`, and the results are presented. Lastly, in Chapter 6 we discuss the model performance, present problems and possible solutions to these problems, and look ahead at possible further work beyond this thesis.

Chapter 2

Models for under-reporting of count data

When it comes to modelling count data, two types of models are especially popular: the Poisson and the Negative Binomial distributions. We will focus on the Poisson model, which can be written as

$$y_i|\lambda_i \sim \text{Poisson}(\lambda_i), \tag{2.1}$$

where y_i are the observed counts and λ_i are the mean expected counts. A Poisson model assumes that the model variance is equal to the expected value.

In both Poisson and Negative Binomial models we assume the count to be fully observed. This is however not always a realistic assumption. Count data can be subjected to under-reporting. This means that if only a fraction of the true counts are observed, and if we do not take this under-reporting into account, that in turn will lead to biased estimates when modelling the observed data. Different statistical models have been proposed to fix this issue and we look more closely at two such models, the Censored Poisson Model and the Compound Poisson Model, in the following sections.

2.1 Censored Poisson Models

Terza 1985 introduced a model that allowed for censored count data, called the Censored Poisson Model (CPM). By censored count data we mean that each data point can either be observed exactly or observed to be in an interval, usually above or below a threshold. For example, a survey could give the respondent alternatives such as “0”, “1” and “2 or more”. If the respondent answered “0”, we know the exact value of the observation, but if the respondent answered “2 or more”, then all we know is that the value is larger than or equal to 2, and saying something about the exact value is impossible. Then, the censoring threshold of that survey question would be 2, as any observation larger than 1 is censored. Let z_i be the observed counts, y_i the true unobserved counts and C the censoring threshold. We can define a censoring indicator as

$$d_i = \begin{cases} 1, & \text{if } y_i < C, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

We are then able to describe the likelihood for all the observations as

$$f(\mathbf{z}|\lambda) = \prod_{i=1}^{n} \begin{cases} f_i(z_i), & \text{if } d_i = 1, \\ 1 - \sum_{j=0}^{C-1} f_i(j), & \text{if } d_i = 0, \end{cases} \tag{2.3}$$

where $f_i(z_i)$ is probability density function of z_i (Terza 1985).

For the model proposed by Terza 1985, the censoring threshold is constant across all observations. This model is therefore not very flexible, as it can be realistic to assume that the censoring can vary, for example across different regions in a spatial model. The CPM was therefore extended to allow for the censoring threshold to vary amongst the observations by Caudill et al. 1995. We denote this varying censoring threshold as C_i. If the sample contains censored and uncensored values, the likelihood function is a product of the density functions presented in Equation 2.3, when $d_i = 1$ and $d_i = 0$.

The likelihood can then be written as

$$f(\mathbf{z}|\mathbf{d},\lambda)=\prod_{i=1}^{n}\left\{\left[f_i(z_i)^{(1-d_i)}\right]\left[\left(1-\sum_{j=0}^{C_i-1}f_i(j)\right)^{d_i}\right]\right\}. \tag{2.4}$$

This model could for instance be applied to under-reporting, if we suspected that only some observations are subject to under-reporting. However, the under-reporting rate is still treated as binary, either present and constant above a predetermined threshold, or not present. We also require ad hoc information to determine which observations are subject to under-reporting.

Oliveira et al. 2017 continued to build on the Censored Poisson Model presented by Caudill et al. 1995 by proposing a random mechanism to specify which observations, or regions in a spatial model, are censored. The new model is called a Random-Censoring Poisson Model (RCPM). The model likelihood is hierarchically obtained as

$$f(\mathbf{z}|\mathbf{d},\lambda)=\prod_{i=1}^{n}\left\{\left[f_i(z_i)^{(1-d_i)}\right]\left[\left(1-\sum_{j=0}^{C_i-1}f_i(j)\right)^{d_i}\right]\right\}, \tag{2.5}$$

$$f(\mathbf{d}|\mathbf{p})=\prod_{i=1}^{n}p_i^{d_i}(1-p_i)^{1-d_i}, \tag{2.6}$$

where $\mathbf{z},\lambda,z_i$, and C_i are as before. d_i is again a censoring indicator assuming the value 1 if the region suffers from under-reporting, and 0 otherwise. The difference now is that d_i is a latent random variable that has a Bernoulli distribution with a censoring probability $p_i\in(0,1)$. Oliveira et al. 2017 show that the posterior estimates from the RCPM model are better if they use informative priors on the probability that an area is censored. Although this is the most flexible approach presented so far, Stoner et al. 2019 arguments that like the other Censored Poisson Models, it does not quantify how severe the under-reporting is for each observation. The RCPM specifies whether an observation has censoring, and therefore potential under-reporting, but it does not tell us anything about how under-reported the observation is. This can lead to biased estimates. The severity of the under-reporting might vary across the observations, and when using the RCPM the estimates for regions with more severe under-reporting will be persistently too low. Regions where the under-reporting is not as severe, will return estimates that are higher than the true value.

2.2 Compound Poisson Models

To introduce more flexibility into modelling under-reporting, we can look at Compound Poisson Models. The Compound Poisson Models permit the strength of under-reporting to vary across the observations, or regions in spatial models. The previously considered Censored Poisson Models looked at a region and determined whether it suffered under-reporting or not, but the severity of this under-reporting did not change from region to region. In contrast, the Compound Poisson Models assumes that all regions suffer from under-reporting, and that the probability of an event being reported is specific to each region (Oliveira et al. 2021).

Let y_i be the true unobserved count. The model assumes that each of the y_i events can be reported or not with probability $\pi_i \in [0, 1]$. We can then express the observed number of events z_i as a binomial distribution with parameters π_i and y_i

$$z_i | y_i, \pi_i \sim \text{Binomial}(y_i, \pi_i). \tag{2.7}$$

The true count y_i is assumed Poisson distributed, written as

$$y_i | \lambda_i \sim \text{Poisson}(\lambda_i). \tag{2.8}$$

Together, this count is then Compound Poisson distributed.

Oliveira et al. 2021 suggests that the biggest challenge under the Compound Poisson Models is to model the reporting probabilities $\pi = (\pi_i, ..., \pi_n)$ of the regions in a sensible way, and different approaches have been discussed in literature. Moreno et al. 1998 modelled the reporting probabilities directly using informative beta distributions, whereas Whittemore et al. 1991, Dvorzak et al. 2016, Stoner et al. 2019 and more modelled the reporting probability as a function of relevant covariates, $\pi_i = f(u_{1,i}, ..., u_{J,i})$. One popular approach to modelling the reporting probability is to assume $f(u_{1,i}, ..., u_{J,i})$ as a logistic function. Then, when adding all these parts together, this specific Compound Poisson Model can be written as

$$z_i | y_i, \pi_i \sim \text{Binomial}(y_i, \pi_i), \tag{2.9}$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{J} \beta_j u_i^{(j)}, \tag{2.10}$$

$$y_i | \lambda_i \sim \text{Poisson}(\lambda_i), \tag{2.11}$$

$$\log(\lambda_i) = a_0 + \sum_{k=1}^{K} \alpha_k x_i^{(k)}. \tag{2.12}$$

Here, y_i are the true counts of region i, whereas z_i are the observed counts. Furthermore, π_i is the probability that a true count is observed, and λ_i is the expected true count. α_0, β_0, α_k and β_j are unknown model parameters, and $x_i^{(k)}$ and $u_i^{(j)}$ are model covariates.

This specific Compound Poisson Model, combining a Binomial model and a Poisson model and letting the reporting probability depend on a logistic function is commonly referred to as a Poisson-Logistic model (Winkelmann et al. 1993). This model has been used in many different real-life applications. It was for example used within the field of economics by Winkelmann 1996 which estimated worker absenteeism in Germany, and the field of natural disasters by Stoner et al. 2018 which applied it to historically recorded volcano eruptions. Other applications include Amoros et al. 2006 exploring the differences between police crash data and the road trauma registry in France and Stoner et al. 2019 using this model to examine the regional differences in the level of under-reporting of tuberculosis in Brazil.

2.3 Reparameterisation of the Poisson-Logistic Model

We choose to focus on variations of the Poisson-Logistic models presented in Equations 2.9-2.12 from now on. This is a hierarchical model, and it is possible to use Bayesian methods to perform inference on this model. Inference on this model is however likely to be quite slow due to its complexity.

The model can be reparameterised to make inference easier. We can show that the model presented in Equations 2.9-2.12 can be re-written as a so called thinned Poisson process, where we have

$$z_i \sim \text{Poisson}(\pi_i \lambda_i), \tag{2.13}$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{J} \beta_j u_i^{(j)}, \tag{2.14}$$

$$\log\big(\lambda_i\big) = \alpha_0 + \sum_{k=1}^{K} \alpha_k x_i^{(k)}, \tag{2.15}$$

where all quantities are the same as in Equations 2.9-2.12.

To see this we look at the marginal probability of $z_i|\pi_i$, that is the probability of the observed counts unconditional on the true counts y_i. This can be written as

$$P(Z=z) = \sum_{y=0}^{\infty} P(Z=z|Y=y)\cdot P(Y=y) \tag{2.16}$$

$$= \sum_{y=z}^{\infty} \binom{y}{z} \pi^z (1-\pi)^{y-z} \cdot \frac{\lambda^y}{y!} e^{-\lambda} \tag{2.17}$$

$$= \sum_{y=z}^{\infty} \frac{y!}{z!(y-z)!} \pi^z (1-\pi)^{y-z} \cdot \frac{\lambda^y}{y!} e^{-\lambda}. \tag{2.18}$$

If we re-write this and move terms out of the sum, we are left with

$$= \frac{(\lambda\pi)^z}{z!} e^{-\lambda} \sum_{y=z}^{\infty} \frac{\Big((1-\pi)\lambda\Big)^{y-z}}{(y-z)!}. \tag{2.19}$$

Part of this expression can be recognised as the power series of an exponential function, and we can then rewrite it as

$$= \frac{(\lambda\pi)^z}{z!} e^{-\lambda} \cdot e^{(1-\pi)\lambda} \tag{2.20}$$

$$= \frac{(\lambda\pi)^z}{z!} e^{-\lambda\pi}, \tag{2.21}$$

which we recognise as a Poisson distribution with parameter $\lambda\pi$.

As the model parameterisation presented in Equations 2.13-2.15 is simpler than the original parameterisation, and therefore more efficient to perform inference on, we opt to using this model from now on. Stoner et al. 2019 uses MCMC simulations to conduct inference on a model similar to that presented in Equations 2.13-2.15. As MCMC simulations tend to be quite computationally intensive, we propose an alternative to this approach using Integrated Nested Laplacian Approximation (INLA) (Rue et al. 2009), and the `inlabru` library in `R`. These methods are presented in Chapter 3.

2.4 Identifiability of Compound Poisson Models

The parameters in Compound Poisson Models are in general not identifiable, and this is also true for the Poisson-Logistic model presented in Equations 2.9-2.12. What this means, is that there exists more than one unique set of parameter values that could produce the same posterior distributions. For instance, when looking at the model presented in Equations 2.9-2.12, a certain observed count z_i could be produced with several combinations of parameter values. The observed count could be identical if the true count y_i was low and the reporting probability π_i was high or if the true count y_i was high but the reporting probability π_i was low. This makes it impossible to correctly identify the intercepts in the two models. The hierarchical framework is also not able to identify whether the model covariates comes from the under-reporting or the count-generating process of the model (Stoner et al. 2019), which makes it impossible to say anything certain about the true count y_i or the reporting probability π_i, and therefore impossible to conduct meaningful inference on the model.

Different methods have been discussed in literature to ensure the identifiability of different Compound Poisson Models. All these methods introduce some prior information into the models, most regularly on the process of reporting. One commonly used method is to introduce a validation set into the model. This validation set is a data set assumed to be without under-reporting, and the level of under-reporting in the count data is then calibrated using this data set. This approach was used by Whittemore et al. 1991 and Dvorzak et al. 2016. The validation set method is however not very versatile. The validation set needs to contain data available for all sampling units in the model, something that can be difficult to obtain (Oliveira et al. 2021). Moreno et al. 1998 and Schmertmann et al. 2018 took a different approach. They looked at the model parameters for the different models in their hierarchical framework, and the prior information on them. To ensure identifiability, they specified the prior information by using the usual conjugate families for the model parameters. This in turn ensured identifiability of the posterior estimates from the model. This approach does require information on the reporting probabilities for each sampling unit in the model, which again can be difficult to obtain. Because this information might not be readily available, Oliveira et al. 2021 took a different approach. They systemically order all sampling units in the region of interest according to their data quality, from the units with the highest data quality to the units with the lowest data quality. They then include a reporting probability for the region with the highest data quality, and then decrease this reporting probability as the data quality decreases. This approach requires a lot less prior information, as only an estimate for the reporting probability of the sample unit with the best data quality is needed. This approach does however assume that the reporting probability decreases evenly across the sampling units, which is not necessarily the case. Stoner et al. 2019 used an informative prior distribution on the mean reporting rate to differentiate between the model parameters and therefore ensure identifiability. As discussed in Section 2.2, Stoner et al. 2019 model the reporting probability as a function of relevant covariates, $\pi_i = f(u_{1,i}, ..., u_{J,i})$, and it is the intercept of this model, denoted by β_0 in Equation 2.10 that is given an informative prior distribution. This means that the approach presented by Stoner et al. 2019 is not dependent on sample-unit specific prior information to ensure model identifiability, only prior information for the mean reporting rate of the region of interest.

2.4.1 Using an informative prior distribution to ensure identifiability

We choose to make the model identifiable by taking the same approach as in Stoner et al. 2019, using an informative prior distribution on one of the parameters in the under-reporting part of the model. Chen et al. 2022 builds on the model by Stoner et al. 2019, and integrate expert knowledge of the mode under-reporting into the model by using a beta prior distribution for the reporting rate π. They use expert knowledge about the reporting rate at the "average" value of all the centred covariates to develop this beta prior, and denote this reporting probability as p_0. By using the relationship between the reporting rate p_0 at the "average" value for all the centred covariates and the intercept in the under-reporting part of the model β_0, we can induce a prior distribution on β_0 from the beta prior on p_0.

Chen et al. 2022 develops a beta prior for the reporting rate p_0 by asking two simple question to an expert:

1. What value is most likely for the reporting probability
2. What value would be considered unusually high?

From this, two important pieces of information is obtained. The mode for the beta prior, and a percentile value. Using these two pieces of information, we can numerically solve for the parameters of the beta prior.

At the "average" values for all centred covariates related to the under-reporting, Equation 2.14 can be written as

$$\operatorname{logit}(p_0) := \log\left(\frac{p_0}{1-p_0}\right) = \beta_0, \tag{2.22}$$

seeing as $\sum_{j=1}^{J} \beta_j u_i^{(j)} = 0$. From this, we can induce a prior distribution on β_0 from the prior distribution on p_0 using the logistic relationship between them. This is done by a transformation of the random variable. Assuming $p_0 \sim f_{p_0}(p_0) = \operatorname{Beta}(a, b)$ and $p_0 = w(\beta_0) = (1 + \exp(-p_0))^{-1}$, we have that the distribution of β_0 is given by $f_{\beta_0}(\beta_0) = f_{p_0}(w(\beta_0))|w'(\beta_0)|$. Since $\beta_0 = \operatorname{logit}(p_0)$, we obtain

$$\begin{aligned} f_{\beta_0}(\beta_0) &= f_{p_0}(p_0) \cdot |w'(\operatorname{logit}(p_0))| && (2.23) \\ &= \frac{1}{\operatorname{Beta}(a,b)} p_0^{(a-1)} (1-p_0)^{(b-1)} \left| \frac{e^{(-\operatorname{logit}(p_0))}}{(1 + e^{(-\operatorname{logit}(p_0))})^2} \right|. && (2.24) \end{aligned}$$

This transformation enables us to specify the distribution parameters on β_0 based on the expert knowledge on p_0.

Chapter 3

Inference on Latent Gaussian Models

This chapter follows a similar structure as Wøllo 2022, as the theoretical basis of Latent Gaussian Models presented in that project is similar to the one presented here.

A Bayesian hierarchical model is a statistical model written in hierarchical form where Bayesian methods are used to perform inference. It most commonly consists of two or three parts, with the model likelihood as one part, the latent field as another and a vector of hyperparameters as a possible third part of the model. We can describe the model likelihood as the joint probability of the observed data, whereas the latent field are the variables that can not be directly observed and has to be estimated through the model. We use the latent field to describe possible dependencies in the data, and these dependencies are controlled through the hyperparameters in the hierarchical model. All the different parts of the model are linked together creating the hierarchical structure, and the generalised structure can be denoted as

$$\mathbf{y}|\mathbf{x}, \theta \sim \pi(\mathbf{y}|\mathbf{x}, \theta) \tag{3.1}$$

$$\mathbf{x}|\theta \sim \pi(\mathbf{x}|\theta) \tag{3.2}$$

$$\theta \sim \pi(\theta). \tag{3.3}$$

Here, $\pi(\mathbf{y}|\mathbf{x}, \theta)$ is the model likelihood, $\mathbf{x}$ is the latent field with distribution $\pi(\mathbf{x}|\theta)$ and the vector θ contains the hyperparameters of the model. The hyperparameters have prior distribution $\pi(\theta)$. When performing inference on the model we are interested in the posterior distributions of the hyperparameters and the latent field, $\pi(\mathbf{x}, \theta|\mathbf{y})$. The main interest usually lies in the posterior marginal distributions of the model, namely $\pi(x_i|\mathbf{y})$ and $\pi(\theta_i|\mathbf{y})$.

