Bayesian Hierarchical Approach in Latent Gaussian Modeling for Tuberculosis Cases in Jimma Zone: Using INLA Method

By:

Endale Alemayehu

A Thesis Submitted to Department of Statistics, College of Natural Sciences, Jimma University as a Partial Fulfillment for the Requirements of the Degree of Master of Science (MSc) in Biostatistics

November 2018

Jimma, Ethiopia

Bayesian Hierarchical Approach in Latent Gaussian Modeling for Tuberculosis Cases in Jimma Zone: Using INLA Method.

By:

Endale Alemayehu

Advisor: Akalu Banbata (PhD Scholar)

Co-advisor: Reta Habtamu (MSc)

November 2018

Jimma, Ethiopia

STATEMENT OF THE AUTHOR

As author of this research study, I declare that the thesis is a result of my genuine work, support of my supervisors and help hands of other individuals. Thus, all those had who participated in the study and sources of materials used for writing this thesis have been duly acknowledged. I have submitted this thesis to Jimma University as a partial fulfillment for the requirements of Degree of Master of Science in Biostatistics. The library directorate of Jimma University can deposit the copy of the thesis in the university library so that students and researchers can refer it. Moreover, I declare that I have not so far submitted this thesis to any other institution anywhere for that award of any academic degree, diploma or certificate and/or to get prove of society's problems. Any brief quotations from this thesis are allowed without requiring special permission if an accurate acknowledgement and citation (after publication) of the source is made. In all other instances, however, permission must be obtained from the author.

Endale Alemayehu

Date: _____

Signature: _____

November 2018

Jimma, Ethiopia

ACKNOWLEDGEMENTS

Above all, I would like to thank the almighty God and his mother St. Mary for the gift of health, wisdom and strength throughout any steps for the achievement of my thesis and more of my entire life.

My deep gratitude goes to my advisor Mr. Akalu Banbata (PhD scholar at Hasselt University) for his professional and deep comments, advice and guidance at any stage of this thesis. I appreciated his friendly contact and his punctuality for commented this paper. Simply, without his frequent supervision, this thesis may not reach at the stage it is now.

I sincerely appreciate my co-advisor Mr. Reta Habtamu (MS.c) for his valuable suggestions and comments for the successful realization of this thesis. It is a great pleasure to express his commitment to share his experience, knowledge and his smooth contacts is also appreciable.

I am also indebted to Dr. Havard Rue (Professor of Statistics and developer of R-INLA) for his surprising continual email contact during prior selection, in any R-INLA function and application of INLA. I appreciated all members of R-INLA group for their view on the questions I posted through the group's email.

My sincere gratitude goes to my family for their tenacious love and encouragement. I do not believe words can express my feelings to my mother who has been my devotee since my earliest childhood. Wishes long life to my mom.

Table of Contents

LIST OF TABLES

LIST OF GRAPHS

ACRONYMS

AFB	Acid-Fast Bacilli
CFR	Case Fatality Ratio
CPO	Conditional Predictive Ordinate
DOTS	Direct Observation Therapy Strategy
GLMM	Generalized Linear Mixed Model
GMRF	Gaussian Markov Random Field
HIV	Human Immunodeficiency Virus
HSTP	Health Sector Transformation Plan
INLA	Integrated Nested Laplace Approximation
KLD	Kullback-Leibler Divergence
LGM	Latent Gaussian Models
MCMC	Markov chain Monte Carlo
MDR-TB	Multidrug Resistant Tuberculosis
PIT	Probability Integral Transform
SDG	Sustainable Development Goal
SLA	Simple Laplace Approximation
TB	Tuberculosis
WAIC	Watanabe Akaike Information Criteria
WHO	World Health Organization

ABSTRACTS

Introduction: Tuberculosis is the long-lasting infectious disease caused by bacteria called Mycobacterium tuberculosis. Globally, in 2016 alone, approximately 10.4 million new cases have occurred worldwide. Africa has shared around 25% of the incidence and specifically in Ethiopia around 82 thousand was caught by Tuberculosis.

Objectives: This study has been aimed to model the counts of Tuberculosis cases using Bayesian hierarchical approach of Latent Gaussian Model (LGM) with Integrated Nested Laplace Approximation method. It is also designed to determine the predictors and see the variation of Tuberculosis incidences across districts of Jimma zone. Moreover, the researcher intends to compare the inbuilt R-INLA default priors and penalized complexity priors so that to assure the robustness of the priors for which Bayesian hierarchical approach of latent Gaussian model was applied.

Methods: The study has been conducted in Jimma zone of entire districts and the data is basically secondary which is obtained from Jimma zone health office. The counts of Tuberculosis cases have been analyzed with factors like gender, HIV co-infection, Population density and age of patients. The Integrated Nested Laplace Approximation (INLA) method of Bayesian approach which is fast, deterministic and promising alternative to MCMC method was used to determine posterior marginal.

Results: The latent Gaussian model of Poisson distributional assumption of Tuberculosis cases that includes both fixed and random effects with penalized complexity priors appeared to be the best model to fit the data based on the Watanabe Akaike Information Criteria and other supportive criteria. Using Kullback-Leibler Divergence criteria, the under-used simplified Laplace approximation indicated that posterior marginal was well approximated by normal distribution. The predictive value of the best model is not far deviated from the actual data based on the Conditional Predictive Ordinate and the probability integral transform.

Conclusions: The hierarchical level of Latent Gaussian Model with Penalized Complexity was found to be the appropriate model. All the variables were significant under this model and the posterior marginal was well approximated by standard Gaussian. The PIT indicated that predictive distribution was less affected by outliers and the model was reasonably well.

Key words: Tuberculosis, Bayesian, LGM, INLA

CHAPTER ONE

1. INTRODUCTION

1.1. Background of the Problems

Tuberculosis (TB) is a chronic infectious disease caused by a bacillus belonging to a group of bacteria grouped in the *Mycobacterium tuberculosis* complex and remains an important public health problem of the 21st century (WHO, 2017). It remains a high-priority communicable disease that causes an enormous burden of morbidity and mortality. Tuberculosis (TB) control and elimination rely on an early detection of active TB cases, prompt anti-TB treatment, identification of persons in risk of exposure and infection and prevention of secondary TB cases. (Lönnroth *et al*., 2015).

Globally, in 2016 alone, approximately 10.4 million new cases (range from 8.8 million to 12.2 million) which are equivalent to 140 cases per 100000 have occurred worldwide. According to the reports of WHO (2017), the most estimated number of TB cases are in the WHO South-East Asia Region (45%), the WHO African Region (25%) and the WHO Western Pacific Region (17%). Similarly, smaller proportions of cases occurred in the WHO Eastern Mediterranean Region (7%), the WHO European Region (3%) and the WHO Region of the Americas (3%) and 1.8 million deaths of tuberculosis were reported worldwide. It is also indicated that of all the cases, 11% of new cases and 0.4 million deaths were people with co-infected of human immunodeficiency virus (HIV) which makes the TB disease more serious top causes of mortality and morbidity (WHO, 2017; Asemahagn *et al*., 2018).

Likewise, there were worsened burden of TB with the estimated 600000 (range, 540 000–660 000) incident cases of Multi-Drug Resistant Tuberculosis (MDR-TB) with cases accounting for 82% (490 000). For the same report, the global number of notified TB cases is estimated to be 350 000 (range, 330 000–370 000) MDR/RR-TB cases among notified TB patients (WHO, 2017). Considering the burden of this infection, WHO has recognized that TB as a global public health emergency and launched direct observation therapy strategy (DOTS) in 1994 because of a

number of peoples exposed to this disease and many deaths was registered (WHO, 2017; Deribew *et al*., 2012; FMoH, 2015).

According to the WHO (2017) report, Africa is not among the regions registered to have a declined in TB mortality rates. The variation in the country in CFR has varied especially from under 5% in a few countries to more than 20% in most countries in the WHO African Region. Thus, this is an indicator for the inequalities in accessing the diagnostics of TB disease. The summary of WHO report indicated that the incidence rate of TB and HIV co-infected in WHO Africa region is estimated to be 41 (34-48) per 100000 population whereas mortality for the same cases is 31(27-36). In 2016, the total notified TB cases in this region was 1303483 with 84% of pulmonary cases which intake an estimated MDR/RR-TB cases of 40000 (ranging from 36000 to 44000) among notified pulmonary TB cases. The estimated TB treatment coverage in the WHO Africa region is only 49% (WHO, 2017).

Different reports and studies certified that Ethiopia has only limited resources to spend on combating tuberculosis and multidrug-resistant tuberculosis. It ranked the ninth among the world most TB burden country and is one of 27 MDR TB high burden countries. In 2016 only 182 (ranging, 128-245) thousand TB incidence, of which 14 (9.6-19) was related to HIV co-infection has occurred in Ethiopia and the estimated notified co-infected people was 103330 (81%). The rate incidence of the cases for the same year is found to be 177/100000. In another way the number of deaths due to TB cases without HIV co-infection, was estimated to be 26 thousand where the death rate is 25/100000 and whereas 4 (2.7-5.4) thousands of HIV co-infected died (WHO, 2017; Deribew *et al*., 2012; FMoH, 2015).

According to different studies, a south west region of Ethiopia was one of the areas recorded to have high risk of Tuberculosis disease. In Jimma zone (south west region) around 10.9% of the cases was recorded annually. The expanding of the diseases was more focused on especially in the two priority risk groups of people living with HIV and children under 5. It shares estimated cases of 25%. Similarly, the most death of TB cases with HIV co-infected is also accounted for this region which is found to be 82% in 2016 (Ali *et al*., 2017; Asemahagn *et al*., 2018).