3.1 Traditional Methods for Inference on Bayesian hierarchical models

When the posterior distribution of a Bayesian statistical model is not available in closed form, Markov Chain Monte Carlo (MCMC) simulations is often used to draw samples from the posterior distributions and further to conduct inference on the model (Ravenzwaaij et al. 2018). The MCMC methodology consists of a collection of different algorithms that allow us to simulate the unknown posterior distributions. The earliest MCMC algorithm was introduced by Metropolis et al. 1953. This algorithm is today known as the Metropolis algorithm, and is still used. The Metropolis algorithm works well on simple models, but when the model parameters are highly correlated, it becomes ineffective. Gibbs sampling is a popular extension to the most basic MCMC algorithms (Ravenzwaaij et al. 2018). Using an MCMC algorithm with Gibbs sampling is the most common approach in Bayesian statistics, made even more accessible by the BUGS project. The BUGS (Bayesian inference Using Gibbs Sampling) project developed software for Bayesian inference using MCMC (*The BUGS Project* 1989), and this software have since been implemented for direct use in `R` through `WinBUGS` and `OpenBUGS`. It has also laid the foundation for `R`-packages like `NIMBLE` and `rjags`.

Another commonly used tool for Bayesian inference using MCMC sampling is the `C++` library `Stan`, implemented for use in `R` in the package `rstan`. Unlike BUGS, `Stan` does not use Gibbs sampling in their MCMC algorithm. Instead, `Stan` uses a No-U-Turn Sampler (NUTS), which is an extension to the Hamiltonian Monte Carlo method for sampling in MCMC algorithms (Hoffman et al. 2014).

MCMC methods are powerful tools for Bayesian inference, but it is not without shortcomings. MCMC algorithms can suffer from slow convergence, and determining when the algorithm has converged can be difficult. Extensive model checking should be done in order to say that the model has properly converged (Ravenzwaaij et al. 2018). MCMC methods using the Metropolis algorithm is dependent on a set of tuning parameters, and these tuning parameters need to be specified well to ensure convergence. Even though the methods using a Gibbs sampler works better with more complex models, it is still dependent on the set of tuning parameters being appropriate for convergence to be reached in a reasonable amount of time. MCMC algorithms based on the No-U-Turn Sampler version of the Hamiltonian Monte Carlo method avoids this all together, by not requiring tuning parameters to run. This speeds up the method significantly, and because of this `Stan` has become a popular alternative to tool built on BUGS. `Stan` is still a method of MCMC sampling, dependent on convergence to be useful. Therefore, even though this large range of MCMC algorithms are flexible and can be applied to a wide range of Bayesian statistical models, their long running times and slow convergence limits their usefulness in certain applications.

Because of this, other methods of inference has been suggested, and for a subclass of Bayesian hierarchical models called Latent Gaussian Models, Integrated Nested Laplacian Approximation has become popular.

3.2 Latent Gaussian Models

Latent Gaussian Models (LGM) are a subclass of hierarchical model where the latent field is Gaussian. As the likelihood is not required to be Gaussian, the posterior distribution $\pi(\mathbf{x}, \theta|\mathbf{y})$ is usually not available in closed form.

We further assume that the observations $\mathbf{y}$ are independent conditional on the Latent Gaussian Field $\mathbf{x}$ and the hyperparameter vector $\boldsymbol{\theta}$ (Martino et al. 2019). That is

$$\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \prod_i \pi(y_i|x_i, \theta). \tag{3.4}$$

We also assume that the distribution of y_i belongs to an exponential family with mean μ_i linked to a linear predictor η_i though a known link function $g(\cdot)$ such that $g(\mu_i) = \eta_i$ (Rue et al. 2009). This linear predictor is additive, which means it can be written on the form

$$\eta_i = \alpha + \sum_j h^{(j)}(u_{ij}) + \sum_k \beta_k z_{ik} + \epsilon_i. \tag{3.5}$$

Here, α is the intercept, and $h^{(j)}$ is a zero intercept function of a covariate u_{ij} modelling its random effects. z_{ik} are known covariates with linear effects, and ϵ_i are unstructured terms in the model. This class of models is very flexible, as the function $h^{(j)}(u_{ij})$ can be defined different ways. We define Gaussian prior distributions for $\alpha, h^{(j)}, \beta_k$ and ϵ. Then, we end up with a Gaussian latent field given by

$$\mathbf{x} = (\boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \mathbf{h}) \tag{3.6}$$

as well as the hyperparameter vector $\boldsymbol{\theta}$. This vector of hyperparameters does not need to be Gaussian (Rue et al. 2009).

3.3 Integrated Nested Laplacian Approximations

Integrated Nested Laplacian Approximation (INLA) was introduced by Rue et al. 2009. It is a deterministic method for Bayesian inference, and can be applied to Latent Gaussian Models that fulfil certain criteria. INLA uses analytical approximations in combination with numerical integration, resulting in accurate deterministic approximations to the posterior marginals $\pi(x_i|y)$ and $\pi(\theta_j|y)$ of a Latent Gaussian Model (Martino et al. 2019). As INLA is a deterministic method and is not dependent on sampling a large number of points, it is fast even for large and complex models (Martino et al. 2019). Unlike MCMC sampling methods, INLA does not struggle with slow convergence or poor mixing (Martino et al. 2019). We can break the INLA scheme for computing posterior marginals into four main steps

1. Exploring the hyperparameter space using Laplace approximations to approximate $\tilde{\pi}(\theta|\mathbf{y})$. Finding the mode and choosing several points $\{\theta^1, ..., \theta^K\}$ close to the mode of $\tilde{\pi}(\theta|\mathbf{y})$.

2. Computing $\tilde{\pi}(\theta^1|\mathbf{y}), ..., \tilde{\pi}(\theta^K|\mathbf{y})$ for the points selected in the previous step.

3. Approximating the density of $x_i|\theta, \mathbf{y}$ as $\tilde{\pi}(x_i|\theta^k, \mathbf{y})$ for all the selected points. This can be done by Gaussian, Laplace or simplified Laplace approximations (Rue et al. 2009).

4. Using numerical integration over θ to obtain the univariate posterior marginals of the model

$$\tilde{\pi}(x_i|\mathbf{y}) = \sum_{k=1}^{K} \tilde{\pi}(x_i|\theta^k, \mathbf{y})\tilde{\pi}(\theta^k|\mathbf{y})\Delta_k. \tag{3.7}$$

   Δ_k is defined as appropriate weights, for more details about these weights see Martino et al. 2019.

The INLA methodology is implemented in an `R`-package named `R-INLA`. See `https://www.r-inla.org` for documentation and examples.

A limitation to the INLA methodology is that it can only be applied to LGMs fulfilling certain criteria. Firstly, the LGMs need to be belonging to a subclass of Latent Gaussian Models named Latent Gaussian Markov Models (LGMM). These are models where the latent field $\mathbf{x}$ has conditional independence properties. This means that the latent field is a Gaussian

Markov random field (GMRF), with a sparse matrix $\mathbf{Q}(\theta)$ (Rue et al. 2005). The Bayesian hierarchical model presented in Equations 3.1-3.3 can be rewritten as a Latent Gaussian Markov Model on the form

$$\mathbf{y}|\mathbf{x}, \theta \sim \prod_i \pi(y_i|\eta_i, \theta) \tag{3.8}$$

$$\mathbf{x}|\theta \sim N(\mathbf{0}, \mathbf{Q}^{-1}(\theta)) \tag{3.9}$$

$$\theta \sim \pi(\theta), \tag{3.10}$$

with $\mathbf{Q}(\theta)$ defined as the precision matrix of the latent Gaussian field (Martino et al. 2019). This precision matrix of the LGMM needs to be sparse for the INLA methodology to be viable, as a numerical integration is performed over the θ space. This is not possible if the vector θ becomes too large. Martino et al. 2019 argues that it is feasible to use the INLA methodology as long as the number of hyperparameters θ is less than $n = 15$. More hyperparameters than this will make the numerical integration computationally infeasible. In addition to the model being a LGMM, the model predictor needs to depend linearly on the unknown smooth function of covariates. Lastly, each of the data points can only depend on the latent field through the linear predictor.

With these restrictions established, we can look back at Section 2, and the thinned Poisson-Logistic presented in Equations 2.13-2.15. Stoner et al. 2019 used MCMC simulations to conduct inference on a model similar to this, but due to the multiplicative term $\pi_i\lambda_i$ making the model predictor non-linear, this model does not fulfil the requirements presented above. INLA can therefore traditionally not be used to conduct inference in this model. Recently, a new extension to the INLA methodology has been proposed. This method is allows for non-linear terms in the model predictor $\eta(\mathbf{x})$, and opens up the possibility of using INLA to conduct inference on the thinned Poisson-Logistic model.

3.4 Linearisation of Non-Linear Predictors using the `inlabru` extension to the INLA methodology

An extension to the INLA algorithm was proposed by Bachl et al. 2019. This method is implemented in the `R`-package `inlabru`. Consequently, this method will be referred to as `inlabru` in this text. `inlabru` is implemented as a wrapper around the `R-INLA` package, and was originally implemented for use on ecological data. To work around the requirement of a linear predictor, `inlabru` adds a linearisation step to the INLA algorithm, using fixed point iteration to approximate a linearisation of the non-linear predictor. This linearisation is then used instead of the non-linear predictor, and the model requirements for using INLA is fulfilled. More formally, if we look at a Latent Gaussian Markov Model defined as in Equations 3.8-3.10, where we input a non-linear predictor $\tilde{\eta}(\mathbf{x})$, then the likelihood becomes

$$\mathbf{y}|\mathbf{x}, \theta \sim \prod_i \pi(y_i|\tilde{\eta}_i, \theta), \tag{3.11}$$

with $\mathbf{x}$ being the latent field. If we let $\bar{\eta}(\mathbf{x})$ be a Taylor approximation of $\tilde{\eta}$ at some $\mathbf{x}_0$, we get

$$\bar{\eta}(\mathbf{x}) = [\tilde{\eta}(\mathbf{x_0}) - \mathbf{B}\mathbf{x}_0] + \mathbf{B}\mathbf{x}, \tag{3.12}$$

with $\mathbf{B}$ being the derivative matrix for $\tilde{\eta}(\mathbf{x})$ at $\mathbf{x}_0$ (Lindgren et al. 2021). If we take this linearisation and input it into the model likelihood, we get

$$\bar{\pi}(\mathbf{y}|\mathbf{x}, \theta) = \pi(\mathbf{y}|\bar{\eta}(\mathbf{x}), \theta) \approx \pi(\mathbf{y}|\tilde{\eta}(\mathbf{x}), \theta) = \tilde{\pi}(\mathbf{y}|\mathbf{x}, \theta). \tag{3.13}$$

This is an approximation of the model likelihood (Lindgren et al. 2021). We input this approximation into the INLA algorithm, and perform inference on the model as before.

Because the `inlabru` approach adds another approximation to the INLA algorithm, the performance of the posterior estimates will depend on how well this linearisation step approximates the original predictor. If the predictor is highly non-linear, then this linearisation might not estimate the predictor well, and the posterior estimates might not be good enough to conduct meaningful inference on the model.

Chapter 4

Model specification and simulation study

We conduct a comprehensive simulation study to investigate if inlabru is suitable for conducting inference on the Poisson-Logistic model for severely under-reported count data. A simulation study using inlabru and the Poisson-Logistic model was conducted in Wøllo 2022. Here the results of performing inference on the Poisson-Logistic model with inlabru were compared to results using MCMC simulations implemented with the NIMBLE library in R. This simulation study treated the under-reporting rate as mild, setting the true reporting rate at $80\% - 95\%$. The results of this simulation study showed that both inlabru and NIMBLE managed to recover the true model well, and the usefulness of inlabru when conducting inference on such a model was confirmed.

Conducting a new comprehensive simulation study enables us to evaluate the usefulness of inlabru and the Poisson-Logistic model when the under-reporting rate is severe. It also allows us to investigate how the model performs on a weak spatial structure like Italy, that has a low number of regions. We do this by looking at a naive example where we manually connect the graph of Italy, and a more involved example where we fit the model using the disconnected graph of Italy. Lastly, we investigate how introducing noise into the under-reporting part of the model affects the model performance. This is done to reflect a real-world application, where a true covariate for the under-reporting might not be available, and a proxy is used instead. Using the results obtained in the simulations study, we discuss whether using `inlabru` to conduct inference on the Poisson-Logistic model can be appropriate in the context of modelling the rate of violence against women in Italy.

To assess the robustness of the Poisson-Logistic model and how sensitive it is to changes in the informative prior distribution on β_0, we also conduct a sensitivity analysis. We do this by changing the values of the informative prior distribution, assessing how this changes the model performance. We choose to focus the sensitivity analysis on modelling with the disconnected graph of Italy, assessing the model performance when applied to both noisy and non-noisy data.

4.1 The Models

For the simulation study, we look at three different cases and compare their performance. We do this by comparing model performance on three different spatial structures. For the three cases, we firstly look at a simple Poisson model and then we compare this to the performance of the Poisson-Logistic model, both with noisy data and data without noise.

This simple Poisson model will then show how a model performs under the naive assumption that the observed count data is in fact the true counts. By looking at an example where we introduce noise into the under-reporting covariate we emulate the real-data case. Here, data for the true under-reporting covariate might not be available and we use a proxy for this as a model covariate instead.

The naive Poisson model can be written on hierarchical form as

$$z_i \sim \text{Poisson}(E_i\lambda_i), \tag{4.1}$$

$$\log\big(\lambda_i\big) = a_0 + \sum_{k=1}^{K} \alpha_k x_{i,k} + \phi_i, \tag{4.2}$$

where z_i is the observed count in region i, λ_i is the expected count in region i, and E_i is an offset relating to population size in a region. The expected count λ_i is modelled through process covariates $x_{i,k}$ modelling the fixed effects and spatial effect ϕ_i modelling the random effect. α_0 is an intercept, and α_k are unknown model parameters.

We want to compare the performance of the naive Poisson model to the performance of the Poisson-Logistic model presented in Section 2.2. Following the discussion in Section 2.3, we can write the Poisson-logistic model as a thinned Poisson model. For the general case with random effects, this model can be written as

$$z_i \sim \text{Poisson}(E_i \pi_i \lambda_i), \tag{4.3}$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{J} \beta_j u_{i,j}, \tag{4.4}$$

$$\log(\lambda_i) = a_0 + \sum_{k=1}^{K} \alpha_k x_{i,k} + \phi_i. \tag{4.5}$$

Here, z_i, E_i, λ_i, α_0, α_k, $x_{i,k}$ and ϕ_i are defined as for the naive Poisson model. π_i is the reporting rate of an event, β_0 is an intercept for the under-reporting part of the model, β_j are unknown parameters and $u_{i,j}$ are covariates related to the under-reporting part of the model.

4.1.1 Models with a structured spatial effect

To capture any spatial variability, we include a spatial random effect ϕ into the models. The structured spatial effect is introduced into the model to explain how the observed count varies within the region of interest, and to pick up any spatial trends in the data. We choose to model the structured spatial effect as an Intrinsic Gaussian Conditional Autoregressive (ICAR) Model (Besag et al. 1991). The naive Poisson model in Equations 4.1-4.2

with one covariate x_i can then be written as

$$z_i \sim \text{Poisson}(E_i \lambda_i), \tag{4.6}$$

$$\log(\lambda_i) = \alpha_0 + \alpha_1 x_i + \phi_i, \tag{4.7}$$

and the Poisson Logistic model presented in Equations 4.3-4.5, with one process covariate x_i and one covariate u_i for the under-reporting part of the model can be written as

$$z_i \sim \text{Poisson}(E_i \pi_i \lambda_i), \tag{4.8}$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 u_i, \tag{4.9}$$

$$\log(\lambda_i) = \alpha_0 + \alpha_1 x_i + \phi_i. \tag{4.10}$$

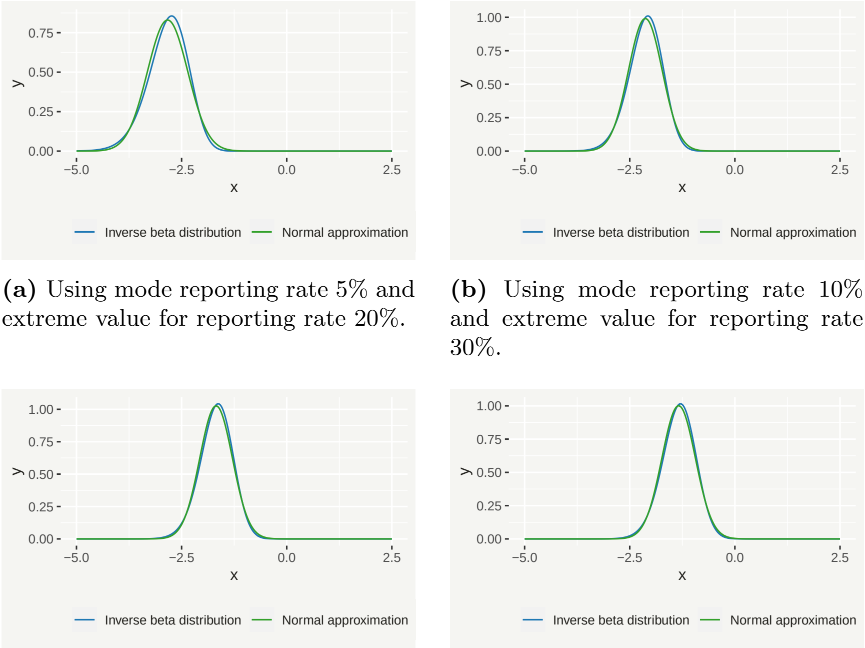
For spatial modelling, it is common to include an unstructured random effect, but as we are primarily interested in the application of the model on the sparse Italian graph with a small number of nodes, we have chosen not to include it. When performing inference using `inlabru` on the naive Poisson and the Poisson-Logistic model including both a structured and an unstructured spatial effect, `inlabru` struggled with convergence issues. It seems that the optimisation step of INLA struggled with separating the structured and the unstructured random effects for such a sparse spatial structure as the graph of Italy. Because of this, we have chosen not to include an unstructured random effect in this simulation study.

4.1.2 Prior specification for an informative prior when using `inlabru` and the INLA methodology

For the Poisson-Logistic model, the issues of identifiability of the model parameters have been discussed in Chapter 2. A solution to this problem has been discussed in Section 2.4.1, and this is the method we want to use in this simulation study. Chen et al. 2022 provides an `R`-function that numerically estimates the parameters of the beta distribution when the mode and a percentile value is given as input. We use this to elicit an appropriate beta prior for the reporting rate at the mean value for all the centred under-reporting covariates, which again is used to induce a prior for β_0 through the transformation shown in Equations 2.23-2.24.

However, there is a problem with this approach. In order to use the INLA methodology for inference, we assume that the Poisson-Logistic model can be written as a Latent Gaussian Model, discussed in Section 3.2. An LGM is a class of models where the latent field is Gaussian. This means that we want to assume a Gaussian prior distribution for all parameters in the latent field. This includes the intercept β_0. The derived prior distribution for β_0, presented in Equation 2.24, is not Gaussian. This means that we can not use this prior distribution directly when performing inference with INLA. Instead, we numerically approximate the mean and the standard deviation of the prior distribution on β_0 from the probability density function, and fit a normal distribution using this mean and standard deviation. To see if this is a reasonable approximation, we plot the originally induced probability density function for β_0 for four different values of mode reporting rate p_0, and then plot the probability density function of the Gaussian distribution estimated using the approximated mean and standard deviation in the same plots. This is shown in Figure 4.1. The four reporting rates shown are some of the reporting rates used for a sensitivity analysis on the β_0-prior later in this chapter.

The Gaussian approximation fits quite well, and we are therefore satisfied with using this as a prior distribution for β_0. We can then incorporate expert knowledge on the reporting rate for the mean of the centred covariates p_0 into the model through the informative prior on β_0 also when using the `inlabru` extension to the INLA methodology to perform inference in the Poisson-Logistic model.



(a) Using mode reporting rate 5% and extreme value for reporting rate 20%.

(b) Using mode reporting rate 10% and extreme value for reporting rate 30%.

(c) Using mode reporting rate 15% and extreme value for reporting rate 40%.

(d) Using mode reporting rate 20% and extreme value for reporting rate 50%.

Figure 4.1: Gaussian approximation to the non-Gaussian prior distribution for β_0 induced from the logistic relationship between β_0 and the reporting rate at the mean values of the centred covariates p_0, and the beta prior distribution on p_0 derived using expert knowledge on the likely reporting rate.

4.2 Data Simulation

To generate the true counts, we set the parameters for the count generating process as $\alpha_0 = 4$ and $\alpha_1 = 1$. For the under-reporting part of the model, we set parameters $\beta_0 = -2$ and $\beta_1 = 2$. The true value of β_0 is chosen to be -2 in order for the observed counts to reflect a model where the reporting rate is low. We simulate both covariates x_i and u_i from a $\text{Unif}(-1, 1)$ distribution. We need a noisy under-reporting covariate $\tilde{u}_i$ for the simulation study, and we define $\tilde{u}_i$ as a covariate having 0.6 correlation with the under-reporting covariate u_i. The population offset E_i is also simulated using a uniform distribution, but this time a $\text{Unif}(\text{Min(FP)}, \text{Max(FP)})$ is used, where FP is a list of the female population size in the regions of Italy. We do this to see how the models perform in environments similar to that of the issue of violence against women in Italy presented in Chapter 5.

The true structured spatial effect ϕ_i is simulated using a numerical method of simulating an ICAR(ν^2) field, presented in the `R`-code attached to Stoner et al. 2019. We set the standard deviation of the ICAR model to $\nu = 0.5$. We then use this synthetic data along with the true parameter values to generate the true expected mean count λ_i and the true reporting probability π_i, as well as the true observed count z_i and the true count y_i. This is done using the Poisson-Logistic model on its original form,

$$z_i | y_i, \pi_i \sim \text{Binomial}(y_i, \pi_i), \tag{4.11}$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 u_i, \tag{4.12}$$

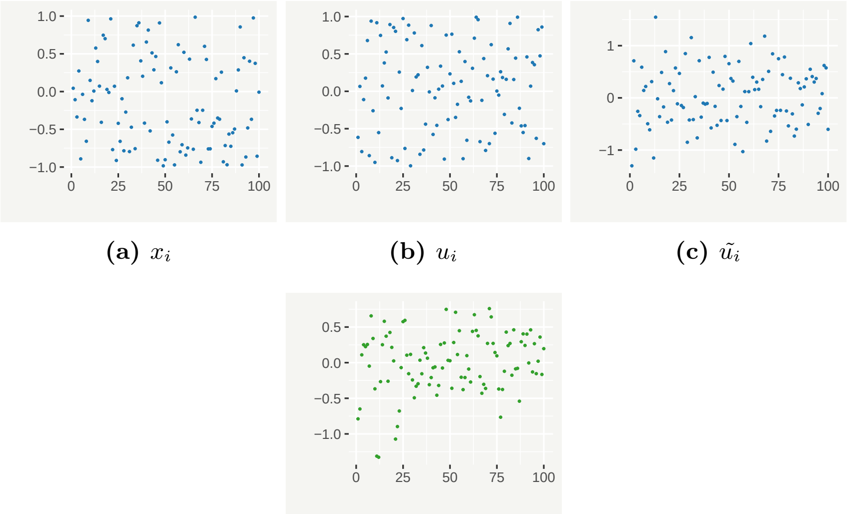
$$y_i | \lambda_i \sim \text{Poisson}(E_i \lambda_i), \tag{4.13}$$

$$\log(\lambda_i) = a_0 + \alpha_1 x_i + \phi_i. \tag{4.14}$$

As we are looking at three examples where the spatial fields are different, we generate data for all three of these spatial fields. A summary of the data generated for one of the simulations in Example 1, discussed later in Section 4.4.1, is shown in Figure 4.2.

4.3 The Spatial Structures

For the simulation study, we firstly look at a regular 10 × 10 lattice, and simulate the spatial effects over this regular lattice. The spatial structure of this lattice is shown in Figure 4.3. We chose this spatial structure as all points in the grid have more than one neighbour, making the spatial structure solid.



(a) x_i **(b)** u_i **(c)** $\tilde{u}_i$

(d) ϕ_i

Figure 4.2: The simulated data plotted against the grid points i.

As we are interested in the application of violence against women in the regions of Italy, having an example with the regions of Italy as the spatial structure would be beneficial. Differences in model recovery and results due to the spatial structures can then investigated. Italy is a long and thin country with many regions having only 0, 1 or 2 neighbouring regions. This weak spatial structure may be a disadvantage, as the regions are not able to borrow as much strength from each other as they would in other regional structures. Italy consists of only 20 regions in total, and the low number of regions might also impact on the modelling results.

4.3.1 Problems with ICAR models on disconnected graphs

All the models are using an ICAR spatial structure to model the spatial dependencies in the data. There is however some considerations that need to be given to the implementation of the spatial structure when using an ICAR model. Sterrantino et al. 2017 presents a spatial graph as a collection

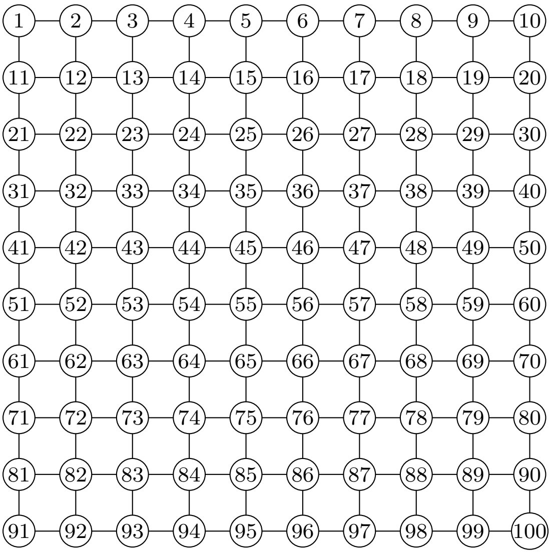


Figure 4.3: Graph of the 10×10 grid used for Example 1 in the simulation study

of nodes belonging to one or more connected components. If all nodes are belonging to only one connected component, we say that the spatial graph is connected. If the graph consists of more than one connected component, (of any size) the graph is disconnected.

If we relate this to the map of Italy, then the mainland is one connected component, as each region in the mainland is connected to at least one other region on the mainland. The Islands of Sardinia and Sicily are however not connected to any other node in the graph, and are therefore referred to as singletons, connected components of size 1. The graph of Italy then consists of three separate connected components, and is therefore disconnected. This leads to problems with the direct implementation of the ICAR model. As explained by Sterrantino et al. 2017, the variance of the ICAR model is directly proportional to the inverse of the number of neighbours n_i a region has. When singletons (regions where $n_i = 0$) are present in the graph, it leads to an infinite prior variance for the singleton, which in turn yields a constant, improper prior for the region. The constant prior for the region makes it difficult for the random effect to shrink to the global mean (Sterrantino et al. 2017). It can also lead to an improper posterior distribution. The marginal variance for the prior distribution depending on the number of neighbours a region has becomes an issue, as the prior for the different components of the graph then has different interpretations.

For our simulation study, we look at two different approaches to dealing with the issues of a disconnected graph of Italy. The first approach is to connect Sicily artificially to the mainland and remove Sardinia altogether, and the other is to take the steps suggested by Sterrantino et al. 2017 to implement the ICAR model on the disconnected graph of Italy.

The connected graph of Italy with 19 regions

For a naive approach where we want to keep the spatial structure simple, we decide to connect the Italian graph. Both Sicily and Sardinia are singletons, not connected to any other regions. Sicily lies close to the mainland, and the most of the traffic over to the mainland goes to the region of Calabria. Adding an artificial connection between these two regions therefore seems sensible. For Sardinia, the connection to the mainland is not as obvious. It lies equally close to several regions, and is connected to the mainland with many different ferries. There is not one connection between a region in the mainland that is more obvious. Correcting the spatial graph by adding artificial connections between singletons and other connected components of the graph is a common practice. We do however want to be careful, as Sterrantino et al. 2017 warns against adding connections between Islands and the mainland where there are not strong indications for doing so. With that in mind we don't artificially add any connections to link Sardinia to the rest of Italy. In order to have a connected graph for this naive approach, we decide to remove the island of Sardinia from the graph all together. We are then able to implement this naive approach on a connected graph that still represents the true spatial structure of Italy. The graph is seen in Figure 4.4.

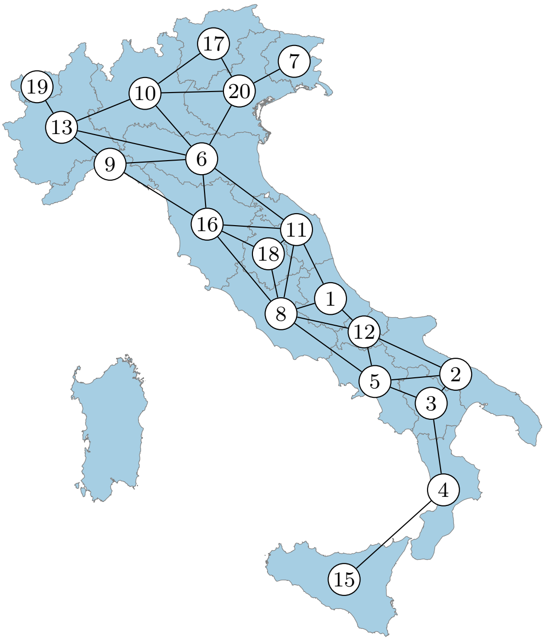


Figure 4.4: The connected graph of Italy used for example 2 in the simulation study. Sardinia is removed, and Sicily is attached to the mainland.

The disconnected graph of Italy with all 20 regions

In order to perform inference on the models using the original disconnected graph of Italy, we perform the steps outlined in Sterrantino et al. 2017 when conducting inference with `inlabru`. Sterrantino et al. 2017 recommends scaling the model, including a sum-to-zero constraint on the connected components of size larger than one and explicitly defining an intercept for all connected components larger than one. They also recommend adding the normalising constant that is scaled for the disconnected graph to the log marginal likelihood estimate.

Firstly, Sterrantino et al. 2017 recommends scaling the graph. This addresses the problem of the marginal variance for the prior distribution, and how this variance is different for each connected component of the graph, as it depends on the number of neighbours n_i a region has. This scaling solution was proposed by Sørbye et al. 2014, and entails scaling the precision matrix of the ICAR spatial model. We scale each connected component with size larger than one independently so that the geometric mean of the marginal variance for each connected component of the graph is 1. For the singletons, we replace the constant prior with a Gaussian prior. Then, the singletons are interpreted as a non-spatial random effect, as in Wakefield 2006. The result of this scaling is that the prior distribution for each connected component of the graph now has the same meaning, independently of the number of neighbours each region has. This makes defining a proper prior distribution possible, and we avoid the problems of an improper posterior distribution (Sterrantino et al. 2017). Next, Sterrantino et al. 2017 recommend including a sum-to-zero constrain on each of the connected components larger than one. This is because the scaled precision matrix of the ICAR model is not full rank for the disconnected graph. Furthermore, they also recommend explicitly defining the intercepts for the ICAR spatial effect. An intercept for the ICAR model is implicitly defined in the model, but to account for the effects of the disconnected graph, the recommendation is to define a separate intercept for each connected component of the disconnected graph with size larger than one. For the graph of Italy, this means that we define an intercept for the Italian mainland. No separate model intercept is needed for the Island of Sicily and Sardinia, as they are singletons.

Both scaling the model and including a sum-to-zero constraint are features available in `inlabru`, and is therefore straight forward to implement. Including a separate intercept is not implemented in `inlabru`, and this was done by looking at the implementation examples in Sterrantino et al. 2017. The normalisation constant is also not computed automatically for the ICAR model in `inlabru`, so this needs to be manually added to the log marginal likelihood estimates, using the implementation described in

Sterrantino et al. 2017. After performing these necessary steps, we can perform inference with `inlabru` using the disconnected graph of Italy. The spatial structure of this graph is shown in Figure 4.5.

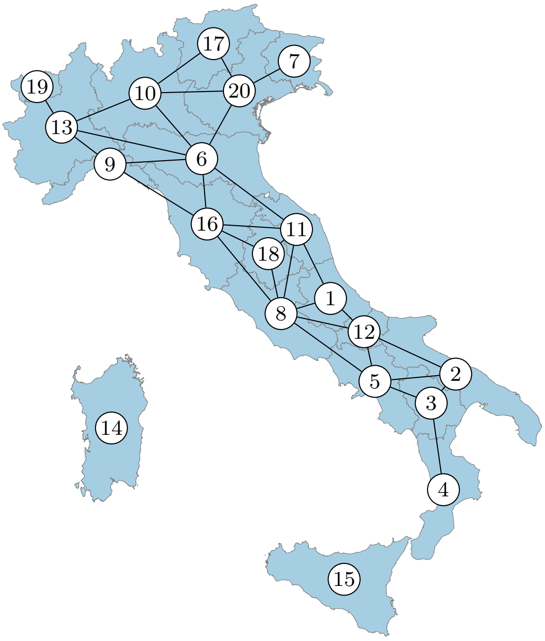


Figure 4.5: Graph of the disconnected graph of Italy used for example 3 in the simulation study. Sardinia and Sicily have no neighbours, and are therefore singletons.