Why Bayesian? Although the application of Bayesian statistics sounds the researchers, it stayed long century with its theoretical definition only because of its difficulties with the integration of the denominator in Bayes theorem. Thanks to simulation-based MCMC methods, the approach got valued to have numerical meaning with the efficient estimation of the application in any fields with some limitation like the burden of time in approximating the posterior and convergence problem (Gelman *et al*. 2009; Berger, 2013). As of 2009, the other news was welcomed with very flexible and fast approximation techniques called Integrated Nested Laplace Approximation (INLA) for Latent Gaussian Model (Rue *et al*., 2009).

With this study, the reason why the Bayesian approach is preferred over the usual frequentist technique is that the power of information obtained from the approach is much better as it is the combination of likelihood data and prior information about the distribution of the parameter. It still empowers the efficiency of the data even when the size of observation may large enough in representing the target population by giving distribution for the unknown parameters. Thus, considering the stated advantages of Bayesian application over classical method and the interesting application of INLA with Latent Gaussian Model (LGM) method are the most key for the motivation to apply it for the data set under this study (Riebler *et al*., 2017; Blangiardo *et al*., 2015).

Latent Gaussian Model (LGM) forms a flexible subclass of Bayesian hierarchical models. Its practical application from a statistical modeling point of view is readily interpretable. Consequently, LGMs have become popular in many areas of statistics and various fields of applications especially in the spatial and spatiotemporal model (Nzabanita, 2012). The Integrated Nested Laplace Approximation (INLA) proposed by Rue *et al*. (2009) is focused on providing a good approximation to the posterior marginal distributions of the parameters in the model of the Bayesian hierarchical framework. In particular, this approximation has been developed for Latent Gaussian models.

1.2. Statement of the Problems

The different study reported from various parts of the country showed that the prevalence of smear-positive cases ranged from 33 to 213.4/100,000 people in Ethiopia. This burden of the diseases were gradually increased till the year of 2016 (Deribew *et al*., 2012; Asemahagn *et al*., 2018). Considering the seriousness of the disease and gaps found with different studies, the researcher has fitted the latent Gaussian model with the Bayesian hierarchical approach using INLA method. Therefore, this study has addressed the gaps seen with previous studies; with especial weight to model gaps used by different researchers.

According to different studies of Bayesian GLMM of TB cases, the model with Bayesian approach have empowered over the frequentist (Jaya *et al*., 2014; Randremanan *et al*., 2010; Gelman *et al*., 2009). However, those studies were based on the application of the simulation-based MCMC method which has the burdensome of time-consuming, convergence problem and Monte Carlo error. Thus, with this study, the deterministic, fast and promising alternative of MCMC called INLA to approximate the posterior marginal has been applicable (Rue *et al*., 2009; Blangiardo *et al*., 2015; Martins *et al*., 2013). Besides, for the study with random effects and Poisson distribution of the observation, the offset variable is considered to adjust the number of events and population size. This concept was ignored with those previous studies are considered under this study (Tonui *et al*., 2018; Ojo *et al*., 2017; Randremanan *et al*., 2010).

The previous studies with INLA methods also were based on the default priors only that sometimes were bad and without further concise with the approximation methods of INLA. This thesis, therefore, addressed the problem of prior assignment by considering the informative Penalized Complexity (PC) prior and intended on the application of different approximation methods of INLA (Riebler *et al*., 2017; Bivand *et al*., 2015; Kipruto *et al*., 2015)

Thus, the study have attempted to answer the basic research questions on: whether there is variation in the distribution of TB cases among the districts of Jimma zone, whether changes in prior assignment really affect the candidate model to be selected and answered the questions on how to apply the latent Gaussian model with INLA methods under the framework Bayesian hierarchical paradigm.

1.3. Objectives

1.3.1. General Objective

The general objective of this study was to model the counts for TB cases in Jimma zone using the Bayesian hierarchical approach of the latent Gaussian model with INLA method.

1.3.2. Specific Objectives

i. To identify the predictors of TB cases and see the variation in the distribution of the cases across districts.
ii. To compare the R-INLA's inbuilt default priors with the informative penalized complexity priors for robustness of the priors.
iii. To fit the latent Gaussian model with INLA methods under the framework of Bayesian hierarchical paradigm.

1.4. The Significance of the Study

The results of this study may help the organization as well as individuals who work in this area to get a clue on to what extent TB distribution is serious across the districts of Jimma zone. It may also be an input to see the trend of TB prevalence by comparing the result of this study with previous studies. The other basic significance of the study is that it may also further assist other researchers interested in this area and they may use it as a benchmark for their future works. In determining the posterior distribution, MCMC simulation technique is the most applicable methods used for a long period of time. But, recently (as of 2009) very fast, convenient and a very fast representative approximation technique called INLA which designed for the latent Gaussian model is availed. With this study, therefore, researchers will benefit by getting familiar with the method and may further help in advertising the approximation technique. The result of this study will also be expected to help those make a policy of any TB concern agendas and strategies.

CHAPTER TWO

2. LITERATURE

2.1. Overview of Tuberculosis Cases

Tuberculosis is a bacterial disease which caused by Mycobacterium tuberculosis, which is spread by airborne droplet nuclei consisting of tiny particles (between 1mm and 5 mm in size) that contains the bacteria and produced by untreated pulmonary TB patient while talking, coughing, sneezing and it remains a high-priority communicable disease that causes an enormous burden of morbidity, mortality which infected an approximately one-third of the global population, and is the second leading cause of death among infectious diseases worldwide(WHO, 2017). It is a rod-shaped, non-spore forming, and is neither gram-positive nor gram-negative aerobic bacteria. Because of its thick cell wall, the bacterium does not decolorize after staining with acid and is therefore known as acid-fast bacilli (AFB) (Pedro *et al*., 2017). A person with active TB can infect 10-15 persons with any of the contacts in the course of a year through close contact (Moghaddam *et al*., 2016)

Tuberculosis primarily affects the lungs (pulmonary TB) in about 80% of the cases but can affect any organ of the body (extra-pulmonary TB) including bones, skin, brain, vertebral spine among others if it is not immediately treated. The very common symptom of TB is a cough for a duration of two weeks or more which is usually accompanied with fever, weight loss, night sweats, chest pain, shortness of breath, tiredness, loss of appetite and in some instances hemoptysis which may serious as the time of infection is being increased (Moghaddam *et al*., 2016). It is, however, a treatable and curable disease, and the common treatment of the disease are the four anti-Tb drugs namely Rifampicin, Isoniazid, Ethambutol and Pyrazinamide (Karumbi *et al.*, 2015). Thus, the initial intensive phase of TB treatment is intended to kill actively growing and semi-dormant TB bacilli. The intensive phase treatment can shorten the duration of infectiousness by rapid smear conversion (80–90%) following two-three months of treatment using fixed-combination dose of the four drugs (WHO, 2015).

In the world, a number of people were suffering from TB disease. Different reports and studies revealed that approximately two to three billion of the world's populations are estimated to be infected with Mycobacterium tuberculosis (WHO, 2015; Hayward *et al*., 2018). Once individuals are infected with TB disease, the probability of time to develop the disease is determined by their age, immunity, and duration of infection. Patients with latent TB cases have a chance of having progressive active TB case 10%-20% during their lifetime (Martinez *et al*., 2017). Even though the disease is serious anywhere in the world, continents have shared a different proportion of being infected. The global health organization report indicated that Asia takes the major burden of the cases (58%) and Ethiopia share 28%; whereas eastern Mediterranean region, America and European have the smaller proportion of the disease risk which is 8%, 3%, 3% respectively (WHO, 2015).

Globally, the absolute number of TB deaths among HIV negative people has been falling since 2000, from 1.7 million in 2000 to 1.3 million in 2016. The TB mortality rate (per 100 000 population) fell by 37% between 2000 and 2016, and by 3.4% between 2015 and 2016. Rates have also been falling in all six of the WHO regions. Since 2010, the fastest average rates of decline in the mortality rate have been in the WHO European Region and the WHO Western Pacific Region (6.0% and 4.6% per year, respectively), and lowest in the WHO Eastern Mediterranean Region (2.2% per year). Trends in mortality rates in the 30 high TB burden countries vary markedly, ranging from substantial reductions since 2000 (e.g. in Cambodia, China, Ethiopia, Myanmar, the Russian Federation, and Viet Nam) to limited changes (e.g. in Angola, Congo, and South Africa). High TB burden countries with rates of decline exceeding 6% per year since 2010 included Ethiopia, the Russian Federation, the United Republic of Tanzania, Viet Nam and Zimbabwe is more affected by the disease (WHO, 2017)

The key purposes like integrated, patient-centered care and prevention, bold policies, supportive systems, and intensified research, as well as innovation, are the main pillars and components of the strategies of Sustainable Development Goals (SDGs). The achievement of these goals warrants a continued effort in both low- and high-incidence countries towards controlling the disease. The SDGs' plan to end the TB epidemic is to achieve an average reduction of TB incidence by 5% per annum until 2025, and then by 4% per annum, in order to reach the 2035

global target but the accomplishment during the year of 2000 to 2014 is too far from the plan which was only 1.5% (Lönnroth *et al*., 2015). With this under-achievement, it seems must conduct technical research to identify the gap so that the futures of the agenda of SDG have to accomplish accordingly.

2.2. Tuberculosis in Ethiopia

Even though Ethiopian health institute has a different structure of working and introduced sectors to prevent communicable disease, but still the issue of controlling the communicable disease remains to be not well addressed. Thus, according to the federal ministry of health report, to do further on this problem the country has formed Health Sector Transformation Plan (HSTP) during the second growth and transformation plan of 2015 by giving more emphasis for the disease like malaria, HIV and TB (FMoH, 2015).

According to the study conducted by (Kebede *et al*., 2014), the history of controlling TB in Ethiopia was started in the early of 1960s in very limited urban areas; but it becomes popular following the establishment of national TB control program office in 1976. Since then, application of different strategies including DOTs extensively exercising with better improvement through time. However, the disease is till serious throughout the country especially following the rapid increment of the population. The FMoH report indicated that Ethiopia is among the 22 high-TB-burden countries in the world. It also accounts the proportion of 1.6% and 12%, for new and retreatment cases respectively of 27 high-MDR-TB-burden countries worldwide and the disease is the third leading cause of hospital admissions and the second top cause of death in the country (WHO, 2015; FMoH, 2015).

The report of world health organization indicated that the number TB infected people in Ethiopia are increased through years and the annual measure of the disease in the country is beyond the average annual incidence of the world. As an indicator, in 2015 the annual incidence of all forms of TB cases in Ethiopia is estimated to be 207, while that of prevalence is 200 per 100,000 people and the result is higher than the annual incidence and prevalence of people in the world which was 174 and 133/100,000 respectively considering for the year. In 2014, the TB case

notification for all forms of TB cases for the respective measure was 123 and 43.3 per 100,000 people for smear-positive TB in Ethiopia (WHO, 2015). Generally, the occurrence of TB cases in Ethiopia has been increasing from year to year; since the chance of being infected is high with the dense population. Hence, the disease has become more serious in the country for which multidirectional prevention methods were established and yet not be well addressed.

2.3. Determinants of Tuberculosis Disease

The progress of TB infection developed to active TB is conditioned by many factors that the development of TB in a person is a two-stage process, in which a susceptible individual is exposed to an infectious TB case and becomes infected, and may later develop active TB (Legido *et al*., 2013). The prevalence of infectious TB cases and the duration of infectiousness are important factors that increase the risk of infection in the general population. An untreated TB case remains infectious unless such patients have access to TB diagnosis and treatment. Close contacts of TB patients such as household contacts and caregivers including healthcare workers are particularly at a higher risk of becoming infected with TB (Hamusse *et al*., 2016). According to the recent studies conducted by (Chalovich *et al*., 2013) which entitled the role of casual contacts in the recent transmission of tuberculosis in settings with high disease burden, the casual transmission of TB can take place within a short contact period in both high and low incidence settings. The impacts of determinants of TB including all factors registered under health office are reviewed as follow.

2.3.1. Gender

The world health organization report of 2017 quantified that the number of male TB infected is greater than that of the female. It indicated that of all 10.4 million estimated TB incidents 6.7 (range, 3.7 to 8.6) million is found to be male. This implies that for every one case of TB incidence among females, there were about 1.3 incidences of TB cases among males for the eastern Mediterranean region to 2.7 western Pacific regions. Globally the ratio of notified TB cases among is 1.7 to those of female (WHO, 2017)

Sulis *et al*. (2016) conducted a study to identify the evolution of TB infection in the world community. Considering control and prospects for reducing tuberculosis incidence, prevalence and deaths globally, the study found that the progression of TB infection to disease is high among women in the reproductive age group compared to men. The socio-cultural factors, including an inability to make decisions on resources, stigma and poor health-seeking behavior may hinder a women's ability to utilize the existing health services. The study also assured that hormonal factors might play a role in the risk of tuberculosis infection and its progression to active disease.

Heunis *et al*. (2017) made a study on Risk factors for mortality in TB patients: a 10-year electronic record review in a South African province. The study was analyzed using the adjusted odds ratio and found that the number of males with TB cases was 53%. But, the adjusted odds ratio indicated that the difference between sexes in a sense was not such significant.

According to the research of Biruk *et al*. (2016) which is to identify the treatment outcomes of tuberculosis and associated factors in an Ethiopian university hospital, the study includes 1584 sample TB infected patients and found that 882 were males and only 702 were female. Moreover, the variable was significant in the model that the odds of the female are less than that of the male. Hence, studies that could help to understand the interaction between the biological, health system and socio-cultural determinants of gender-based variances are needed to be undertaken.

2.3.2. TB/HIV Co-infection

HIV is known to be the most powerful determinant in increasing the risk of TB infection and its progression to active disease. Approximately 80% of the total of estimated HIV associated TB is found in the countries of Sub-Saharan Africa. Of the 9.27 million TB cases existed globally, 15% (1.37 million) cases were HIV-related tuberculosis, of which 79% were from the African region which is almost 1.1 million cases. Because of this interaction, these two diseases form a vicious cycle (Getahun *et al*., 2010; WHO, 2017). Similarly, Getahun *et al*. (2010) also studied on co-infection of TB/HIV resulted as there was a positive relationship between the prevalence of TB

and HIV. The result of the study suggested that the prevalence of HIV infection was significantly associated with the incidence of TB. Thus, the strong spearman correlation($r=0.69$, $p<0.01$) also certified the association.

Sreenivasulu *et al*. (2018) found that the co-infection of TB-HIV is a major problem of public health importance. They concluded that HIV is a major determinant in the treatment outcome among the tuberculosis patients and if patients infected with HIV, also then the death rate increases by 25% more compared to those tuberculosis patients who are not infected with HIV. Among HIV patients also tuberculosis is the most common opportunistic infection. HIV is acting as an important hurdle in the targets of TB treatment success. This is especially applicable for countries with a high burden of HIV.

In Sub-Saharan Africa, the TB epidemic has worsened because of its interaction with HIV and AIDS (Corbett *et al.*, 2003). The results indicate that tuberculosis infections among immune competent individuals remain asymptomatic and become latent infections. Still, there is a high risk of for the progression of a primary infection to active TB among HIV infected individuals. Likewise, the risk of a latent TB progression to an active one among HIV-infected individuals is roughly 20-30 times more likely compared to non-HIV infected ones. Behind this, the treatment of active TB among HIV co-infected TB patients are complicated due to adverse drug reactions, drug interactions, and less favorable patient outcomes, with an increased likelihood of mortality, lower cure rates, and lower treatment success rates compared to non-HIV infected TB patients (WHO, 2014).

Hayward *et al*. (2018) have found that Infection with HIV is the strongest known risk factor for the development of TB disease. According to their study, TB-HIV co-infection synergistically worsens both conditions, leading it to be termed the cursed duet and HIV increases both the risk of rapid progression to active disease following infection and reactivation with an increased risk of TB throughout the course of HIV disease and incidence rate ratios >5 when averaged across all levels of immunodeficiency.

According to the assessment report of Datiko *et al*. (2009), producing health extension worker in Ethiopia, aimed to prevent the communicable disease, the extension health workers have been

successfully involved in TB identification: collecting sputum and storing and transporting it to the nearest microscopic centers for TB diagnosis. This resulted in improved case detection: 122.2% in intervention areas compared to 69.4% in control areas, with women showing the highest rise in case detection which further benefited the HIV co-infected not to develop active TB cases. Generally, since many studies have supported as HIV is a major factor that hurries the TB cases to be active, it is also considered in this study as one factor of TB cases.

2.3.3. Age

Differences in TB infection and disease burden across various age groups have been reported from different parts of the world. Addisu *et al*. (2018) conducted study entitled factors associated with poor treatment outcome of tuberculosis in Debre Tabor northwest Ethiopia. To identify the associated risk factors adjusted the odd ratio of logistic regression was used with which Age was found to be a significant factor. For this study patient of age, less than 14 was treated as reference categories and all other categories of their age were indicated significant difference as compared to the reference group.

The risk of infection increases from early infancy to early adult life, possibly due to the increasing number of social interactions and frequency of contacts. The incidence of TB rises from early infancy to pre-adolescence but falls as the time from of infection increases. Moreover, young children often developed disseminated TB such as TB meningitis and extra-pulmonary TB, which affected organs other than the lungs as a result of immature cellular immunity to localized TB bacilli. Tuberculosis is predominantly reported to be a disease among the adult population in the productive age group from 15 to 49 years (Pahlavanzadeh *et al*., 2016; Gugssa *et al*., 2017).

2.3.4. Population Density

The Population density has a significant effect on the existence of TB especially as the TB transmits airborne particles in the increase for the dense area. Since the infected person has the chance of transmitted the disease to around 5-10 person, the number is inflated due to the density of people in the limited land. According to the study conducted by Padmanesan(2013) that has the objective of identifying the risk factors of TB cases, crowded living conditions were found to

have significant effect to increase the rate of TB spread through increased contact between infectious and susceptible individuals.

The different study had observed to have high TB rates in areas of the high population (Jacirema *et al*., 2009; Girum *et.al*., 2018; Legido *et al*., 2013). However, the study conducted by Harling *et al*. (2014) indicated that the observed that population density was a predictor of high tuberculosis rates regardless of poverty and urban residence which was partially mediated by higher rates of HIV/AIDS rates in the municipalities. High population density is associated with outdoor residential crowding experienced in cities especially in urban slums and informal settlements characterized by lack of basic sanitation, poor housing, and overcrowding, high levels of congestion and urban air pollution as a result of increased vehicular movements, industrial pollution, effluent from generating sets and household fuel combustion. These situations may contribute to increased respiratory illness including TB (McMichael, 2000).

2.4. Overview of Bayesian Modeling

The Bayesian statistical methodology presents a well-established framework for making an inference from observed data for quantities of interest by using an underlying probability model for a comprehensive overview of modern Bayesian statistical analysis. The Bayesian methodology differs from the classical frequentist approach in that all of the unknown parameters in the underlying probability model are treated as random variables, as opposed to unknown constants in the classical frequentist approach. As such, the unknown parameters are assigned prior distributions which are based on a priori subjective beliefs or scientific knowledge about the unknown parameters. In other words, prior distributions serve as probabilistic descriptions of what is known about the unknown parameters before observational data are collected and analyzed (Gelman *et al*., 2013; Berger, 2013).