4.4 Simulation study

For the simulation study, we are interested in three different cases. These are

$$\text{C1} \sim \text{Naive Poisson} \tag{4.15}$$

$$\text{C2} \sim \text{Poisson-Logistic model} \tag{4.16}$$

$$\text{C3} \sim \text{Noisy Poisson-Logistic model.} \tag{4.17}$$

Firstly, case 1 is the naive Poisson model presented in Equations 4.6-4.7. Then, case 2 is looking at the Poisson-Logistic model presented in Equations 4.8-4.10. For case 3, the same model as in case 2 are used, but we introduce noise into the under-reporting covariate u_i. This is done to mimic the real data application and the belief that we don't have access to a true under-reporting covariate u_i, but instead only a stand-in for this covariate, $\tilde{u}_i$.

For the simulation study, we need prior distributions for all fixed effects and hyperparameters of the models. Firstly, we need an informative prior distribution for β_0, and we choose this from expert knowledge of the likely reporting rate. We choose this likely reporting rate to be 10%, and set the 99.99th percentile to be 30%. This elicits a $\text{N}(-2.130, 0.403^2)$ prior distribution for β_0. The other unknown parameters α_0, α_1 and β_1 are given an uniformative prior distribution of $\text{N}(0, 10^2)$. For the random effect, we assign a $\text{Gamma}(1, 0.0005)$ prior distribution to the ICAR structured spatial effect ϕ.

For each of the three examples, we simulate 100 data sets m, $\{m = 1, \ldots, 100\}$. We report on the average posterior bias, the root mean squared error and the coverage probability of the 95% credible intervals for all model parameters, as well as the mean expected count $\lambda = \frac{1}{n}\sum_{i=1}^{n}\lambda_i$ and the mean reporting probability $\pi = \frac{1}{n}\sum_{i=1}^{n}\pi_i$.

Average Bias

The bias is a measure of the difference between the expected value of an estimator and the true value. Here, the bias for the model parameters is the difference between the posterior mean value of the unknown parameter and the true value. For example, the bias for the process intercept for simulation m would be

$$\text{Bias}(\alpha_{0,m}) = \hat{\alpha}_{0,m} - \alpha_0, \tag{4.18}$$

where α_0 is the true value, and $\hat{\alpha}_0$ is the estimated posterior mean. To get the average bias we average over all the simulations, giving

$$\text{Average Bias}(\alpha_0) = \frac{1}{100}\sum_{m=1}^{100} \hat{\alpha}_{0,m} - \alpha_0. \tag{4.19}$$

To get the average bias for the expected count λ_i and the reporting probability π_i, we follow the approach given above in Equations 4.18 and 4.19. We are then left with the average bias for each node i of the spatial structure. In order to summarise this calculate the mean bias across all the regions i and report this average value.

Root Mean Squared Error (RMSE)

We also report the root mean squared error for all model parameters as well as λ and π. The root mean squared error is defined as the square root of the average of the squares of the bias. This mean we square the bias for the parameter estimate in each simulation, and then average over all the simulations. For α_0, the root mean squared error is defined as

$$\text{RMSE} = \sqrt{\frac{1}{100}\sum_{m=1}^{100} (\hat{\alpha}_{0,m} - \alpha_0)^2}. \tag{4.20}$$

Again, we calculate the RMSE of λ and π separately for each region i, and calculate the average.

Coverage probability

The coverage is referring to the coverage probability of the 95% credible interval of an estimator. The credible interval is a calculated interval, where we estimate that there is a 95% probability that the true value lies within this interval. We calculate this credible interval (CI) from the posterior distribution of each model parameter, as well as λ and π. From this, we check for each case whether the true value actually lies in the credible interval, expecting it to lie in the interval 95% of the time. If the coverage is lower than 0.95, this suggests that the model is struggling to obtain an accurate posterior distribution for the parameter. If the coverage is higher than 0.95, this can indicate that the variance of the posterior estimate is very high, creating a larger than expected 95% credible interval. The coverage for α_0 is calculated as

$$\text{Coverage}(\alpha_0) = \frac{1}{100} \sum_{m=1}^{100} \text{Coverage}_m, \tag{4.21}$$

where Coverage_m is written as

$$\text{Coverage}_m = \begin{cases} 1, & \text{if } \alpha_0 \text{ lies within the credible interval of } \hat{\alpha}_{0,m}, \\ 0, & \text{if } \alpha_0 \text{ lies outside the credible interval of } \hat{\alpha}_{0,m}. \end{cases} \tag{4.22}$$

4.4.1 Example 1: Inference on a 10×10 grid

For the simplest of the spatial structures, the 10×10 grid, the results from running the simulation study is shown in Table 4.1.

Parameter	α_0	α_1	β_0	β_1	λ	π
True Value	4	1	-2	2	–	–
			Average Bias			
C1	-2.189	$7.9e-05$	–	–	-58.67	–
C2	0.058	0.007	-0.052	0.012	142.3	$-1.0e-04$
C3	-0.119	0.008	0.194	-0.734	81.4	0.031
			RMSE			
C1	2.192	0.207	–	–	75.32	–
C2	0.404	0.045	0.473	0.114	1512	0.004
C3	0.875	0.161	1.118	0.807	391.1	0.202
			Coverage			
C1	0.00	0.89	–	–	0.00	–
C2	0.93	0.94	0.96	0.98	0.93	0.96
C3	0.92	0.93	0.96	0.63	0.87	0.91

Table 4.1: Results from 10×10 grid. Summaries of average bias, root mean squared error and coverage for the model parameters for the three cases C1, C2 and C3.

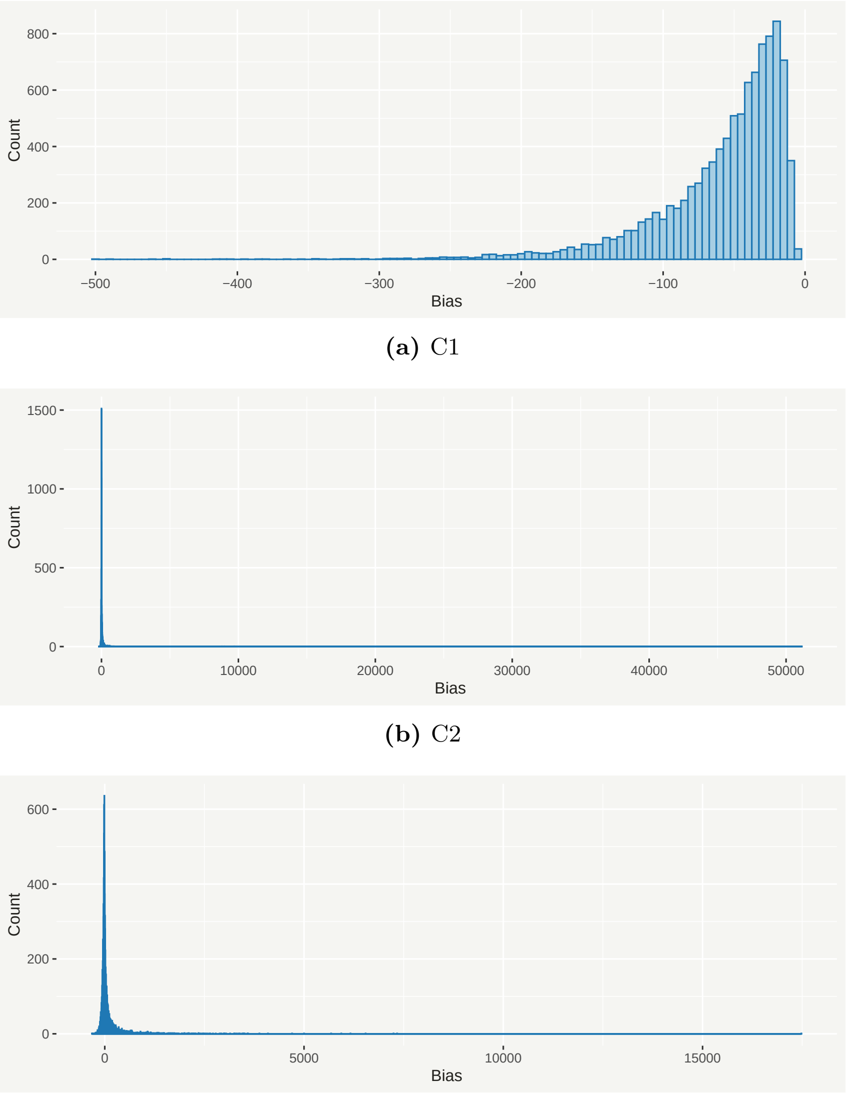
Looking at the intercept α_0 relating to the count process, we see that the naive Poisson case C1 performs badly. The average bias and RMSE is much higher for C1 than C2 and C3, and the coverage is 0. This shows that when not correcting for under-reporting on the 10×10 grid, the mean of the posterior distribution on α_0 is far too low. This again indicates that the model will consistently predict a posterior count that is lower than the true count y_i, making the model severely biased. We see this from the posterior samples of λ, which have a large negative posterior bias. When comparing the posterior estimates of α_0 of C1 with C2 and C3, the impact of correcting for under-reporting becomes obvious. The average posterior bias for the model parameters of C2 and C3 are much smaller than that of C1. The coverage of the credible intervals of α_0 from C2 and C3 is 0.92 and 0.93, indicating that the models correcting for under-reporting manages to recover the true value of α_0 quite well.

Next, looking at all the parameters returned for C2 and C3 we see that the average bias is low for both cases. Comparing the two, we do see that the bias and and RMSE increases for most parameters when noise is introduced through C3. This is most prominent in the β_1 parameter. This is not surprising, as parameter is directly impacted by the noise introduced in the under-reporting covariate. We also see how the coverage for β_1 drops from 0.98 for C2 down to 0.63 for C3. For the rest of the model parameters, α_0, α_1 and β_0, the coverage values look good for C2 and C3. We do however see that the coverage values for λ and π is significantly lower for C3 than for C2. The values are still acceptable, but below the desirable 95% coverage. This tells us that the noisy Poisson-Logistic model struggles more compared to the non-noisy model.

Both C2 and C3 perform well for all parameters and π, returning low bias and RMSE as well as a high coverage. The behaviour of the bias and RMSE for λ is however strange. As discussed, the coverage values looks good. When it comes to the bias and RMSE, results are not as expected however. C2 has average bias 142.3 and RMSE 1512. This is very high, and does not reflect the coverage results showing the model performs well. Looking at C3, the results are similar here. The average bias is also here higher than C1, and the RMSE is more than five times higher than C1. Still, C3 returns a coverage of 0.87, and C1 has coverage 0. This indicates that something unexpected has happened, and this needs to be investigated further.

Investigating the posterior estimates for λ

We look at all posterior λ estimates returned for each simulated data set, and find some unexpected results. A low number of simulations return very high posterior estimates for λ, which in turn gives high bias and squared error estimates. Because these estimates are so skewed, they significantly influence the results averaged over all simulations showed in Figure 4.1. To see the magnitude of the problem, we look at all the λ bias results returned in the 100 runs of the simulation study for C1, C2 and C3, for each region i. This is showed in Figure 4.6.



(a) C1

(b) C2

(c) C3

Figure 4.6: Histogram of all bias for all posterior estimates of λ_i for all 100 simulated data sets, for cases C1, C2 and C3.

4.6b and 4.6c clearly shows that the bias estimates from C2 and C3 are mostly low, but the histograms show a very large tail, with a few very high bias values. The bias estimates returned from C1 also have some high values, but far smaller than the results from C2 and C3. To investigate it this is a large problem, influencing most of the posterior estimates, we also report the median bias values from C1, C2 and C3. These results are shown in Table 4.2.

Parameter	C1	C2	C3
Mean λ	-58.67	142.3	81.4
Median λ	-44.41	2.200	-5.496

Table 4.2: Mean and Median bias for C1, C2 and C3 on the 10×10 grid.

Here, we see that the median bias of C2 and C3 is very different from the average bias, and that the median bias for λ is much lower for C2 and C3 than for C1. This, along with the histograms seen in Figure 4.6, indicates that there are a small number of runs returning these very highly bias estimates. Overall, as there is only as small number of simulations that return highly biased estimates and the coverage values for λ look good for both C2 and C3, we conclude that these strange results will not impact modelling results as long as we take necessary precautions and ensure that the model has converged. More on this in Section 4.4.4.

We can therefore conclude that when comparing all models on the 10×10 grid, the naive Poisson model is not suitable to model under-reported count data. Both the case using the Poisson-Logistic models with no noise and the case with noise performs well in most cases, but as expected the model with no noise performs better.

4.4.2 Example 2: Inference on the connected graph of Italy

The results from performing inference on the connected graph of Italy are shown in Table 4.3.

Parameter	α_0	α_1	β_0	β_1	λ	π
True Value	4	1	-2	2	$-$	$-$
			Average Bias			
C1	-2.177	-0.034	$-$	$-$	-60.53	$-$
C2	0.314	-0.035	-0.297	0.038	170.36	0.001
C3	-0.611	-0.023	1.019	0.839	$6.6e+11$	0.167
			RMSE			
C1	2.190	0.490	$-$	$-$	88.95	$-$
C2	0.967	0.152	1.109	0.303	954.9	0.019
C3	0.884	0.361	1.569	1.621	$7.5e+12$	0.298
			Coverage			
C1	0.00	0.93	$-$	$-$	0.00	$-$
C2	0.92	0.93	0.98	0.96	0.93	0.96
C3	0.80	0.93	0.97	0.99	0.79	0.90

Table 4.3: Results from connected graph of Italy, using 19 regions. Summaries of average bias, root mean squared error and coverage for the model parameters for the three cases C1, C2 and C3.

We see the same general trends here as we did on the 10 × 10 grid. The naive Poisson model is still suffering a large bias, and is not suitable for modelling under-reported count data. Focusing on C2 and C3, the posterior bias for the parameters are in general still higher for C3 that for C2. When looking at the coverage probability for α_0, we see a difference between C2 and C3. Here, the coverage is significantly lower when using the noisy model, C3, than when using model C2. The coverage for C2 is still okay, albeit a little lower than for Example 1. When using C2 the coverage is 0.92. For C3 however, the coverage lies at 0.80. Another noteworthy difference is the coverage for β_1 for the noisy Poisson-Logistic model C3, which is is higher here than in Example 1. The reason for this is unclear. When comparing C2 and C3, the coverage probability for λ is lower for C3. This is again different from Example 1, where the coverage was more similar for C2 and C3.

We again see that the RMSE for C2 and C3 is much larger than expected. These results are similar to what we saw in Example 1, and when further investigating this, we see that this is a result of the same modelling trouble as in Example 1. A small number of runs return very large bias estimates, whereas most runs return good estimates. For C3, the average bias and RMSE values are extremely large. We discuss reasons as well as possible solutions in Section 4.4.4, but again conclude that this will not affect the overall model performance if necessary percussion is taken. The median bias for λ is shown in Table 4.4, showing that C2 and C3 return good estimates for most on the runs in the simulation study.

Parameter	C1	C2	C3
Mean λ	-60.53	170.36	$6.6e+11$
Median λ	-43.62	3.052	1.186

Table 4.4: Mean and Median bias for C1, C2 and C3 on the connected graph of Italy.

Overall, we again see that the naive Poisson model, C1, is not suitable for the task of modelling under-reported count data. We also see that the noisy Poisson-Logistic model C3 performs worse that the model without noise C2 when modelling on the connected graph of Italy.

4.4.3 Example 3: Inference on the disconnected graph of Italy

The results from performing inference on the disconnected graph of Italy where all 20 regions are included are shown in Table 4.5. Because of the disconnected graph, we add a region-specific intercept α_{cc} on the ICAR spatial model, as discussed in Section 4.3.1. Posterior estimates for this intercept is also seen in Table 4.5.

Parameter	α_0	α_{cc}	α_1	β_0	β_1	λ	π
True Value	4	0	1	-2	2	–	–
			Average Bias				
C1	-2.198	-0.127	-0.064	–	–	-59.77	–
C2	0.152	-0.076	-0.004	-0.091	0.081	97.28	0.018
C3	0.396	0.294	-0.080	-0.556	-0.479	581.1	-0.039
			RMSE				
C1	2.346	0.126	0.506	–	–	79.42	–
C2	0.934	0.077	0.138	1.069	0.345	454.9	0.148
C3	0.974	0.295	0.366	1.399	1.225	2517	0.192
			Coverage				
C1	0.44	1.00	0.91	–	–	0.00	–
C2	0.93	1.00	0.91	0.91	0.97	0.87	0.90
C3	0.99	1.00	0.91	1.00	0.95	0.98	0.98

Table 4.5: Results from disconnected graph of Italy, using all 20 regions. Summaries of average bias, root mean squared error and coverage for the model parameters for the three cases C1, C2 and C3.

For this example, we see that the results from C1 again are similar to Example 1 and 2, and does not lend itself to useful modelling of under-reported count data. As for C2 and C3, the posterior bias of the parameters are similar to that of Examples 1 and 2. The bias is somewhat higher for C3 than for C2, as we also saw in the previous examples. What is most different with Example 3 compared to Example 1 and 2 is the coverage results for C3. The coverage for all parameters and the spatial effects of C2 are similar to the results of example 2, which is to be expected as the spatial graph is similar, with a similar structure and number of regions. For C3 however, the coverage probabilities are very good, and better than for C2. The coverage probabilities for the model parameters $\alpha_0, \alpha_1, \beta_0$ and β_1, as well as the coverage for the posterior of λ and π in Example 3 are also higher than for C3 in Example 1 and 2. This is surprising, and the reason for this is unclear. We still see the same strange results for the average bias and RMSE of the posterior estimate of λ as in Example 1 and 2, and the median bias values for λ are shown in Table 4.6. From this, we again conclude that this is due to a small number of runs, and that this does not impact the overall model performance noteworthy. We do however see that the median bias for C3 is now quite large. This is discussed in further detail in Section 4.4.4.

Parameter	C1	C2	C3
Mean λ	-59.77	97.28	581.1
Median λ	-43.93	-0.375	49.59

Table 4.6: Mean and Median bias for C1, C2 and C3 on the disconnected graph of Italy.

4.4.4 Discussion

Overall, the simulation study shows that performing inference on the Poisson-Logistic model using `inlabru` returns good results. We implemented the model with a structured spatial effect ϕ. We chose not to include an additional unstructured random effect, as this made `inlabru` struggle with convergence issues for the weaker spatial structures of Example 2 and 3. The naive Poisson model returns estimates with a large negative bias, indicating that it is not suitable for inference on severely under-reported count data. The Poisson-Logistic model takes this under-reporting into account, and returns far more accurate estimates. We see that the model performance is not as good when introducing noise into the model, but the model coverage is still acceptable. When comparing the three examples, we see that the 10×10 grid spatial structure performs best. This is not surprising as this is the strongest spatial structure, with the largest amount of nodes. The disconnected graph of Italy returns weaker results than the 10×10 grid, but the results are still very good. This suggests that this spatial structure can be used in model application. Lastly, the disconnected graph of Italy also returned good results, particularly for the noisy model. The reason for this is unsure. The results for the non-noisy model were acceptable, suggesting that this spatial structure is strong enough to recover the true model parameters. This leads us to conclude that the Poisson-Logistic model with an informative prior distribution on β_0, even with noisy data for under-reporting, is strong enough to perform inference. All three spatial structures were strong enough for good posterior estimates to be returned, suggesting that we can use Poisson-Logistic model to perform inference on the connected or the disconnected graph of Italy.

We did however see some unexpected results for the posterior estimates of λ from C2 and C3 in this simulation study. The coverage results looked acceptable, which led us to believe that a small number of simulations might skew the average bias and RMSE returned from all $m = 100$ simulations. After investigating this, and studying the bias returned from all regions i for all m simulations, we conclude that a small number of runs give severely skewed posterior estimates for λ. To understand the possible reasons for this, we take a closer look at the affected simulations. We see that the posterior distribution for both the process intercept α_0 and the under-reporting intercept β_0 have a much higher variance here than for the runs where the bias is low. The model also struggles with recovering the true value for α_0, returning a posterior mean that is much higher than the true value. We also notice that `inlabru` struggles more with convergence of the optimisation step, and requires a larger number of iterations to converge than for runs returning good estimates. This information makes us suspect that the model in some cases is still not totally determined, even with an informative prior distribution on β_0. `inlabru` seems to struggle with separating and identifying the effects coming from the two different parts of the model. This again seems to result in posterior distributions for α_0 and β_0 that are fairly flat, making the parameters highly variable. The reason why this is seen in the posterior estimates for λ, but not for α_0 and β_0, is because of the exponential relationship between λ and α_0.

When the posterior estimates of α_0 are higher than the true value, with a high variance, this can result in very large posterior estimates for λ. These results show that there are instances where `inlabru` struggles to identify the different parts of the model, even with an informative prior distribution on β_0. One possible solution to this is to introduce an informative prior distribution also on the process intercept α_0. It is worth noting that these results were not seen in Wøllo 2022, which might mean that `inlabru` struggles more with identifiability when the rate of under-reporting is severe. In this investigation, we saw that `inlabru` only struggled in a few of the simulations, something reflected in the median bias estimates returned for λ. Consequently we do not deem it necessary to take any further steps in order to apply the Poisson-Logistic model to a real application. This is however something that needs to be investigated more in further works, to ensure the reliability of the Poisson-Logistic model with severely under-reported data.

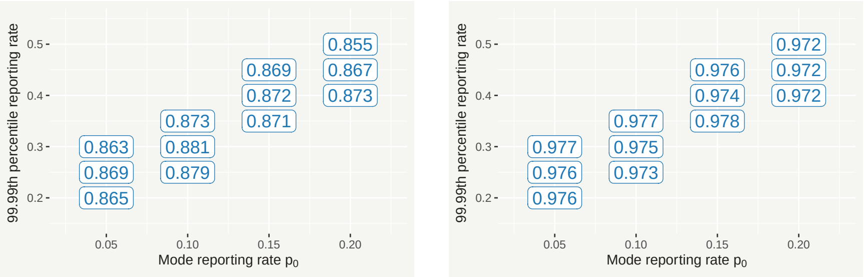
4.5 Sensitivity analysis for model prior selection

To investigate how the Poisson-Logistic model is affected by the prior distribution on β_0, we conduct a sensitivity analysis. If a model is very sensitive, then small changes in the prior distribution of β_0 will lead to the posterior estimates changing. This is not desirable, as this makes for a volatile model very governed by how the prior distribution on β_0 is defined. This can lead to poor model performance. The contrast would be if the model is robust. Then it will be able to stand up to changes in the β_0-prior, and will therefore be less dependent on its prior being defined exactly correct for the model to produce sensible posterior estimates. To investigate the sensitivity of the model, we run the simulation study several times, changing only the prior distribution on β_0. To define which prior distributions would be interesting to use in the sensitivity analysis, we look to the application of the model on the issue of violence against women in Italy. Estimates suggest that the likely reporting probability of incidents of violence lies somewhere between 5% and 20%, as described in Chapter 1. From this expert knowledge we induce a Gaussian prior for β_0 in the same way as in Section 4.1.2. The approximated parameters for this Gaussian prior is seen in Table 4.7.

Expert knowledge		Parameters for the approximated Gaussian prior on β_0	
Mode (%)	99.99th percentile (%)	Mean	Standard deviation
5	20	-2.833	0.481
5	25	-2.779	0.576
5	30	-2.721	0.658
10	25	-2.154	0.323
10	30	-2.130	0.403
10	35	-2.101	0.475
15	35	-1.697	0.324
15	40	-1.680	0.389
15	45	-1.660	0.451
20	40	-1.363	0.279
20	45	-1.351	0.339
20	50	-1.337	0.399

Table 4.7: Transformation of expert knowledge values on the reporting rate at the average of the centred covariates p_0 into parameters values for the informative Gaussian prior on β_0

We choose to focus the sensitivity analysis on the disconnected graph of Italy, as we will be using this spatial structure when modelling the rate of violence against women in Italy in Chapter 5. We look at the performance of the Poisson-Logistic model with and without noise added to the under-reporting part of the model. The noisy model is included to simulate a real-data application, where we might not have assess to the true under-reporting covariate and instead use a proxy. The results from the sensitivity analysis is shown in Figure 4.7.



(a) C2, modelling without noise

(b) C3, modelling with noise

Figure 4.7: Coverage for λ plotted on the grid with the expert knowledge values used to induce a prior distribution for β_0. Modelling with the disconnected graph of Italy, including all 20 regions.

Looking at models C2 and C3, we see that the sensitivity analysis returns high coverage values for all the prior distributions of β_0. This indicates that both models are robust with regards to changes in the informative prior distribution on β_0 on the disconnected graph of Italy. We see that the coverage values for λ returned from the noisy model C3 is higher that the coverage values returned from the non-noisy model C2. This is similar to the results seen in the simulation study, and the reasons for this is still not clear. One possible explanation for this is that the variance of the posterior estimates are higher for the noisy model than the non-noisy model. This will give a wider credible interval for λ, which in turn can result in a higher coverage probability. This is however not possible to conclude without further investigations. Based on the coverage for λ, we can conclude that both models are robust, and perform well for a range of prior distributions on β_0. This is desirable, as it shows that the model is less dependent on a perfectly accurate specification of the β_0-prior. Again, this suggests that we can use this model to model the rate of violence against women in Italy.

Chapter 5

Application on incidence rate of violence against women in Italy

Italy is a long and thin country consisting of 20 regions. The country we today know as Italy was unified in the years between 1861 and 1870, with the coming together of the southern and northern parts, with Venice and Rome being the last parts of the country to unify. Ever since this unification, the north-south differential in Italian development has been an issue Abramo et al. 2016. The northern regions are still to this day more prosperous than the southerns regions of Italy. In 2019, the average GDP per capita in the north-east of Italy was around 37000 Euro, with the average GDP per capita in the southern regions of Italy being around 17000 Euro Istat 2021. This divide shows no sign of slowing down or reversing, as the economic development is still stronger in the north than the south Istat 2021. Some of the regions of Italy are large, consisting of both larger cities and rural areas. Other regions such as Acosta Valley are smaller in size, with low population density. Overall, the population density is higher along the coast and in the larger cities, and lower in the far north as well as in areas further south far from the coast.

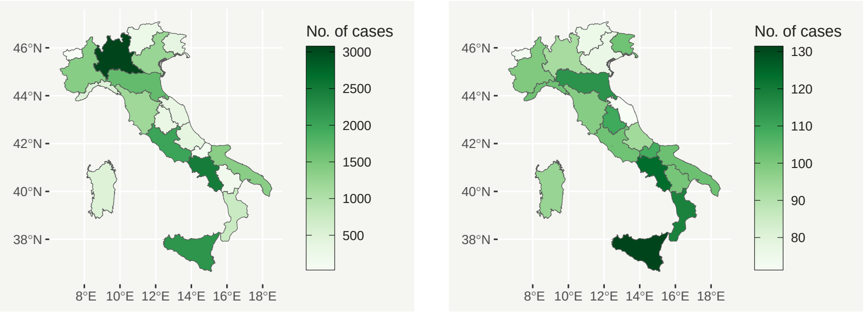
5.1 The Data

For this investigation, we use both registry data and survey data. The data used is gathered and compiled by ISTAT - The Italian National Institute of Statistics. ISTAT aims to produce statistics "in the service of citizens and policy-makers" (Istat 2022a), and is the largest producer of official statistics in Italy.

Response variable

As the response variable we consider the number of crimes against women registered by the Italian police in 2020. This data is gathered and published by ISTAT (Violenceagainstwomen.Stat 2022a). Crimes are divided into four categories: "Intentional homicide", "Battering", "Stalking" and "Rapes". We group these crimes into one category, as they are all considered violence against women according to the definition by the United Nations (United Nations 1993). The data is aggregated at a regional level, as privacy concerns prohibits the police from making data at a more detailed spatial aggregation level.

In this investigation, we further narrow the response variable by looking only at women in between the ages of 14 and 65. The total numbers and the incidence rate per 100000 women are plotted in Figure 5.1. We see that the crime incidence rate, shown in Figure 5.1b is slightly higher in the south of Italy than in the north of Italy. The region of Emilia-Romagna is an exception to this, as it is located in northern Italy, but still has among the higher incidence rates of crimes against women.



(a) In total. **(b)** Per 100,000 women.

Figure 5.1: Number of registered crimes against women in 2020. Total numbers to the left, and numbers per 100,000 women to the right.

Model covariates

The Italian National Institute of Statistics publishes "The report on equitable and sustainable well-being (BES)" every year, as well as an updated set of indicators (Istat 2022b). The report includes data on 153 indicators divided into 12 categories. These categories include health, security, environment and education and training. The data the indicators are calculated from comes from different sources. These sources include population surveys conducted by ISTAT, administrative registers and data collected by other governmental bodies and public agencies. We have looked more closely at some of the indicators in the report, and investigated whether they could be used as covariates in the model to explain the spatial variation of the registered incidence rate of gender crimes throughout Italy. Seven indicators from the BES report were chosen as model covariates, with advice from Silvia Polettini (Polettini 2022) and Serena Arima (Arima 2022), and an overview of these covariates are seen in Table 5.1, with indicator descriptions in Table 5.2. All seven of these indicators were gathered from yearly surveys conducted by Istat.

To model the spatial variability in the under-reporting part of the model, we also include a covariate called helpline calls. This data is also gathered by ISTAT, and published in the ViolenceAgainstWomen.Stat-database (Violenceagainstwomen.Stat 2022b). The covariate is explained in more detail in Table 5.2, and an overview of the covariate is shown in Table 5.1. The model assumption here is that this helpline is accessible all over Italy, and it is more likely for someone to call this helpline when experiencing violence than reporting it to the police.

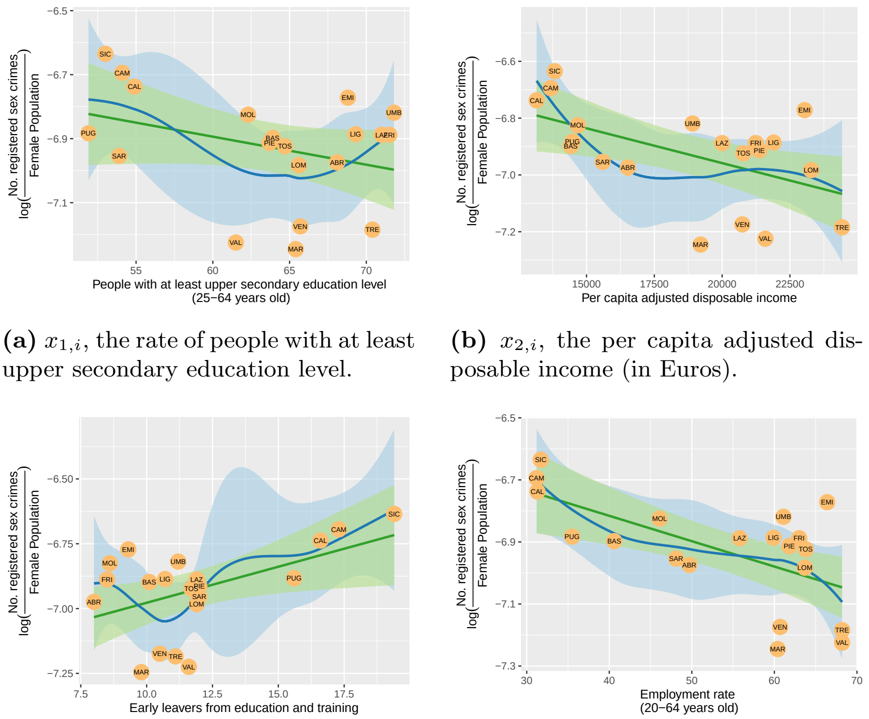
Overview Indicators						
Indicator	Indicator Name	Min	Max	Mean	Median	SD
$x_{1,i}$	Upper Secondary Education (%)	51.3	71.3	63.1	64.6	6.61
$x_{2,i}$	Per Capita Adjusted Disposable Income (Euro)	13160	24423	18643	19602	3710
$x_{3,i}$	Early Leavers From Education And Training (%)	8.4	22.4	12.6	11.2	4.1
$x_{4,i}$	Employment Rate (%)	44.5	76.6	63.8	68.5	10.8
$x_{5,i}$	People at risk of poverty (%)	6.1	41.4	18.8	14.0	10.8
$x_{6,i}$	Severe Material Deprivation Rate (%)	1.7	17.8	7.1	4.7	4.9
$x_{7,i}$	Alcohol consumption (%)	9.6	23.4	17.9	17.9	3.7
u_i	Helpline calls	68	9292	2635	1406	2521

Table 5.1: Overview of indicators used as covariates in the model.