According to the study conducted by Geirsson *et al*. (2014) the temperature is depending on where and when it is measured. The study was aimed to test the spatial modeling of annual minimum and maximum temperature in Iceland, the observed data, consisting of measurements of temperature, exhibit a latent dependence structure in the sense that the temperature is

dependent on where and when it is measured. It also indicated that including the random effects were significant.

Gelman *et al*. (2009) defined Bayesian hierarchical modeling as a systematic modeling methodology to capture the latent dependence structure of observed data. Moreover, the Bayesian paradigm is flexible to handle multiple parameters where the model parameters are related or dependent in a systematic manner. The resulting joint probability model should thus reflect the dependence structure of the parameters. Further, it is considered natural to structure a model of this type hierarchically, with observable data modeled conditionally on a certain set of model parameters, which in turn can be potentially dependent on another set of model parameters.

Jaya *et al*. (2014) conducted a study to model Bayesian conditional auto-regression (CAR) and map tuberculosis cases in India with Win BUGS software to assess the spatial pattern of TB. The study observed that states in the North-eastern region had a higher risk of tuberculosis compared with other regions in the country The Bayesian CAR model, in addition, provided a smoothed map of the Standardized Incidence Ratio (SIR) that had fewer extreme relative risk values compared with the raw unsmoothed SIR values.

Randremanan *et al*. (2010) in Antananarivo city, Madagascar have used a combination of a Bayesian approach and generalized linear mixed model in order to spatially model TB cases and identify the potential risk factors of the disease. The Bayesian spatial modeling approach in Markov Chain Monte Carlo (MCMC) used the Win BUGS software to detect clusters of TB. Tuberculosis was associated with households with more than one TB cases and households who had a TB patient that had been lost to follow-up. In comparing the two approaches (the spatial scan statistics and the Bayesian approach) in Antananarivo, it was observed that the spatial scan method detected general regions with significantly high risk for TB. However, the spatial scan method generated larger clusters than expected thus had more high false positive areas than the Bayesian approach. The Bayesian approach, on the other hand, identified neighborhoods that significantly contributed to the scan statistics circle but with a lower false positive rate.

Musenge *et al.* (2013) conducted Bayesian Spatio-temporal analysis using the INLA package in R software in the study of childhood TB/HIV mortality in South Africa. The study identified three hotspots areas in the central, southeasterly and southwesterly regions. The factors protective of childhood mortality were the number of adults in the household, the number of antenatal clinic attendance and mother being alive. Households with a higher socioeconomic status had significantly lower childhood deaths compared with poorer households.

Kipruto *et al*. (2015) undertook a spatiotemporal modeling of tuberculosis in the 47 regions in Kenya using the Bayesian Hierarchical generalized linear mixed model in the INLA package in R software (Rue *et al*., 2009). The study identified TB hot-spots in 11 regions namely Nairobi, Mombasa, Marsabit, Isiolo, Lamu, Machakos, Kajiado, Makueni, Kisumu, Siaya and Homabay. In addition, the study found a significant association between TB and risk factors such as age, gender, and HIV.

CHAPTER THREE

3. DATA AND METHODOLOGY

3.1. Study Area

Jimma is one of the zones in the Oromia regional state of Ethiopia and is named for the former kingdom of Jimma, which was absorbed into the former province of Kaffa in 1932. The capital town of the zone is Jimma which is the largest city in south-west Ethiopia. The zone has a latitude and longitude of $7^{o}40'N\ 36^{o}50'E/7.66^{0}N\ 36.833^{0}E$ and the temperature at Jimma are in a comfortable range, with the daily mean staying from 20 to 25 degree Celsius. Recently the zone includes around 22 districts. Based on the 2007 census conducted by the CSA, this zone has a total population of 2,486,155 and has an area of 15,568.58 square kilometers. It has a population density of 159.69.

3.2. Source of Data

The data for this study was mainly based on the secondary data that has been obtained from Jimma zone and Jimma district health office except for data related to population density. All the cases registered on the data base of the office have been considered. The population size has been taken from the central statistical agency. Since the latest census in Ethiopia was held in 2007 which seems a bit old; population projection on the current was determined by considering the expected annual increase of the population in Jimma zone. All data on any forms of TB cases and identified covariates except population density have been obtained from the district and zone health office registered from September 2016 to August 2017 which was recorded for a year.

3.3. Variables under Study

3.3.1. Response Variable

The dependent variable of this study was the **count of TB cases** (all forms) in each district of Jimma zone recorded under the health office from September 2016 to August 2017.

3.3.2. Explanatory Variables

According to the different reviewed study discussed in literature parts, the explanatory variables considered for this study which has been registered in the health office were gender, HIV co-infection, population density and age of patients.

3.4. Methods of Data Analysis

In any research design, an appropriate data analysis plays a crucial place for the relevance of data under consideration. Thus, to fit the data well, the researcher has been passed through different stages of data analysis for which the techniques were presented under sub-sections here below.

3.4.1. Bayesian Hierarchical Model

The Bayesian approach to inference allows parameter estimation using information coming from the data via the likelihood function as well as information coming from other sources (i.e. previous studies, subjective judgments) which is formalized via prior distributions. These probability statements are conditional on the observed value. Generally, the Bayes theorem is termed as:

$$P(\theta/y) = \frac{P(y/\theta)P(\theta)}{\int P(y/\theta)P(\theta)} \qquad [3.1]$$

The so-called Bayesian hierarchical models are very recent and attractive as they provide a unified approach to data analysis (Samuels *et al.*, 2015). They are usually characterized by three stages of observations and parameters. The first stage consists of distributional assumptions for the observations. For this data, the TB counts y_i (i =1, . . , 21) for i geographic districts within a pre-specified time period were applied. It is assumed that y_i followed the Poisson distribution with the rate λ_i . The parameterλ_i denotes the relative risk for the TB cases in district i. The y_i's are conditionally independent given all $\lambda_i's$. On the second stage, the priors were defined for all the parameters $\lambda_i's$ or, more often, a specific transformation of them. For the Poisson distribution, it is common to use log(λ_i) =η_i The variable η_i is called the linear predictor and is usually an additive term of unknown random components (Octavianty *et al*., 2017).

The unknown random components are of different types like spatial random effects and linear or smooth effects of covariates. High flexibility can be obtained by assigning Gaussian priors to all components of the linear predictor. Such models are also called latent Gaussian models (Rue *et al*., 2009). The third stage consists of prior distributions for unknown hyper-parameters $\theta 1,\ldots,\theta r$, which typically are variances or correlations for random effects within $\eta = (\eta_1 \ldots \eta_i)^T$. The hyper-parameters are not expected to necessarily have Gaussian distribution. The prior assignment for all parameters in the model has been discussed below in section 3.4.6. Bayesian hierarchical models with latent Gaussian layers have proven very flexible in capturing complex stochastic behavior and hierarchical structures in high-dimensional data.

3.4.2. Likelihoods

The likelihood is the function of the distributional assumption of the observation. For this dataset, since it is characterized to count data, Poisson distribution has been used. The observed number of TB patients in each district of Jimma zone which was y_i; i=1, 2,…,21 has been considered. Thus, the distribution of TB cases y_i~Poisson (λ_i) where λ_ia function of relative risks for the disease under consideration was taken as the distribution of the TB cases with log canonical link function.

In order to account for different population sizes of districts, the researchers preferred to compute the expected number of patients in each district, which has been used as a scale factor. This technique of scaling is known as an offset variable. The offset variable is mainly used to adjust different sizes of the population in each district so that to inline the effect of the population size of the districts with its corresponding TB cases (Srinivasan *et al*., 2014). Moreover, the attractive application of this offset variable is that rather than adjusting variation due to population size, it is not presented as the explanatory variable since it always has a coefficient value of 1. Then, the expected number of patients in each district can be determined as:

$$E_i = \theta_i \frac{\sum_i y_i}{\sum_i \theta_i}$$

Where y_i the number of TB counts for the ith district, and θi is all population under the risk of TB disease in districts i

3.4.3. Latent Gaussian Model

The structured latent Gaussian regression models amenable to INLA-based inference can be defined in terms of three layers: hyper-parameters, latent Gaussian field, likelihood model. The univariate likelihood captures the marginal distribution of data and is often chosen as an exponential family similar to the framework of generalized linear models.

For those exponential family models, the link function is used to have the linear relationship with the response variable. Hyper-parameters can appear in the likelihood as dispersion parameters like the variance of the Gaussian distributions (Opitz, 2016). Formally, the Latent Gaussian Model (LGM) can be written as:

$y/x, \theta_i \sim \prod p(y_i/\eta, \theta_2)$ Likelihood [3.2]

$x/\theta_1 \sim p(x/\theta_1) = N(0, Q)$ Latent Field [3.3]

$\theta = [\theta_1, \theta_2]^T \sim p(\theta)$ Hyper-priors [3.4]

The dimension of the latent field x can be large (10^2 -10^5) and the number of hyper-parameters is not expected to exceed 6. This helps to reduce the complexity of the model (Hosseini *et al*., 2011).

Thus, considering the latent Gaussian model, the specific generalized linear mixed model for cases of TB counts has form of:

$y \sim \prod_i^N p(y_i/\lambda_i)$

$\log(\lambda_i) = \beta_0 + bo + \beta_1 * Sex + \beta_2 * density + \beta_3 * HIV + \beta_4 * Age + \log(offset) + \varepsilon_i$

Where y is an observed dataset (count of TB cases), x is the joint distribution of all parameters in the linear predictor (including itself) which is η, β_0, β_1, β_2, β_3, β_4. The coefficients $\beta 0$ stands for the total average TB occurrences keeping covariates at reference categories for categorical variables and constant for continues covariates, bo is the average TB occurrence in district level and, β_1, β_2, β_3, β_4 were the coefficients of sex, HIV co-infection, population density, and Age respectively. All fixed covariates have the joint Gaussian distribution of mean zero and small variance and whereas the random effect

follows the Gaussian distribution with mean zero and inverse variance precision matrix which denoted by θ that are the hyper-parameters of the latent field which is not necessarily Gaussian. (Rue *et al.*, 2009; Opitz, 2016).

Thus, the model is said to be Latent Gaussian Model (LGM), provided that if and only if there is strong assumption that the parameters have joint Gaussian distribution and it can be achieved by assigning normal priors for each element of latent fields. With this, we mean that x is the joint distribution of the parameters of the linear predictor including itself.

$$x = [\eta,\ \beta_0,\ \beta_1, \beta_2,\ \beta_3,\ \beta_4, b_0] \sim N(0, Q) \qquad [3.5]$$

This is an indicator that the joint parameters under consideration ascertain for the existence of latent Gaussian model in which its precision matrix (inverse of the variance) can handle the variability among the random effect (Rue *et al.*, 2009). If we assumed conditional independence in x, then this latent field x is a Gaussian Markov Random Field (GMRF).

3.4.4. Gaussian Markov Random Field

We can form a large vector x, which consists of the linear predictor vector η^{T} and all its additive components. As the η_{i}'s are on the first I positions in the vector x, each observation y_i depends directly only on the corresponding i$^{\mathrm{th}}$ element xi in x. Furthermore, since Gaussian priors are assigned to all components of x as mentioned above, the vector x is also Gaussian and forms a so-called Gaussian Markov random field (GMRF) (Bolin, 2015; Rue *et al.*, 2005).

A GMRF is a random vector following a multivariate normal distribution with Markov properties: for $i \neq j,\ x_i \perp x_j / x_{-ij}$. The notation $-ij$ refers to all elements of x other than i and j. The conditional independence between two components x_i and x_j of a GMRF can directly be read off from its so-called precision matrix Q. It holds that: $x_i \perp x_j | x_{-ij} \Leftrightarrow Q_{ij} = 0$; more formally we can define a random vector $x = (x_1, \ldots x_n)^T$ as GMRF with mean μ and

positive definite precision matrix Q. Sturtz (2008) provide a description of methods for efficient computation of GMRF which can be used to speed up computations and provide fast approximations. GMRF is the key to providing good Gaussian approximations for the posterior marginal if its density of the above equation [5] has the form:

$$\pi(x) = (2\pi)^{-\frac{n}{2}}|Q^{\frac{1}{2}}|\exp(-1/2(x-\mu)^T Q(x-\mu)) \qquad [3.6]$$

The covariance $\sum = Q^{-1}$ of the GMRF is the inverse of the precision matrix. At the second stage within a hierarchical model, GMRFs provide a flexible tool to model the dependence between latent effects and thus, implicitly, the dependence between the observed data (TB cases) (Rue *et al*., 2009).

3.4.5. Integrated Nested Laplace Approximation

Integrated nested Laplace approximation (INLA) is a recent approach to Bayesian statistical inference for latent Gaussian Markov random field models introduced by (Rue *et al*., 2009). It provides a fast, deterministic alternative to Markov chain Monte Carlo (MCMC) which, at the moment, is the standard tool for inference in such models of Bayesian inference.

The main advantage of the INLA approach over MCMC is that it is much faster to compute and is promising accurate; it gives answers in minutes and seconds where MCMC requires hours and days. The fundamental idea of INLA consists in applying the device of Laplace approximation to integrate out high-dimensional latent components. This theoretical foundation is combined with efficient algorithms and numerical tricks, and approximations to ensure a fast yet accurate approximation of posterior marginal densities of interest like those of the predictors or of hyper-parameters. This approximation technique is specifically designed for the latent Gaussian model (Rue *et al*., 2009).

The main goal of the approximation techniques used in the analysis of the latent Gaussian model is to compute posterior marginal for each component of *x* of expression [3.5]. Generally, the marginal posterior distributions for each element of the parameter vector can be formulated as:

$$\pi(xi/y) = \int_{\theta} \pi(xi/\theta, y)\pi(\theta/y)\, d\theta \quad [3.7]$$

And the marginal posterior distribution for each element of the hyper-parameter vector:

$$\pi(\theta_j/y) = \int \pi(\theta/y) d\theta_{-j} \quad [3.8]$$

Now, our intention was to compute $\pi(\theta/y)$ from which all the relevant $\pi(\theta_j/y)$ can be obtained and to determine $\pi(x_i/\theta, y)$ which is needed to compute the parameter marginal posteriors $\pi(x_i/y)$

Assuming $\tilde{\pi}(x_i/\theta, y)$ and $\tilde{\pi}\ (\theta/y)$ were approximations of $\pi(x_i/\theta, y)$ and $\pi(\theta/y)$ respectively, we further demonstrated different techniques of the approximation. The approximation for the hyper-parameter can be approximated as:

$$\tilde{\pi}(\theta/y) \propto \frac{\pi(x,\theta,y)}{\tilde{\pi}_G(x/\theta,y)}|_{x=x^*(\theta)} \quad [3.9]$$

Where $\tilde{\pi}_G(x/\theta, y)$ is the Gaussian approximation to the full conditional of x and $x^*(\theta)$ is the mode of a full conditional for x for a given θ.

To compute out the posterior marginal of the latent field, we started by determining the Gaussian approximation $\tilde{\pi}_G(x/\theta, y)$ and to approximate the density of $x_i/\theta, y$ as

$$\tilde{\pi}(x_i/\theta, y) = N(x_i;\ \mu_i(\theta), \sigma_i^2(\theta)) \quad [3.10]$$

Then, the integral for each element of the latent Gaussian to determine posterior marginal can be approximated with the finite sum as.

$$\tilde{\pi}(xi/y) = \sum_k \tilde{\pi}\ (x_i/\theta_k, y)\tilde{\pi}(\theta_k/y)\Delta_k \quad [3.11]$$

$\tilde{\pi}(x_i/\theta, y)$ and $\tilde{\pi}(\theta/y)$ denote approximations of $\pi(x_i/\theta, y)$ and $\pi(\theta/y)$ respectively. Finally, the sum is evaluated at support points θ_k by numerical integration using appropriate weights Δ_k.

Generally, the approximation for the expression [3.5] can be computed in three steps. The first step approximates the posterior marginal of θ by using the Laplace approximation. The second

step computes Laplace approximation or the simplified Laplace approximation for x and finally, the two was combined with numerical integration (Blangiardo *et al*., 2015; Rue *et al*., 2009).

3.4.5.1. Approximating the latent field parameters

Besides of approximating the posterior marginal of θ, the other ultimate goal of doing with INLA was to get the accurate approximation for $x_i's$ of equation [3.5] that are conditioned on the θ.

Considering $\tilde{\pi}(x_i/y,\theta_k)$ denote the approximation of the latent field parameters, three approximation techniques; i.e. Gaussian approximation, Laplace approximation and simplified Laplace approximation have been applicable.

The denominator in the equation [3.9] is the simplest Gaussian approximation of the posterior marginal for the latent field. This Gaussian approximation often gives reasonable results, but there can be errors in the location and/or errors due to the lack of skewness (Rue *et al*., 2007; Blangiardo *et al*., 2013). The second and most natural of computation is the so-called Laplace approximation (LA). It can have the general form of:

$$\tilde{\pi}_{LA}(x_i/\theta,y) \propto \frac{\pi(x,\theta,y)}{\tilde{\pi}_{GG}\left(\frac{x_{-i}}{x},\theta,y\right)}\Big|_{-x_{-i}=x^*_{-i}}(x_i,\theta) \qquad [3.12]$$

This method of approximation is computationally expensive; because the Gaussian approximation in the denominator of expression [3.12] must be recomputed for each value of x_i and θ since the precision matrix depends on such parameters.

The other most efficient and computationally simplest method is simplified Laplace approximation (SLA). The technique is based on Taylor's series expansion up to the third order of both numerator and denominator for $\tilde{\pi}(x_i/\theta,y)$. The series expansion was effectively correcting the Gaussian approximation for location and skewness to increase the fit to the required distribution (Rue *et al*., 2009; Baio, 2013; Blangiardo *et al*., 2013).

To identify which approximation method (LA or SLA) is better in approximating the posterior marginal, the popular comparison technique in INLA called Kullback-Leiber divergence has been applied. It was the automatic output in any model results of INLA. The value of the Kullback-Leibler divergence (KLD) describes the difference between the standard Gaussian and the Simplified Laplace approximation to the marginal posterior densities or between standard Gaussian and full Laplace approximation keeping Simplified Laplace approximation as first desire if the criteria were met to reduce computation burden. The smaller the KLD, the better approximation of the posterior distribution is. The default approximation method in INLA was simplified Laplace approximation (Baio, 2013).

3.4.6. Priors Assignment for the Distributions of Parameters

To do with Bayesian inference, the choice of prior distribution is a vital issue as it represents the information that is available for the parameters of interest. The prior may be selected based on the previous study and from the knowledge of experts (informative) or if there is no such option, after all, one may select prior using non-informative techniques without seeing the data. For convenience purpose, there are few models with which the distribution of the posterior is naturally known based on the conjugate of prior parameter distribution and the distribution of the likelihood.

The latent Gaussian model of INLA application assumed that all the fixed effects follow the Gaussian distribution with mean zero and the small number of precision matrix Q. Hence, only the parameters in the precision matrix of the random effect need a prior which was considered as a hyper-parameter (Baio, 2013).

3.4.6.1. Priors Comparison for Robustness of the priors

With this study, to go over the effect of the prior assignment, the researchers were interested to fit the model under two different priors namely default priors in INLA of the latent Gaussian model with the Poisson distribution of the data and Penalized Complexity (PC) priors.