Description Indicators	
Indicator	Description
$x_{1,i}$	People in the age group $25-64$ years old that has completed at least upper secondary education.
$x_{2,i}$	A ratio between the total disposable income of a consumer household, and the number of residents in the household.
$x_{3,i}$	The population (in the age group $18-24$) who have not completed upper secondary education, and is currently not included in a training program.
$x_{4,i}$	The population (in the age group $20-64$) who are currently employed.
$x_{5,i}$	The population that has an equivalised income less than or equal to 60% of the equivalised income in the region. Equivalised income as defined by eurostat is a measure of household income adjusted for differences in household size and composition (Eurostat 2021).
$x_{6,i}$	The share of the population that that lives in a household that is unable to fulfil at least four of the following nine scenarios: 1) paying rent or utility bills, 2) keeping the home adequately warm, 3) facing unexpected expenses (in 2014, this rate was set to 800 Euros), 4) eating meat, fish or another equivalent protein source at least every second day, 5) a week holiday from home, or that is unable to afford 6) a car, 7) a washing machine, 8) a colour TV or 9) a telephone.
$x_{7,i}$	Share of population, aged 14 or older, showing risk behaviour connected to alcohol use. These risk factors are determined in agreement with the Italian National Institute of Health, taking into consideration the definitions adopted by the World Health Organisation (WHO) and recommendations from the Italian National Institute on Food and Nutrition (INRAN). "At-risk" consumers are people with at least one of the risk behaviours: exceeding the daily consumption of alcohol or consuming six or more units of any alcoholic beverage on one occasion (binge drinking).
u_i	Recorded number of telephone calls to the National Hotline Service 1522, where the caller was experiencing gender related crime, or called on behalf of someone experiencing this. This helpline is open 24/7, and can be accessed anonymously (1522 2022).

Table 5.2: Description of indicators used as covariates in the model.

In Figure 5.2, the chosen model covariates are plotted against the log of the response data divided by the female population in each region, with each point in the plot representing a region of Italy. A linear and a smoothed trend line is added to the plots. We see no indication that the smoothed trend line pick up a trend in the data better than the linear trend line.



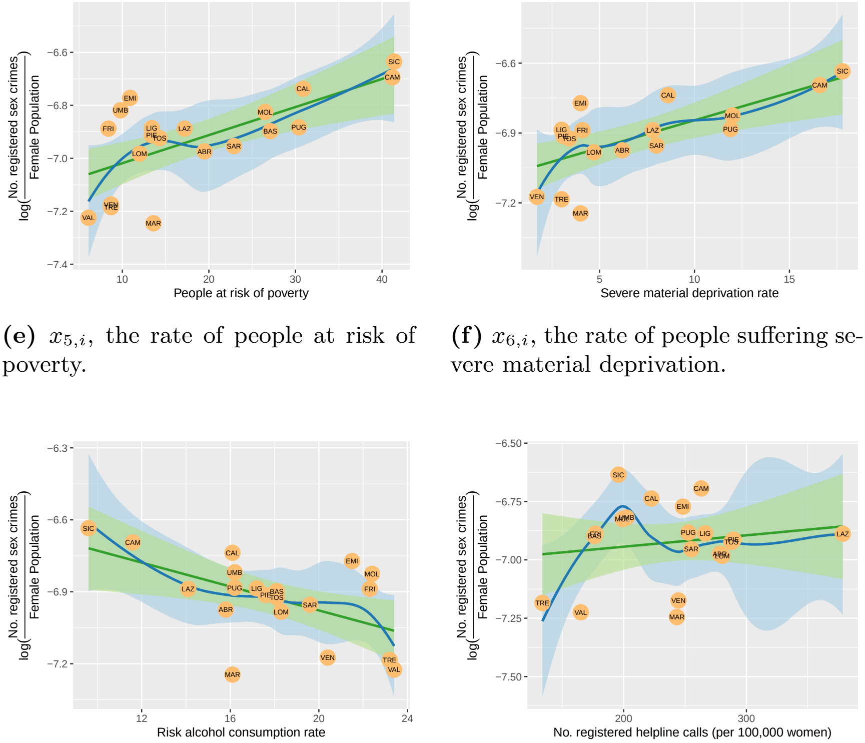
(a) $x_{1,i}$, the rate of people with at least upper secondary education level.

(b) $x_{2,i}$, the per capita adjusted disposable income (in Euros).

(c) $x_{3,i}$, the rate of early leavers from education and training

(d) $x_{4,i}$, the employment rate.

Figure 5.2: Scatter plots of the log of the response data plotted against model covariates, with linear and smoothed trend line fitted. (Figure continues on next page)



(e) $x_{5,i}$, the rate of people at risk of poverty.

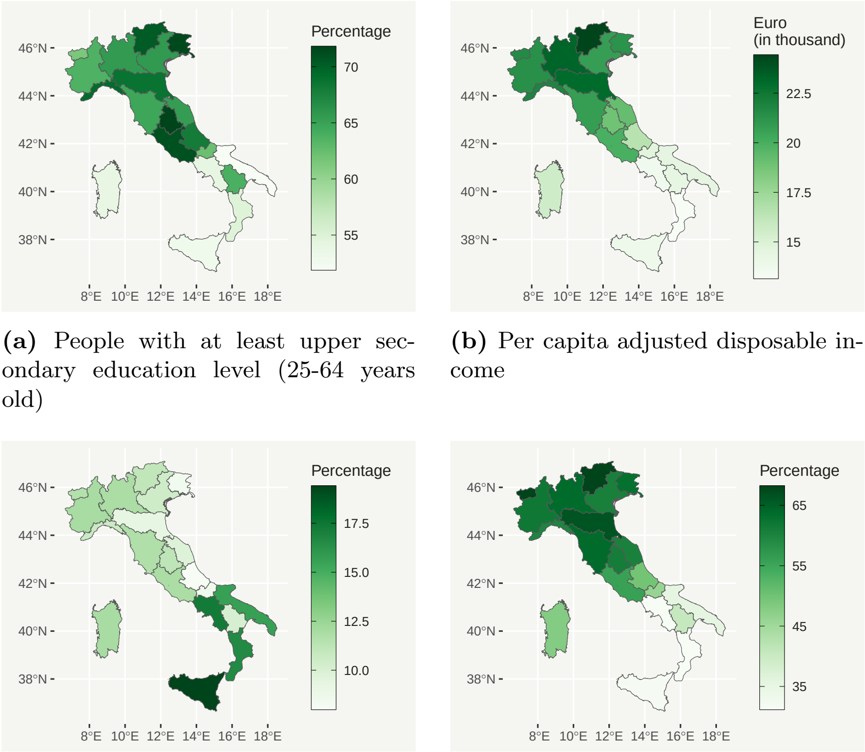
(f) $x_{6,i}$, the rate of people suffering severe material deprivation.

(g) $x_{7,i}$, the rate of people showing at least one risk behaviour connected to alcohol consumption.

(h) u_i, the number of women per 100000 calling the violence helpline 1522.

Figure 5.2: Scatter plots of the log of the response data plotted against model covariates, with linear and smoothed trend line fitted. (Figure continued from previous page)

All the model covariates are again plotted on a map of Italy in Figure 5.3. There seems to be a spatial trend to the covariates, where the education level, income and employment rate is higher in northern Italy, and the rate of early education leavers as well as the risk of poverty and material deprivation is higher in the southern parts of Italy. This trend seem to follow the mentioned north-south Italian economic divide, with the northernmost regions being more prosperous.



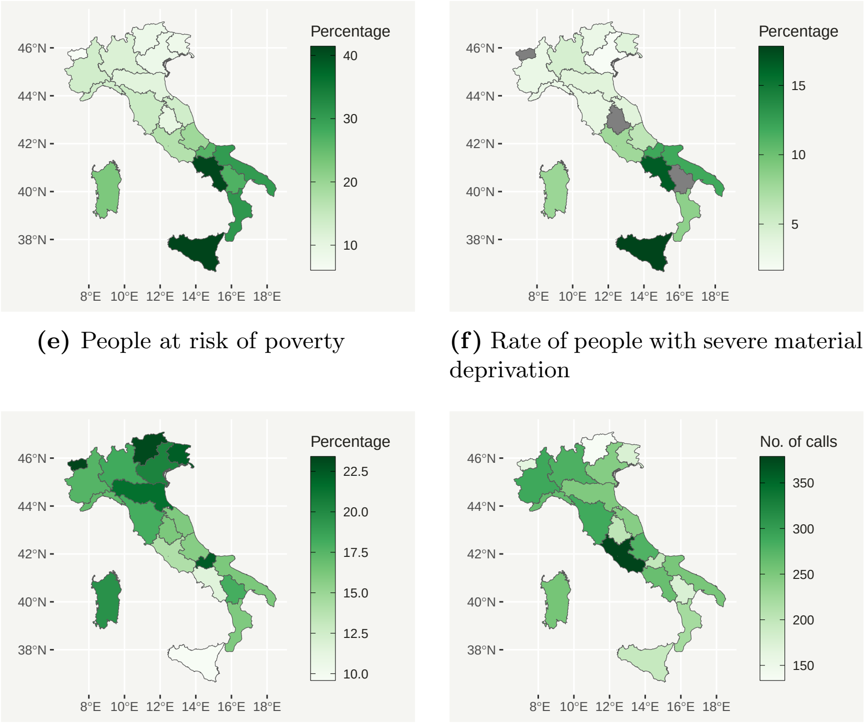
(a) People with at least upper secondary education level (25-64 years old)

(b) Per capita adjusted disposable income

(c) Early leavers from education and training

(d) Employment rate (20-64 years old)

Figure 5.3: The model covariates plotted on a map of Italy. (Figure continues on next page)



(e) People at risk of poverty

(f) Rate of people with severe material deprivation

(g) Rate of people showing risk behaviour connected to alcohol consumption

(h) Helpline calls per 100000 women.

Figure 5.3: The model covariates plotted on a map of Italy. (Figure continued from previous page)

In Figure 5.4, the correlation between all covariates and the response variable is shown. The response variable, as well as the helpline calls adjusted for the number of women in a region is also included in the correlation matrix. We clearly see that there is correlation between the seven covariates used to model the number of gender related crimes. This makes sense, as these are all indicators connected to work and life situation. None of the covariates are fully correlated however, so we choose to include all in the model.

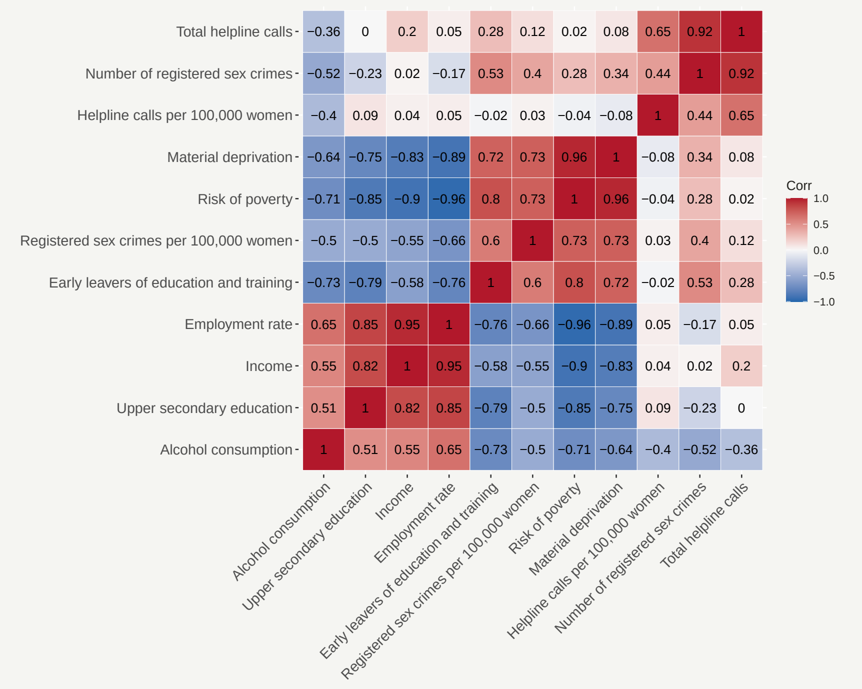


Figure 5.4: Correlation Matrix for observed counts of violence against women z and all model covariates.

5.2 The Model

We assume that the rate of violence against women in Italy is under-reported, as evidence from FRA - European Union Agency for Fundamental Rights 2015 and Department for Equal Opportunities 2014 suggests. As the simulation study performed in Chapter 4 suggests, using a naive Poisson distribution on the observed count of violence against women would then lead to severely biased estimates. The Poisson-Logistic model performed well on severely under-reported count data, even on the sparse spatial structure of Italy, using the disconnected graph with all 20 regions. The model is robust, and manages to recover the true parameters even when we apply it to noisy data. From these results, we concluded that the Poisson-Logistic model presented in Equations 2.13-2.15 with an informative prior distribution can be used for this application. We investigate the model performance using the covariates presented in Section 5.1, with $x_{1,i}$ to $x_{7,i}$ as process covariates and u_i as the under-reporting covariate. Looking at Figures 5.2 and 5.2, we did not see any strong indications to model the covariates as anything other than linear effects in the model. This also keeps the model simpler, and less complex. The resulting model can then be written as

$$
\begin{aligned}
z_i &\sim \text{Poisson}(E_i \pi_i \lambda_i), && (5.1)\\
\log\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \beta_1 u_i, && (5.2)\\
\log(\lambda_i) &= \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} \\
&+ \alpha_4 x_{4,i} + \alpha_5 x_{5,i} + \alpha_6 x_{6,i} && (5.3)\\
&+ \alpha_7 x_{7,i} + \phi_i,
\end{aligned}
$$

where ϕ_i is an ICAR random effect, defined as in the simulation study in Chapter 4. z_i is the observed count in region i, E_i is the population of women in region i, included into the model as a population offset. π_i is the reporting probability, λ_i is the expected count, and $\alpha_0, \ldots, \alpha_7$, β_0 and β_1 are unknown parameters. We also choose to use the disconnected spatial structure of Italy with 20 regions correcting for the effects of the disconnected graph, as this gave satisfactory results in the simulation study.

As in the simulation study, we need an informative prior distribution for the model to be identifiable. We use expert knowledge on the reporting rate and use this to elicit an informative prior on β_0. We do this by using the transformation described in Equations 2.23-2.24 and the normal approximation described in Section 4.1.2. We set the mean reporting rate at the mean of the centred covariates, p_0, as 10%, and the 99.99th percentile as 30%. This reporting rate is chosen as Department for Equal Opportunities 2014 estimates a 12.2% reporting rate for intimate partner violence against women, and a 6% reporting rate for non-partner violence against women in Italy. There are no joint estimates for the reporting rate, and we therefore choose a reporting rate that is between the two estimates. The 10% reporting rate elicits a $\mathrm{N}(-2.130, 0.403)$ prior distribution for β_0. For the other unknown parameters $\alpha_0, ..., \alpha_7$ and β_1 we define a non-informative $\mathrm{N}(0, 10^2)$ prior distribution. The spatial effect ϕ is given a $\mathrm{Gamma}(1, 0.0005)$ prior. Like in the simulation study, `inlabru` is used to conduct inference on the model.

5.3 Results

To assess the model performance and how the suggested covariates influence the models, we first run the model excluding all the process covariates. We denote this model as M0. Then, we new add covariates to the model using the method of forward selection. This means that we include the covariate that improves model performance the most into the model, as this covariate is the most influential. We label these models M1, . . . , M7. Note that the under-reporting covariate u_i is included for all models. We compare diagnostics to assess the usefulness of the introduced covariate.

5.3.1 Diagnostic tools to access model performance

In order to compare the performance of the different models and perform the model selection, we calculate and compare information criterion values for the models. We compare both the Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002, Linde 2005) and the Widely Available Information Criterion (WAIC) (Watanabe 2010). These information criterion are both easily available in `inlabru`. Both DIC and WAIC are rooted in the Akaike Information Criterion (AIC) (Akaike 1973). AIC is an estimator of prediction error for statistical models given a data set. We assume that no model is a perfect representation of the process it is describing, and some information will almost always be lost. The different information criterion is then a measure of how much information is lost through the modelling. This measure is relative, meaning that the value has no meaning on its own, but can be used to compare the relative performance of several models, judging which model gives the best fit.

The Deviance Information Criterion

The Deviance Information Criterion is a generalisation of the Akaike Information Criterion, and is used for hierarchical models. DIC is a popular choice for Bayesian modelling. If we define the deviance as

$$D(\theta) = -2log(p(y|\theta)) + C, \tag{5.4}$$

where y is the data, and θ are the unknown parameters. $p(y|\theta)$ is the model likelihood. C is a constant that cancels out when comparing models. The Deviance Information Criterion is defined as

$$\mathrm{DIC} = D(\bar{\theta}) + 2p_D, \tag{5.5}$$

where $\bar{\theta}$ is the expectation of θ, and p_D is the effective number of parameters. The effective number of parameters is calculated as $p_D = \overline{D(\theta)} - D(\bar{\theta})$ (Spiegelhalter et al. 2002). Because a larger effective number of parameters makes it easier for the model to fit the data, we need this penalisation in the calculation of the DIC.

The Deviance Information Criterion does not come without difficulties however. The proposed measure is attempting to obtain an unbiased and accurate measure of the prediction error that is valid for a broad class of models (Gelman et al. 2014). This is very difficult and Gelman et al. 2014 shows that the DIC struggles when the posterior distribution is not well summarised by its mean. Because of this, we also look at a different information criterion, namely the Widely Available Information Criterion (WAIC).