The default priors are the inbuilt priors of R-INLA packages of INLA function in which the researchers need not to further assign the other priors. It was widely used by different researchers (Bivand *et al*., 2015; Martins *et al*., 2013; Rue *et al*., 2009; Blangiardo *et al*., 2015). According to the study by Rue *et al*. (2009) (who was the developer of R-INLA packages and INLA function), these default priors were considered as weak informative priors that were checked under different conditions before officially encoded under the R-INLA packages. Thus, all the parameters are assumed to follow Gaussian distribution. If the observation is assumed to follow the Poisson distribution, for intercept, INLA assign zero for both mean and precision; i.e. Normal(0,0) and all the fixed parameters are assigned mean zero and precision 0.001; meaning that they have Normal(0,0.001) priors. Whereas the random effect (district for our case) is also Gaussian with mean zero and precision parameter. Finally, the precision parameter in the random effects is also assigned to other distribution of log gamma which is log-gamma(1, 0.00001) (Rue *et al*., 2009; Blangiardo *et al*., 2015).

The other and very applicable recent priors are the one called Penalized Complexity prior (PC) priors. It was developed by Simpson *et al*. (2014) and was an informative prior. PC priors are general enough to be used in realistically complex statistical models and are straightforward enough to be used by general practitioners. Using only weak information, PC priors represent a unified prior specification with a clear meaning and interpretation. With this type of priors, researchers were agreed with its advantages of controlling the heterogeneity in random effect as it defined with the results of the standard deviation of residuals in the fixed effect.

In the PC priors, the distribution of intercept and all fixed effects are same to that of the value in the default priors. However, after the random effect was assigned to follow $Normal(0, Q)$. Thus, the precision value Q can be determined as:

$$\pi\left(Q^{-1/2}\right) = \frac{\lambda}{2} Q^{-3/2} \exp\left(-\lambda Q^{-\frac{1}{2}}\right), \quad Q > 0, \lambda > 0 \qquad [3.13]$$

This also has the form of exponential distribution of standard deviation with λ that determines the magnitude of the penalty for deviating from the base model. Here the idea is to specify (U, α)

that requires to determine λ. So that $Prob\left({}^{1}/_{\sqrt{Q}} > \ U\right) = \alpha$. After some mathematical arrangement: $\lambda = \ln(\alpha)/_{U}$.

Finally, the marginal standard deviation of random effect, after type-2 Gumbel distribution for Q is integrated out, is about 0.3*U when $\alpha = 0.01$. U is the standard deviation of residuals of the data (TB cases) (Simpson *et al*., 2014; Simpson *et al*., 2017). The proof and derivation of PC priors were found in the Appendix 3.

3.4.7. Posterior Distribution

A great advantage of working in a Bayesian framework is the availability of the entire posterior probability distribution for the parameter(s) of interest. Obviously, it is always possible and useful to summarize it through some suitable synthetic indicators.

In a Bayesian model, we generally want posterior distributions for our models (the distribution of the parameters given the data), or predictive posterior distributions (for prediction/forecasting - the distribution of new values given the observed ones) (Kruschke, 2008).

Finally, we obtained a posterior distribution for the parameter for which we can provide summary statistics (median, median, or mode) and quantiles to directly obtain credible intervals. The summary statistic typically used is the posterior mean, which, for a hypothetical continuous parameter of interestθ, is:

$$E(\theta/y) = \int_{\theta\in\vartheta} \theta P(\theta)\, d\theta \qquad [3.14]$$

Where ϑ all the possibilities that the variable θ can assume and the integral becomes summation if the value of θ is assumed to be discrete. Moreover, it is also possible to determine the indicators of which divide the probability in a very convenient way. Thus, the posterior median of $\theta_{0.5}$ is defined as the value which divides the probability distribution into two equal halves and can be determined as:

$P(\theta \leq \theta_{0.5}/y) = 0.5$ and similarly $P(\theta \geq \theta_{0.5}/y) = 0.5$.

While the 95% credibility interval (CI) is defined as the pair of θ values ($\theta_{0.025}$ and $\theta_{0.975}$) so that

$P(\theta \leq \theta_{0.025}/y) = 0.025$ and $P(\theta \geq \theta_{0.025}/y) = 0.025$.

However, the interpretation of credibility interval is completely different from that of the confidence interval. In frequentist approach, the $(100 - \alpha)\%$ confidence interval suggests that if we could repeat the same experiment, under the same conditions, for a large number *M* of times, then the real value of θ would fall out of the intervals only $\alpha\%$ of the times. This convoluted statement is not equivalent to asserting that the probability of θ lying within the confidence interval is$(100 - \alpha)\%$, since the parameter is considered a fixed, unknown value, and it is not a random variable characterized by a probability distribution. Conversely, within the Bayesian approach, a credibility interval explicitly indicates the posterior probability that θ lies within its boundaries; $P(\theta \epsilon CI/y)$; this is made possible by the fact that the parameter of interest is associated with a probability distribution, so that we can make probabilistic statements and take the underlying uncertainty into account (Rue *et al*., 2009; Blangiardo *et al*., 2015).

3.5. R-INLA Packages

R-INLA is the R package to implement approximate Bayesian inference using the INLA approach (Rue *et al*., 2009). In order to fulfill this aim, the r-INLA package (http://www.r-inla.org) was created as an R interface to the INLA program, which is itself written in C. The syntax for the r-INLA package is based on the inbuilt glm function in R, which highlights the effectiveness of the INLA method as a general solver for generalized linear (mixed) models. The key to the computational efficiency of the r-INLA program is that it is based on GMRFLib, a C library written by H°avard Rue for performing efficient computations on Gaussian Markov random fields. As such, r-INLA is particularly effective when the latent Gaussian field has the Markov property. This covers the case of spline smoothing (in any dimension), as well as conditional autoregressive models and some Matern random fields (Lindgren *et al*., 2011).

3.6. Bayesian Model Checking and Selections

An important aspect of Bayesian modeling is the assessment of its plausibility and checking which model better fit of the dataset. Two major aspects have been considered here. The first is model criticism, which can be get defined as the evaluation of which variables have included in the model, which assumptions to make on the parameters (e.g., exchangeability, independence, etc.), which prior distribution to assign on parameters and hyper-parameters. Are they plausible? Do they provide reasonable posterior inference? The second was called model selection which emphases on finding the best fit for the data in hand and compares model differing for the variables included the assumptions on parameters and likelihood and the prior distribution on parameters and hyper-parameters (Ferkingstad *et al*., 2017)

With this study, three different models were compared to identify the best model that fit the data well. This comparison was done in two different sessions; meaning that the first model including only fixed effects was compared with other models which has an additional random effect. The comparison was under the assumption of default priors discussed under sub-sub-title 3.4.8.1 above. The comparison has advantages so that to whether there may TB distribution differences across the district. Moreover, since the selection of priors was the crucial issues in the Bayesian framework, the researcher further compared the full model under two different priors assignment called default priors and penalized complexity priors.

According to Blangiardo *et al*. (2015), it would be possible to perform sensitivity analysis setting up a joint distribution of both model criticism and selection which is a mixture of all models to be checked. However, practically it is not such feasible and two different approaches are commonly used: the first is based on the predictive distribution and the second uses functions of the deviance. The function of deviance is also not theoretically attractive for model selection; thus the researcher was preferred to wrapping widely applicable information criteria instead of DIC after showing how it was efficient and theoretically attractive over it.

3.6.1. Predictive Distribution Methods

This method is based on the assumption of classifying the observation says y into two groups, so that $y = (y_{f,}\, y_c)$, where y_f is used to fitting the model and to estimate the posterior distribution of the parameters and y_c is used to perform model criticism. The idea on how to classify the sample data and techniques undergone for the application of model selection and criticism for predictive distribution can be seen in two major ways called cross-validation and posterior predictive check(Piironen *et al*., 2017).

3.6.1.1. Cross-Validation

Once the data has been splitting into the two groups, the posterior distribution for the parameters have the form of $p(x/y_f)$; x is a vector including all the parameters. This technique is called leave one out cross-validation, which assumes $y_f = y_{-i}$ and $y_c = y_i$. To evaluate the goodness of the model in this perspective, the Conditional Predictive Ordinate (CPO) and the Probability Integral Transform (PIT) are the two indices applicable with this method. The CPO facilitates the computation of the cross-validated log-score for model choice and PIT histograms can be computed to assess calibration of out-of-sample predictions (Czado *et al.*, 2009; Vehtari *et al*., 2016).

According to Czado *et al*. (2009), numerical problems may occur when the CPO and PIT indexes are computed. To this regards, the function in the R-INLA has provided automatically a sign for failure vector which contains a 0 or 1 value for each observation. The value equal to 1 in the output indicates that for the corresponding observation, the predictive measures are not reliable due to some problems in the calculation (Schrödle, 2011).

3.6.1.2. Posterior Predictive check

The posterior predictive checks are working based on the assumption that $y_c = y_f = y$; so all the observations are used for the model estimate and checking. In particular two quantities are of interest called posterior predictive distribution and posterior predictive p-value has been used here. Using the proved formula of Gelman *et al*. (2013), the value of posterior predictive P-value

near 0 or 1 indicates that the model fails to fit the data and it should be reconsidered. Unusually small values of indicate observations that come from the tails of the assumed distribution and can be classified as outliers. If this happens for many values, this suggests that the model is not adequate for the data in hand.

3.6.2. Watanabe Akaike Information Criteria

The most commonly used measure of model fit based on the deviance for Bayesian models is the deviance information criterion (DIC), proposed by (Spiegelhalter *et al*., 2002). It is a generalization of the Akaike information criterion (AIC), developed especially for Bayesian model comparison.

Though DIC has gained popularity in recent years, in part through its implementation in the graphical modeling package BUGS (Spiegelhalter *et al*., 2002), but it is known to have some problems, which arise in part from not being fully Bayesian in that it is based on a point estimate (Van , 2005).