The Widely Available Information Criterion

The Widely Available Information Criterion (more commonly referred to as the Watanabe-Akaike Information Criterion) is a fully Bayesian method. This method estimate the out-of-sample expectation, and to do this it starts with the log pointwise posterior predictive density. It then, like the DIC, corrects for the effective number of model parameters in the model by adding a penalisation. This is done to adjust for overfitting. The WAIC can be written as

$$\text{WAIC} = -2\sum_{i=1}^{n} \log \int p(y_i|\theta)p_{post}(\theta)d\theta + p_W, \tag{5.6}$$

where p_W is the effective number of parameters, y_i is the data, θ the model parameters and $p_{post}(\theta)$ is the posterior distribution. We use $p_W = \sum_{i=1}^{n} \text{variance}(\log(p(y_i|\theta)))$, as suggested by Gelman et al. 2014. As the WAIC averages over the posterior distribution, unlike DIC that conditions on a point estimate, it is preferred by many over DIC. Gelman et al. 2014 does however show that the WAIC can struggle with structured models, like the spatial models we are looking at in this application.

Because neither DIC or WAIC is without limitations, we decide to use them both as diagnostic tools. We can then compare the models using both tools, and look at similarities and discrepancies.

5.3.2 Comparing diagnostic results of different models

Modelling with no spatial effect

We first perform modelling without the spatial effect ϕ. The results from this is shown in Table 5.3. Here, we see that the DIC and WAIC values are high for M0, and as covariates are added to the model, the results of the diagnostic tests become better. The WAIC keeps decreasing until M3, whereas the DIC is lowest for M5. This leads us to conclude that M3 and M5 are the models returning the best results when no spatial effects are included.

Model	DIC	Effective number of parameters DIC	WAIC	Effective number of parameters WAIC	Included covariate
M0	642.98	2.0063	635.29	60.854	–
M1	371.17	3.0277	398.96	26.778	$x_{5,i}$
M2	342.40	4.0350	370.35	27.198	$x_{1,i}$
M3	327.33	5.0422	360.74	31.858	$x_{4,i}$
M4	321.31	6.0489	369.78	43.262	$x_{7,i}$
M5	310.06	7.0561	366.01	49.071	$x_{3,i}$
M6	310.67	8.0632	368.62	51.146	$x_{2,i}$
M7	310.54	9.0703	373.86	55.223	$x_{6,i}$

Table 5.3: Results from models with no spatial effects

All three covariates included in M3 return significant posterior estimates, with all three estimates being positive. This is somewhat surprising for covariates $x_{1,i}$ (Education level) and $x_{4,i}$ (Employment rate). $x_{5,i}$ (People at risk of poverty) is the most significant covariate in the model, providing a stronger positive effect than the two other model covariates. The under-reporting covariate is also significant in M3, with a weak negative effect. For M5, the results are similar. M5 uses the same covariates for modelling as M3, but also includes $x_{7,i}$ (Alcohol consumption) and $x_{3,i}$ (Early leavers from education and training). Again, all process covariates return significant, positive posterior estimates with $x_{5,i}$ giving the largest effect. For M5, the posterior estimates of the under-reporting covariate u_i is not significant.

Modelling with a structured spatial effect

Next, we include a structured spatial effect ϕ into the model, with the intent that this spatial effect will pick up on any spatial dependencies in the data, thus improving the model performance. Again, we use forward selection to include the covariates into the model one after the other. The results from all the models with a spatial effect is shown in Table 5.4.

When including a spatial effect ϕ into the model, both the DIC and the

Model	DIC	Effective number of parameters DIC	WAIC	Effective number of parameters WAIC	Included covariate
M0$_\phi$	205.50	18.674	202.14	10.983	–
M1$_\phi$	205.49	18.831	201.77	10.843	$x_{1,i}$
M2$_\phi$	205.36	18.861	201.41	10.708	$x_{4,i}$
M3$_\phi$	205.40	19.040	201.08	10.580	$x_{7,i}$
M4$_\phi$	205.10	18.960	200.48	10.329	$x_{3,i}$
M5$_\phi$	205.30	19.130	200.58	10.373	$x_{2,i}$
M6$_\phi$	205.82	19.269	201.36	10.658	$x_{6,i}$
M7$_\phi$	206.24	19.252	202.39	11.058	$x_{5,i}$

Table 5.4: Results from models with a structured random effect ϕ.

WAIC give very different results than when the spatial effect is not included. M3 and M5 were the non-spatial models returning the best results, but all models with a spatial random effect ϕ perform better than this. The model performance is also very similar for all the models, regardless of the number of process covariates that are included in the model. This suggests that the model covariates does not help us predict the incidence rate of violence against women in Italy when a spatial effect is added.

Looking closer at the results from each model, we also notice that only a few of the model covariates are significant in any of the models when including a spatial effect. The under-reporting covariate is not significant in any of the models. Possible reasons for this will be discussed in detail in Chapter 6. The none of the process covariates are significant in M1 and M2, but $x_{4,i}$ returns significant posterior estimates for M3$_\phi$,. . .,M7$_\phi$, with $x_{4,i}$ having a negative effect. $x_{1,i}$ also returns a significant positive effect for M4$_\phi$.

5.3.3 Posterior estimates of model M4$_\phi$

To understand the modelling results better we look closer at M4$_\phi$, the model returning the lowest DIC and WAIC. This model includes four process covariates, with $x_{1,i}$ (Upper Secondary Education), $x_{4,i}$ (Employment rate), $x_{7,i}$ (Alcohol consumption) and $x_{3,i}$ (Early leavers from education and training). We look at the posterior estimates of the spatial effect, as well as posterior estimates for the expected count λ_i and the reporting rate π_i.

For the disconnected graph of Italy, we included a region-specific intercept for the ICAR model, as described in Section 4.3.1. The posterior spatial effect then becomes $\alpha_{cc} \cdot \phi$. This effect is seen in Figure 5.5. We see that there is a no prominent spatial trend, and the structured spatial effect behaves more like an unstructured random effect. This is surprising, as there were strong spatial trends in the covariates. It does seem to indicate that there is no spatial indications that lead to violence against women in Italy. As the model performance is significantly better when a spatial effect is included, this spatial effect does pick up variability in the model that are not explained by the covariates, but this variability does not seem to have a strong spatial trend.

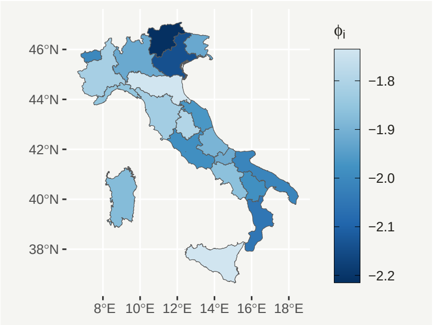
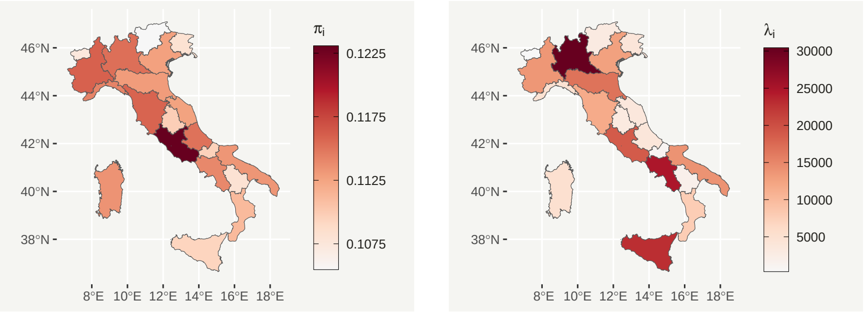


Figure 5.5: The posterior estimate for the structured spatial effect ϕ of model M4$_\phi$, plotted on the map of Italy



(a) Posterior estimates of the reporting probability π_i.

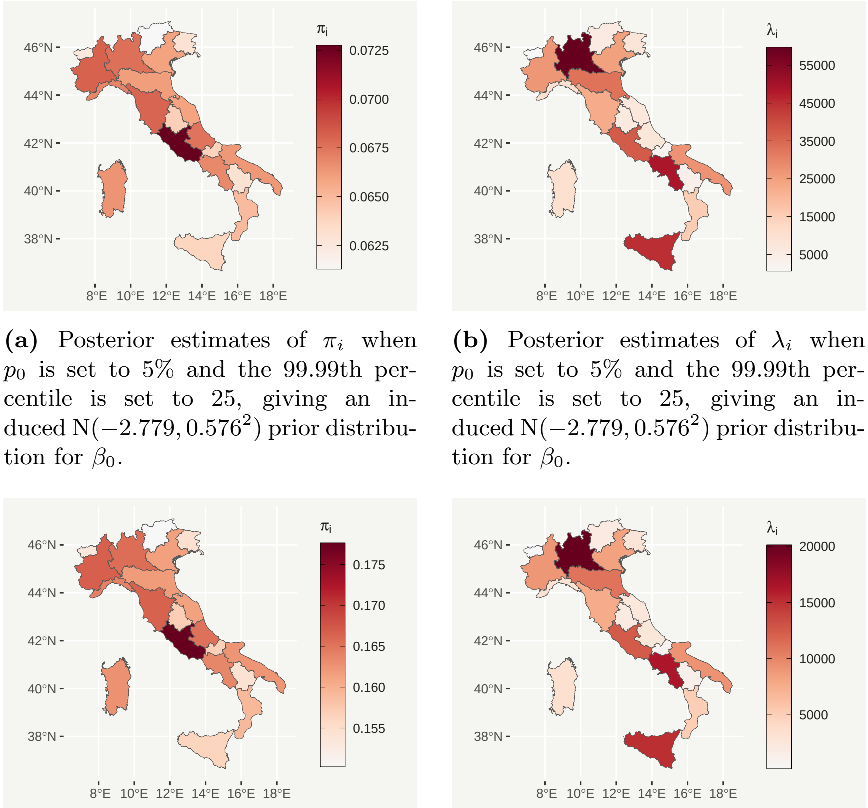
(b) Posterior estimates of the expected count λ_i.

Figure 5.6: Posterior estimates from M4$_\phi$

Looking at the posterior estimates of λ_i, seen in Figure 5.6b, we again see that the expected count λ_i looks similar to the observed count z_i, seen in Figure 5.1a. The values are not the same, but the spatial variation in the posterior estimates are very similar to the spatial variation in the observed count. If we look at the posterior estimates for the reporting rate π_i, shown in Figure 5.6a, we understand why that is. The reporting rate is almost constant between the regions of Italy, and almost identical to the prior of 10% reporting rate. There are small regional differences with a similar spatial structure to that of u_i, but the effect of this is small. This suggest that the posterior estimates of π_i is very influenced by the prior distribution on β_0, and not much by the under-reporting covariate u_i.

The consequences of the informative prior distribution on β_0

Seeing as the model seems more dependent on the observations z_i and the prior distribution of β_0 that the data $\mathbf{x}$ and u, we are interested in looking at how the model changes with a different prior distribution on β_0. The simulation study showed that the model was robust even with noisy covariates, and not sensitive to small changes in the prior distribution on β_0. As the posterior estimates of π_i are so similar to the expert suggested mean reporting rate p_0 of 10%, it is interesting to see how this changes if we change p_0. Figure 5.7 shows the posterior estimates for π_i if we set p_0 to 5% and 15% and induce a prior on β_0 from this.



(a) Posterior estimates of π_i when p_0 is set to 5% and the 99.99th percentile is set to 25, giving an induced $\mathrm{N}(-2.779, 0.576^2)$ prior distribution for β_0.

(b) Posterior estimates of λ_i when p_0 is set to 5% and the 99.99th percentile is set to 25, giving an induced $\mathrm{N}(-2.779, 0.576^2)$ prior distribution for β_0.

(c) Posterior estimates of π_i when p_0 is set to 15% and the 99.99th percentile is set to 35, giving an induced $\mathrm{N}(-1.697, 0.324^2)$ prior distribution for β_0.

(d) Posterior estimates of λ_i when p_0 is set to 15% and the 99.99th percentile is set to 35, giving an induced $\mathrm{N}(-1.697, 0.324^2)$ prior distribution for β_0.

Figure 5.7: Posterior estimates for π_i and λ_i from models with different prior distributions on β_0.

We see from the Figure 5.7 that the posterior estimates of π_i and λ_i is almost completely governed by the prior specification. That indicates that the model is not robust at all. In both Figures 5.7a and 5.7c and Figure 5.6a, the posterior estimates are higher than the informative prior specifies, indicating that some information is imparted through the under-reporting covariate u_i, although not much. From the simulation study we saw that the model was robust even when a noisy under-reporting covariate was introduced into the model, suggesting that this under-reporting covariate u_i provides very little information about the true underlying process that we are modelling.

5.3.4 Conclusions on modelling the rate of violence against women in Italy

If we look at the results all together, they are not convincing. We see that the diagnostic testing returns the best estimates for the models with a structured spatial effect ϕ. When looking closer at the posterior estimates from the best of these models however, we see that it is almost fully governed by the prior specification and not the data. This suggests that none of the covariates we have looked at, neither the process covariates $\mathbf{x}$ nor the under-reporting covariate u are useful for modelling the rate of violence against women in Italy. Possible reasons for this will follow in Chapter 6.

Chapter 6

Discussion and further work

In this thesis, we have applied the Poisson-Logistic model to the application of violence against women in Italy, performing inference on the model using `inlabru`. We have conducted a comprehensive simulation study, investigating how the model performs when the reporting rate is low, and when the spatial structure of interest is weak with a low number of regions. We have also investigated how adding noise to the under-reporting part of the model affected the model performance, as well as how robust the model is to changes in the informative prior on β_0. From the simulation study, we concluded that the model was robust, and appropriate to use for modelling the rate of violence against women in Italy. There are however several results from the simulation study and the application to modelling the rate of violence against women in Italy that needs to be discussed further.

6.1 Simulation study and model definition

The simulation study returned good results for most of the $m = 100$ runs, but we saw that `inlabru` struggled with identifying the two different parts of the model for some of the runs, leading to the posterior distributions for the model intercepts α_0 and β_0 being quite flat, and for the posterior mean of α_0 to be far to high. This was very noticeable in the posterior estimates for λ for those particular instances, as there is an exponential relationship between α_0 and λ, making small inaccuracies in α_0 very noticeable in λ. As this problem only occurred for a small number of runs of the simulation study, we did not deem it to be a significant problem, and concluded that the model was still appropriate to use on applications with severe under-reporting.

Although we still found the model good enough to apply to a real-data application, this problem of identifiability is something that needs to be investigated further. One possible solution to this problem is to apply an informative prior also to the process-part of the model, for instance on the process intercept α_0. This can be done in a similar way as for β_0, using a Beta prior distribution derived from expert knowledge on the expected count λ, and then induce an informative prior distribution for α_0 based on this, using a normal approximation in order to implement it in `inlabru`. This is left for future work.

6.2 Application on incidence rate of violence against women in Italy

In Chapter 5, we applied the Poisson-Logistic model to the issue of violence against women in Italy using `inlabru`. The results from this modelling suggests that the posterior estimates from the model is more governed by the informative prior distribution on β_0 and the response variable z than the data included in the model.

The Poisson-Logistic model has been used with success in many applications, as discussed in Section 2.2. Chen et al. 2022 used the Poisson-Logistic model with spatial effects in an application with severely under-reported data, namely the detection of Covid-19 in the early days of the pandemic when access to reliable testing was scarce. These earlier applications of the Poisson-Logistic model, as well as the simulation study performed in Chapter 4 suggest that the model i sensible in the application of severely under-reported count data.

From this, we conclude that the poor modelling results might have other reasons behind it. Two possible reasons might be the weak spatial structure of Italy, and the quality of the data used for modelling. Italy is a long and thin country, with many regions having only 0, 1 or 2 neighbouring regions. A stronger spatial structure provides the regions with the ability of "borrowing" strength from its neighbours. Chen et al. 2022 also looked at severely under-reported data, but in the context of USA. The graph of USA has a stronger spatial structure than Italy, which might make the model more robust. Additionally, Italy consist of only 20 regions. This low number of regions might influence the model robustness, as it means each covariate has relatively few data points, making inference more difficult. Looking back at the simulation study in Chapter 4, we saw that the Poisson-Logistic model managed to recover the true parameters, even when using the disconnected graph of Italy with 20 regions. This in turn suggests that the weak spatial structure of Italy and low number of regions should be good enough to perform inference, as long at the data provided to the model is good enough. Even when introducing noise into the under-reporting covariate, the model performance was good. This again suggest that it is not only the weak spatial structure that causes the poor modelling results.

This leads us to the model covariates. The covariates were chosen after advice from Arima 2022 and Polettini 2022. The choice of several of the process covariates were also grounded in literature on violence against women, with risk factors as alcohol use and education level discussed in WHO 2005 and Vugt et al. 2022. As models including a spatial effect ϕ returned the best results, these are the models we looked closer at. The results revealed that neither the process covariates $\mathbf{x}$ nor the under-reporting covariate u were significant for any of the models including a spatial effect. This signifies that the chosen covariates are not useful for modelling violence against women. There can be several reasons for this. Firstly, this can mean that none of the chosen process covariates has any correlation with the incidence rate of violence against women. There is not a lot of research into the risk factors of violence against women. Some of the research done is discussed in Chapter 1, and included alcohol use and lacking education as possible risk factors. Data on this is included here, but were not significant in any of the models. This may suggest that more research into these risk factors are needed in order to make a informed decision about which covariates to include in the model.

The covariate u used to model the reporting probability was also not significant. This was the registered calls to a national hotline service for victims of violence. The hope was that this hotline was widely available to people, and easier to contact than to go to the police. This helpline would then pick up on the regional differences in how aware the population in a region was of the problem of violence against women, and that this could be used as a proxy for the regional reporting rate. From the modelling results, this covariate did not give much information about the regional under-reporting rate. More investigation into what factors stop victims of violence from going to the police is needed in order to find data that better describes this. WHO 2005 found that increases in education level amongst women decreases the risk of them being a victim of violence. It could be interesting to see if this also increases the awareness of available reporting resources, and therefore would be an appropriate covariate for the under-reporting part of the model.

One possible reason that the data included in the model was non-significant is the aggregation level. We looked at the 20 regions of Italy, but as Italy is a large country with a population of almost 60 million people, this aggregation is very general. Many of the regions in Italy are large and heterogeneous, consisting of both larger cities and more rural counties. All model covariates as well as the response variable z are aggregated up to this regional level. It is likely that information is lost in this aggregation, as all the data is averaged over these large and diverse regions. Looking at the covariates shown in Figures 5.3 and 5.3, we do see a spatial trend in the data between the northern and southern regions of Italy, but it is possible that this regional-level data becomes too general in order to pick up on smaller regional differences. The simulation study showed that the

Poisson-Logistic model is robust even with noisy data, but it is likely that the regional data does not provide an accurate enough picture of Italy for the model to use the data effectively. To improve on this, less aggregated data is needed. Italy has three administrative levels, with the regional level being the highest level. The second level is the provincial level. Each region is divided into 1 to 12 provinces, and there are 107 provinces in total. A further investigation after this work could be to perform inference using the Poisson-Logistic model on the same data on a provincial level, and see if this returned more informative posterior estimates. It is possible that the data aggregated up to the provincial level more accurately captures the regional differences between the provinces than data on the regional level. Other added benefits to modelling with data on the provincial level is the strengthening of the spatial structure, as well as the larger number of data points. The reasons why data for the provincial level has not been used in this investigation are data availability and privacy concerns. Because we are dealing with sensitive data of reported crimes and personal well-being, the data needs to be aggregated in such a way that no personal information is revealed. Now, only data on the regional level is available to the public through ISTAT. Whether provincial data would be possible to obtain while still taking privacy concerns into account will need to be further investigated.

There is another data quality concern that also needs to be addressed in this investigation. The report on equitable and sustainable well-being (BES) (Istat 2022b), where all the process covariates are gathered from, heavily relies on survey data. It is very difficult to obtain unbiased estimates that accurately represent the whole population using surveys. Both the quality of the survey questions and how people are selected for survey participation needs to be considered. There is also a number of people that chooses not to answer a population survey. Knowledge of whether this is a random subset of the population or if certain groups of people are less likely to answer is important in order to possibly correct for this effect and give estimates that reflect the whole population. Lastly, some topics are more difficult to speak publicly about, something which might result in questions about these topics not being answered truthfully. Because of these factors, the quality of survey data varies. The BES report is a yearly report, and gathers data from specialised and reputable surveys. The data quality of the process covariates should therefore be good, but it is worth considering. The under-reporting covariate u_i and the response variable z_i are gathered from registry data from reliable sources, so the same quality concerns does not apply to them.

As mentioned in Section 6.1, it is possible to introduce another informative prior distribution into the model. `inlabru` had no convergence problems when applied to the issue of violence against women in Italy, but we did see how the posterior estimates of the reporting rate π_i were almost completely dependent on the prior distribution of β_0. Introducing an informative prior on the process intercept α_0, which can be interpreted as the mean rate of violence against women, could allow us to weaken the prior on β_0. This may in turn allow the posterior estimates of π_i to be more governed by the data and less by the prior distribution.

Although the model application to real data did not return good results, the simulation study still proves the effectiveness of using the `inlabru` extension to the INLA methodology on the Poisson-Logistic model. This was not possible before the development of `inlabru`, due to the non-linear model predictor of the model. Chen et al. 2022 used `Stan` when conducting inference on the Poisson-Logistic model with an informative prior on β_0 as an alternative to `NIMBLE` used by Stoner et al. 2019, and showing how `Stan` was a faster method giving the same results. With this simulation study using `inlabru`, we have seen how the INLA methodology can be used for inference on the model, providing a much faster method of inference than `Stan`.

To ensure that the Poisson-Logistic model was identifiable, we used a similar approach as in Stoner et al. 2019, by inducing an informative prior distribution on β_0. We did this through expert knowledge on the reporting rate π, as seen in Chen et al. 2022. We did however have to adapt this approach for use with `inlabru`, as INLA assumes Gaussian priors for all fixed parameters, and this was done by numerically approximating a normal distribution to the induced prior for β_0. This approach worked well as seen in Figure 4.1, and as the results from the simulation study showed. This allowed us to use a the information from a Beta prior distribution derived based on expert knowledge in the context of Bayesian inference using INLA.

From the simulation study, we have seen that it is possible to use `inlabru` to conduct inference on the Poisson-Logistic model where the reporting rate is low. Applying the model to the incidence rate of violence against women in Italy resulted in a poor modelling fit. This is likely the result of highly aggregated data, what is summarised over large, heterogeneous regions in Italy. For further work, we recommend that data aggregated up to the provincial level be used, in order to conduct meaningful inference. More investigation into the risk factors of violence against women also needs to be performed, so a more informed choice can be made on the appropriate model covariates. Lastly, it would be interesting to look at different possible covariates to model the under-reporting of violence, other than calls to the helpline service 1522. One such possible covariate could be the female education level. It is estimated that 31.5% of all Italian women will experience violence from either an intimate partner or someone else during their lifetime. As this causes large societal consequences, it is an important field of research that needs to be further developed.

Bibliography

1522 (2022). *1522 - Numero Anti Violenza e Stalking*. URL: {https://www.1522.eu/cose-1522/?lang=en} (visited on May 28, 2022).

Abramo, G, C. A. D'Angelo, and F Rosati (2016). "The north-south divide in the Italian higher education system". *Scientometrics* 109.3, pp. 1588–2861. DOI: 10.1007/s11192-016-2141-9.

Abramsky, T et al. (2011). "What factors are associated with recent intimate partner violence? Findings from the WHO multi-country study on women's health and domestic violence". *BMC Public Health* 11, p. 17. DOI: 10.1186/1471-2458-11-109.

Akaike, Hirotugu (1973). "Information Theory and an Extension of the Maximum Likelihood Principle".

Amoros, E, Martin J.L., and Laumon B (2006). "Under-reporting of road crash casualties in France". *Accident analysis and prevention* 38.4, pp. 627–635. DOI: 10.1016/j.aap.2005.11.006.

Arima, Serena (2022). Personal communication. Department of History, Society and Human Studies. University of Salento.

Bachl, F. E. et al. (2019). "inlabru: an R package for Bayesian spatial modelling from ecological survey data". *Methods in Ecology and Evolution* 10.6, pp. 760–766. DOI: https://doi.org/10.1111/2041-210X.13168.

Besag, J, J York, and A Mollie (1991). "Bayesian image restoration, with teo applications in spatial statistics." *Annuals of the Institute of Statistical Mathematics* 43, pp. 1–20. DOI: https://doi.org/10.1007/BF00116466.

Caudill, BS and FG Mixon Jr (1995). "Modelling household fertility decisions: estimation and testing of censored regression models for count data". *Empirical Economics* 20, pp. 183–196. DOI: https://doi.org/10.1007/BF01205434.

Chen, J., J.J. Song, and J.D. Stamey (2022). "A Bayesian Hierarchical Spatial Model to Correct for Misreporting in Count Data: Application to State-Level COVID-19 Data in the United States". *International Journal of Environmental Research and Public Health* 19.3327, p. 15. DOI: https://doi.org/10.3390/ijerph19063327.

Department for Equal Opportunities (2014). *Violence against women in and outside the family*. URL: {https://www.istat.it/it/files//2019/11/Violence-against-women-_2014.pdf} (visited on June 20, 2022).

Dvorzak, M and Wagner H (2016). "Sparse Bayesian modelling of underreported count data". 16.1, pp. 24–46. DOI: https://doi.org/10.1177/1471082X15588398.

Eurostat (2021). *Glossary: Equivalised income*. URL: {https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Equivalised_income} (visited on Mar. 27, 2022).

FRA - European Union Agency for Fundamental Rights (2015). *Violence against women: an EU-wide survey: main results*. Publications Office of the European Union, xix, 87 p. DOI: doi/10.2811/981927.

Gelman, A, J Hwang, and A Vehtari (2014). "Understanding predictive informaton criteria for Bayesian models". *Statistics and Computing* 24, pp. 997–1016. DOI: 10.1007/s11222-013-9416-2.

Hoffman, M. D. and A Gelman (2014). "The Nu-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". *Journal of Machine Learning Research* 15.47, pp. 1593–1623. URL: http://jmlr.org/papers/v15/hoffman14a.html.

Istat (2021). *REGIONAL ACCOUNTS, Weak growth in all areas of the country*. URL: {https://www.istat.it/en/archivio/252224} (visited on June 25, 2022).

— (2022a). *Istituto Nazionale di Statistica*. URL: {https://www.istat.it/en/organisation-and-activity} (visited on Mar. 27, 2022).

— (2022b). *report on equitable and sustainable well being*. URL: {https://www.istat.it/en/well-being-and-sustainability/the-measurement-of-well-being/bes-report} (visited on Apr. 4, 2022).

Linde, A. van der (2005). “DIC in variable selection”. *Statistica Neerlandica* 59.1, pp. 45–56. DOI: https://doi.org/10.1111/j.1467-9574.2005.00278.x.

Lindgren, F and F. E. Bachl (2021). *Iterative INLA method.* Accessed: 18.01.2022. URL: https://inlabru-org.github.io/inlabru/articles/method.html.

Martino, S and A Riebler (2019). “Integrated Nested Laplace Approximations (INLA)”. DOI: https://arxiv.org/abs/1907.01248.

Metropolis, N et al. (1953). “Equation of State Calculations by Fast Computing Machines”. 21.6, pp. 1087–1092. DOI: https://doi.org/10.1063/1.1699114.

Moreno, E and J Giron (1998). “Estimating with incomplete count data: A Bayesian approach”. 66, pp. 147–159.

Oliveira, GL, RH Loschi, and RM Assunção (2017). “A random-censoring model for underreported data”. *Statisitcs in Medicine* 36.12, pp. 4873–4892. DOI: 10.1002/sim.7456.

Oliveira, GL et al. (2021). “Bias Correction in Clustered Underreported Data”. *Bayesian Anal. Advance Publication*, pp. 1–32. DOI: https://doi.org/10.1214/20-BA1244.

Polettini, Silvia (2022). Personal communication. Department of Methods and Models for Economics, Territory and Finance. Sapenzia University of Rome.

Ravenzwaaij, D. v., P Cassey, and Brown S. D. (2018). “A simple introduction to Markov Chain Monte-Carlo sampling”. 25, pp. 143–154. DOI: 10.3758/s13423-016-1015-8.

Rue, H and L Held (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall - CRC Press. DOI: https://doi.org/10.1201/9780203492024.

Rue, H, S Martino, and N Chopin (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2, pp. 319–392. DOI: https://doi.org/10.1111/j.1467-9868.2008.00700.x.

Sanz-Barbero, B, L Otero-Garcia, and C Vives-Casses (2018). "Factors Associated With Women's Reporting of Intimate Partner Violence in Spain". *Journal of Interpersonal Violence* 33.15, pp. 2402–2419. DOI: 10.1177/0886260515625512.

Schmertmann, C. P and M. R Gonzaga (2018). "Bayesian Estimation of Age-Specific Mortality and Life Expectancy for Small Areas With Defective Vital Records". 55, pp. 1363–1388. DOI: https://doi.org/10.1007/s13524-018-0695-2.

Spiegelhalter, D. J. et al. (2002). "Bayesian measures of model complexity and fit". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.4, pp. 583–639. DOI: https://doi.org/10.1111/1467-9868.00353.

Sterrantino, Anna, Massimo Ventrucci, and Håvard Rue (May 2017). "A note on intrinsic Conditional Autoregressive models for disconnected graphs". *Spatial and Spatio-temporal Epidemiology* 26. DOI: 10.1016/j.sste.2018.04.002.

Stoner, O and T Economou (2018). "Correcting Under-Reporting in Historical Volcano Data". *Proceedings of the 33rd International Workshop on Statistical Modelling* 1, pp. 1482–1483.

Stoner, O, T Economou, and GD Silva (2019). "A Hierarcical Framework for Correcting Under-Reporting in Count Data". *Journal of the Americal Statistical Associasion* 114.528, pp. 1481–1492. DOI: https://doi.org/10.1080/01621459.2019.1573732.

Stöckl, H et al. (2013). “The global prevalence of intimate partner homicide: a systematic review”. *Lancet* 382, pp. 859–865. DOI: http://dx.doi.org/10.1016/S0140-6736(13)61030-2.

Sørbye, S.H. and H Rue (2014). “Scaling intrinsic Gaussian Markov random field priors in spatial modelling”. *Spatial Statistics* 8. Spatial Statistics Miami, pp. 39–51. ISSN: 2211-6753. DOI: https://doi.org/10.1016/j.spasta.2013.06.004.

Terza, JV (1985). “A Tobit-type estimator for the censored Poisson regression model”. *Economics Letters* 18.4, pp. 361–365. DOI: https://doi.org/10.1016/0165-1765(85)90053-9.

The BUGS Project (1989). Accessed: 28.01.2022. URL: https://www.mrc-bsu.cam.ac.uk/software/bugs/.

The World Bank (2022). *World Bank Country and Lending Groups.* URL: {https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups} (visited on May 10, 2022).

United Nations (1993). *General Assembly Resolution 48/104, Declaration on the Elimination of Violence against Women.* URL: {undocs.org/en/A/RES/48/104} (visited on May 9, 2022).

Violenceagainstwomen.Stat (2022a). *Alleged offenders and victims of crimes reported by the police forces to the judicial authority: Violent crimes, age - regions.* Data retrieved from ViolenceAgainstWomen.Stat database, filtering on ”Investigation, prosecutions, convictions” and then ”Police reporting”. URL: {http://dati-violenzadonne.istat.it/?lang=en} (visited on May 30, 2022).

— (2022b). *Users to 1522 (anti-violence and stailking number).* Data retrieved from ViolenceAgainstWomen.Stat database, filtering on ”Special victim services”. URL: {http://dati-violenzadonne.istat.it/?lang=en} (visited on May 30, 2022).

— (2022c). *Women who have suffered violence - demographic characterisics and habits.* Data retrieved from ViolenceAgainstWomen.Stat database, filtering on ”Data on the phenomenon”, then ”Violence inside and outside the family” and finally ”Type of perpetrator and region”. URL: {http://dati-violenzadonne.istat.it/?lang=en} (visited on May 30, 2022).

Vugt, L van and I.A. Pop (2022). “Status mismatch and self-reported intimate partner violence in the European Union: does the country’s context matter?” *European Societies.* DOI: 10.1080/14616696.2022.2068184.

Wakefield, Jon (June 2006). “Disease mapping and spatial regression with count data”. *Biostatistics* 8.2, pp. 158–183.

Watanabe, S. (2010). “Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory”. *Journal of Machine Learning Reasearch* 11.116, pp. 3571–3594. URL: http://jmlr.org/papers/v11/watanabe10a.html.

Whittemore, A.S and G Gong (1991). “Poisson Regression with Misclassified Counts: Application to Cervical Cancer Mortality Rates”. 40.1, pp. 81–93. DOI: https://doi.org/10.2307/2347906.

WHO, World Health Organization (2005). *WHO multi-country study on women’s health and domestic violence against women : initial results on prevalence, health outcomes and women’s responses.*

— (2013). *Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence.* World Health Organization, vi, 51 p.

— (2021). *Violence against women prevalence estimates, 2018: global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence against women.* World Health Organization, xix, 87 p.

Winkelmann, R (1996). “Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism.” *Empir-*

ical Economics 21.4, pp. 575–587. DOI: https://doi.org/10.1007/BF01180702.

Winkelmann, R. and K.F. Zimmermann (1993). *Poisson logistic regression.* Münchener Wirtschaftswissenschaftliche Beiträge. Volkswirtschaftliche Fak., Ludwig-Maximilians-Univ. URL: https://books.google.no/books?id=EbxtNAEACAAJ.

Wøllo, Sara E. (2022). *Correcting Under-reporting in Count Data using INLA*. https://github.com/saraew/Prosjektoppgave.git.