WAIC (the widely applicable or Watanabe-Akaike information criterion) is a promising alternative and can be viewed as an improvement on the DIC for Bayesian models (Watanabe, 2010). Compared to DIC, WAIC has the desirable property of averaging over the posterior distribution rather than conditioning on a point estimate and does not rely on posterior means of parameters. This is especially relevant in a predictive context, as WAIC is evaluating the reductions that are actually being used for new data in a Bayesian context. WAIC works also in the singular models and this is particularly helpful for models with hierarchical and mixture structures in which the number of parameters increases with sample size and where point estimates often do not make sense.

WAIC can be easily determined analytically as well. The first step is to write the predictive density for each data point; $p_{post}(y_i) = N(\frac{y_i}{\bar{y},}1 + \frac{1}{n})$ and summing the terms for the data points, we get,

$$\sum_{i=1}^{n} log p_{post}(y_i) = -\frac{n}{n}\log\left(1 + \frac{1}{n}\right) - \frac{1}{2}\frac{n(n-1)}{n+1}s_y^2 \qquad [3.15]$$

Next, we determine the two forms of the effective number of parameters. The first effective number of the parameter can be calculated with the formula:

$$pWAIC1 = \frac{n-1}{n+1} s_y^2 + 1 - nlog(1 + \frac{1}{n}) \quad [3.16]$$

To evaluate the next effective numbers of the parameter, we based on the variance of the posterior distribution and averaging over it. Then, the effective parameter can be determined as:

$$pWAIC2 = \frac{n-1}{n} s_y^2 + \frac{1}{2n} \quad [3.17]$$

Then from the combination of equation 3.10, 3.11 and 3.12, the WAIC is formulated as:

$$WAIC = -2 \sum_{i=1}^{n} logp_{post}(y_i) + 2pWAIC \quad [3.18]$$

Generally, the empirical and theoretical reason based formulation of WAIC makes it more acceptable over that of DIC. Especially, since it used the point-wise predictive density averaging of each term over the entire posterior distribution rather than conditional on a point estimate, its acceptability and efficiency in comparing the model is quite reasonable (Gelman *et al*., 2013; Piironen *et al*., 2017).

3.7. Ethical Consideration

The Research Ethics Review Board of Jimma University has provided an ethical clearance for the study. The data was taken from Jimma zone health office, and the formal cooperation letter was written from college of natural science research office to the Jimma zone health office where data was obtained. The study conducted without individual informed consent; because it relied on retrospective data. The one year computer based recorded data was obtained with their corresponding covariates.

CHAPTER FOUR

4. RESULTS AND DISCUSSION

4.1. Results

Under this section of data analysis, the researcher tried to answer the basic research questions and attained to address the objectives by modeling the data with the appropriate model fit. In order to further go for the model, we have started with the simplest frequency table which has the power to intend the appropriate candidate model.

Thus, using the concept of INLA of the Bayesian framework, the results of the models with different fixed and random parameters considering the assignment of priors have been discussed stepwise here below. The results obtained from the different model of this study were compared by using standard statistical tools of model selection and comparison so that to filter out the relative best model in approximating the posterior marginal well.

4.1.1. Descriptive Data Analysis

Table 4.1 presents the counts and percentages of TB patients in each district of Jimma-zone. It is indicated that, without considering the effect of sex and ages, Nono bench district accounted minimum (2%) TB cases, whereas Seka chokorsa recorded to have the highest (12%). The numbers of male cases in each district were greater than those of females, except for districts of Agaro, Gomma, Limmu Seka, and Kersa.

It is also indicated that age class of the two edges of extreme (0-14 and >54) have lower TB cases which found to be 10% and 14% of the total cases; whereas the two middle class ages (15-34 and 35-54) were relatively more affected group; which is 30% and 46%. While the middle age of each district was developed more TB cases, people of aged 35-54 with the cases in Setema district have been recorded to have a higher number of cases (59%) in comparative of infected people of other districts of the same age group. However, with this same age category, in Boter district, only 22% of the cases were grouped under age 35-54. There were 1446 males and 1260 females of TB infected peoples in Jimma zone. Specifically, Mencho district has largest number

(64%) of males with TB cases considering the number of TB cases in the district as compared to other districts; whereas Agaro district has lowest (40%) number of Males infected with TB cases.

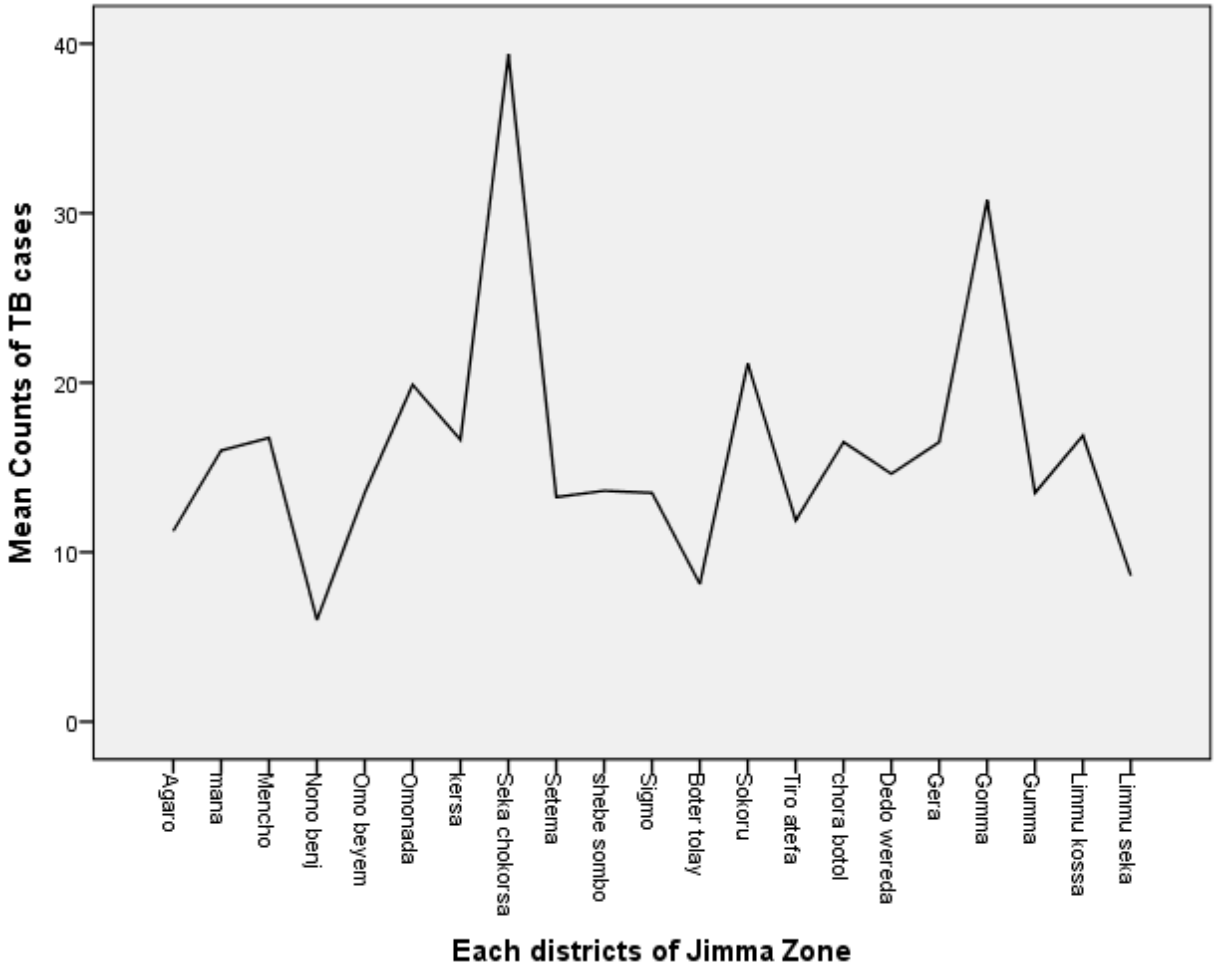
Table 4.1: Summary of TB counts of each district by considering Sex and Age

Districts	Tuberculosis Cases By Sex Counts (%)		Tuberculosis Cases by Ages Counts(Percentage)				TB per each District (%)
	Female	Male	0-14	15-34	35-54	>54	
Agaro	54(60)	36(40)	18(20)	32(36)	27(30)	13(14)	90(3)
Boter Tolay	31(48)	34(52)	9(14)	33(50)	14(22)	9(14)	65(2)
Chora Botol	52(39)	80(61)	25(19)	49(37)	35(27)	23(17)	132(5)
Dedo	57(49)	60(51)	6(5)	40(35)	59(50)	12(10)	117(4)
Gera	63(48)	69(52)	11(8)	59(45)	49(37)	13(10)	132(5)
Gomma	125(51)	121(49)	22(9)	84(34)	117(48)	23(9)	246(9)
Gumma	47(44)	61(56)	9(8)	26(24)	54(50)	19(18)	108(4)
Limmu Kossa	67(50)	68(50)	17(13)	36(27)	60(44)	22(16)	135(5)
Limmu Seka	38(55)	31(45)	12(17)	10(15)	33(48)	14(20)	69(3)
Mana	62(48)	66(52)	17(13)	35(27)	61(48)	15(12)	128(5)
Mencho	48(36)	86(64)	14(10)	48(36)	55(41)	17(13)	134(5)
Nono Bench	22(46)	26(54)	9(19)	12(25)	19(40)	8(17)	48(2)
Omo Beyem	46(43)	62(57)	7(6)	21(19)	60(56)	20(19)	108(4)
Omo Nada	75(47)	84(53)	20(12)	38(24)	71(45)	30(19)	159(6)
Kersa	73(55)	60(45)	10(8)	41(31)	66(49)	16(12)	133(5)
Seka Chokorsa	156(50)	159(50)	29(9)	87(28)	156(50)	43(13)	315(12)
Setema	41(39)	65(61)	8(8)	32(30)	63(59)	3(3)	106(4)
Shebe Sombo	46(42)	63(58)	7(6)	33(30)	53(49)	16(15)	109(4)
Sigmo	47(44)	61(56)	5(5)	24(22)	63(58)	16(15)	108(4)
Sokoru	72(43)	97(57)	20(12)	44(26)	62(37)	43(25)	169(6)
Tiro Itefa	38(40)	57(60)	5(5)	25(26)	57(60)	8(9)	95(3)
Total TB cases at Zone level	1260(47)	1446(53)	280(10)	809(30)	1234(46)	383(14)	2706(100)

The graphical presentation of Figure 4.1 was the results of the total counts of all forms of TB cases in each district of Jimma zone. It was supported by the results obtained in Table 1 above that Seka Chokorsa district has the highest number of TB cases as compared to the other districts and Nono Bench district seem to have fewer counts of the cases.

The graph clearly showed that the counts of the cases were varied from one district to the other and leads to looking in a way that how to handle the heterogeneity in such difference. It also still empowers the researcher to bear in account whether the number of population of each district may also have effect for the variability of the cases across the districts.

Fig 4.1: Counts of all forms of TB cases in each district of Jimma zone



4.1.2. Model Based Data Analysis

The outcome in Table 4.2 below is the summary results of LGM of Poisson distributional assumption for TB cases with fixed effects only model under default priors. For the convenience and get the easy understanding of the interpretation, the researcher has interpreted the exponentiation value of the coefficients. The intercept exp(0.723)=2.06 represents the average rate of yearly TB cases in Jimma zone when Age=0-14 and the effect of HIV co-infected and population density were getting zero. The model indicated that all covariates except sex were found to have a significant effect. Thus, sex as a factor of determining the TB cases was not significant under this model; since the credible interval (CI) (-0.012, 0.144) has included zero.

The coefficient value of the age categories of 15-34 was exp(1.138)=3.121 with (2.678, 3.536) CI. This is interpreted as the yearly incidence rates of TB cases in Jimma zone aged 15-34 was 3.121 times greater than those of age 0-14 holding the other factors at constant. Similarly, the coefficients of age 35-54 and >54 can be written as exp (1.679)=5.36 and exp(0.914)=2.494, and meaning that compared to those aged 0-14, the yearly incidence rates of TB cases were 5.36 and 2.94 times greater for age 35-54 and >54 respectively.

The value exp(0.098)=1.103 is for a unit increase in the number of HIV infected who screened for the TB cases, the yearly incidence rates of all forms of TB was increased by 1.103. In the same way, exp(0.005)=1.005 represents the yearly increase in TB incidence rate for a one unit increase in population density.

Table 4.2: Summary results of LGM for fixed effects only model with default priors

Fixed effects	Post. Estimate	St.dev	Media	95% CI	Mode	Kld
Intercept	0.723	0.138	0.724	(0.450, 0.992)	0.725	0
Sex(Male)	0.066	0.040	0.066	(-0.012, 0.144)	0.066	0
Age(15-34)	1.138	0.079	1.138	(0.985, 1.263)	1.137	0
Age(35-54)	1.679	0.088	1.678	(1.508, 1.852)	1.677	0
Age(>54)	0.914	0.107	0.913	(0.705, 1.123)	0.913	0
HIV co infect	0.098	0.009	0.098	(0.080, 0.115)	0.098	0
Pop. Density	0.005	0.001	0.005	(0.004, 0.007)	0.005	0

The summary presented in Table 4.3 below was the results of LGM with default prior under the consideration of generalized linear mixed model. Here the effect of the district was also included as the random effect so that to handle the variability from one district to the other. With this model, all the covariates were found to be significant since all the CI has not included zero.

The researcher preferred to take advantage of interpreting the exponentiating results. The intercept exp(1.039)=2.826 was the average yearly incidence rates of TB in Jimma zone when sex=female, age=0-14, and holding HIV co-infected and population density at constant.

The coefficients value of male patients was exp(0.1075)=1.113. This is to mean that the yearly TB incidence rates of the male were 1.113 times greater than those of female for the same cases. The same interpretation was drawn for the age factor. The exponentiated value of age 15-34 was exp(1.144)=3.139 and meaning that those aged 15-34 were 3.139 times greater than patients aged 0-15 in getting yearly TB incidence rates. Similarly, the yearly TB incidence rates of those aged 35-54 and >54 were 5.209 and 2.05 times greater than people aged 0-14 respectively.

For a one unit increase in the number of HIV infected who screened for the TB cases, the yearly incidence rates of all forms of TB increased by 1.04 (4%). And for a one unit increase in population density, the yearly incidence rates of TB were increased by 1.0036.

Table 4.3: Summary results of the LGM model with both fixed and random effects with default priors

Fixed effects	Post. Estimate	St.dev	Media	95% CI	Mode	kld
Results of Fixed Effects						
Intercept	1.0392	0.2467	1.0373	(0.5601, 1.5285)	1.0334	0
Sex(Male)	0.1075	0.0398	0.1075	(0.0293, 0.1857)	0.1075	0
Age(15-34)	1.1440	0.0917	1.1437	(0.9651, 1.3248)	1.1430	0
Age(35-54)	1.6503	0.1175	1.6501	(1.4200, 1.8812)	1.6499	0
Age(>54)	0.7180	0.1551	0.7184	(0.4123, 1.0214)	0.7192	0
HIV co infect	0.0435	0.0105	0.0435	(0.0227, 0.0640)	0.0436	0
Pop. Density	0.0036	0.0012	0.0036	(0.0013, 0.0058)	0.0036	0
Results of Random Effects						
Precision of Districts	11.71	4.489	12.09	(5.7710, 2 3.240)	10.88	-

The summary data in table 4.4 below was the outcome of Poisson distributional assumption of Tb cases under LGM which include both fixed and random effects with penalized complexity priors. The intercept exp(1.0703)=2.916. When the sex=female, age=0-14 and covariates HIV co-infection and population density were held constant, the average incidence rates of TB in Jimma zone was found to be 2.916. This is because the intercept was interpreted under the reference categories of categorical covariates and assumed when the effect of continues variables were zero.

The incidence rate of TB with the male was 1.114 times greater than those of female. This is to mean that around 11.4% more diseases with males. Each category of ages was also significant for the occurrences TB cases. Compared to those aged 0-14, people with age 15-34 developed TB incidence rates by 3.12 times more. In the same fashion, the incidence rates of people aged 35-54 was 5.15 times greater than those of aged 0-14 and those in the age interval of >54 were 2.01 times more likely to have TB incidence rates than those ranged in the age 0-14.

For a one unit increase in the HIV co-infected people, the TB incidence rate was increased by 1.044(4.4%). This result has an implication that the number of HIV co-infected in this study seems not to such signs in determining the relative risk of TB cases. On the other hand, the incidence rate of TB was increased by 1.0034(0.34%) as population density was increased by one unit. Even though the population density was found to be significant, under this model, the coefficient value indicated that this variable was not such potential in determining the occurrence of TB cases for this study.

The Kullback-Leibler divergence (KLD) describes the difference between the standard Gaussian and the under-used Simplified Laplace Approximation (SLA). Therefore, with this model, since the values of KLD irrespective of all covariates were zero, the researcher can generalize that the marginal posterior densities were well approximated by the Normal distribution. Thus, the under-used SLA which is the default approximation method in INLA function, in determining the densities of posterior marginal was defined as having good (small) error rate and no need to use the more computationally intensive technique full Laplace approximation.

Table 4.4: Summary results of LGM for a model including both fixed and random effects with Penalized Complexity priors

Results of Fixed Effects						
Fixed effects	Post. Estimate	St.dev	Media	95% CI	Mode	kld
Intercept	1.0703	0.2530	1.0684	(0.5788, 1.5720)	1.0645	0
Sex(Male)	0.1080	0.0398	0.1079	(0.0297, 0.1861)	0.1079	0
Age(15-34)	1.1377	0.0923	1.1373	(0.9575, 1.3197)	1.1366	0
Age(35-54)	1.6393	0.1190	1.6391	(1.4062, 1.8732)	1.6387	0
Age(>54)	0.6992	0.1574	0.6995	(0.3893, 1.0073)	0.7001	0
HIV co infect	0.0427	0.0105	0.0427	(0.0219, 0.0633)	0.0428	0
Pop. Density	0.0034	0.0012	0.0035	(0.0011, 0.0057)	0.0035	0
Results of Random Effects						
Prec. of district	12.51	3.911	11.946	(4.5260, 19.730)	8.858	-

The summarized result of table 4.5 below is the posterior marginal distribution of districts variation of tuberculosis with penalized complexity priors. The interpretation of posterior marginal for the precision of the random effect district in Table 5 is more general and bit difficult to interpret because it is on the scale of 1/variance. On the other hand, it is not possible to take the reciprocal of the (square-rooted) summaries to obtain information about the posterior distribution of the standard deviation, because the transformation is not linear.

Thus, the researcher preferred to compute posterior marginal with the scale of standard deviation. The average standard deviation of the variation of TB cases across districts was 0.294 with (0.207, 0.416) credible interval. Besides, the table 4.8 in Appendix 2 indicated that Seka Chokorsa has the highest incidence of TB cases and whereas Nono Benj district has less affected with TB cases. The diseases with the other were also varied in between. Generally, the appendix indicated that there is variation in TB cases across districts of Jimma zone.

Table 4.5: Posterior marginal distributions of standard deviation for random effect under PC priors

Posterior distribution	Mean	St.dev	Media	95% CI
St.dev for districts	0.294	0.053	0.287	(0.207, 0.416)

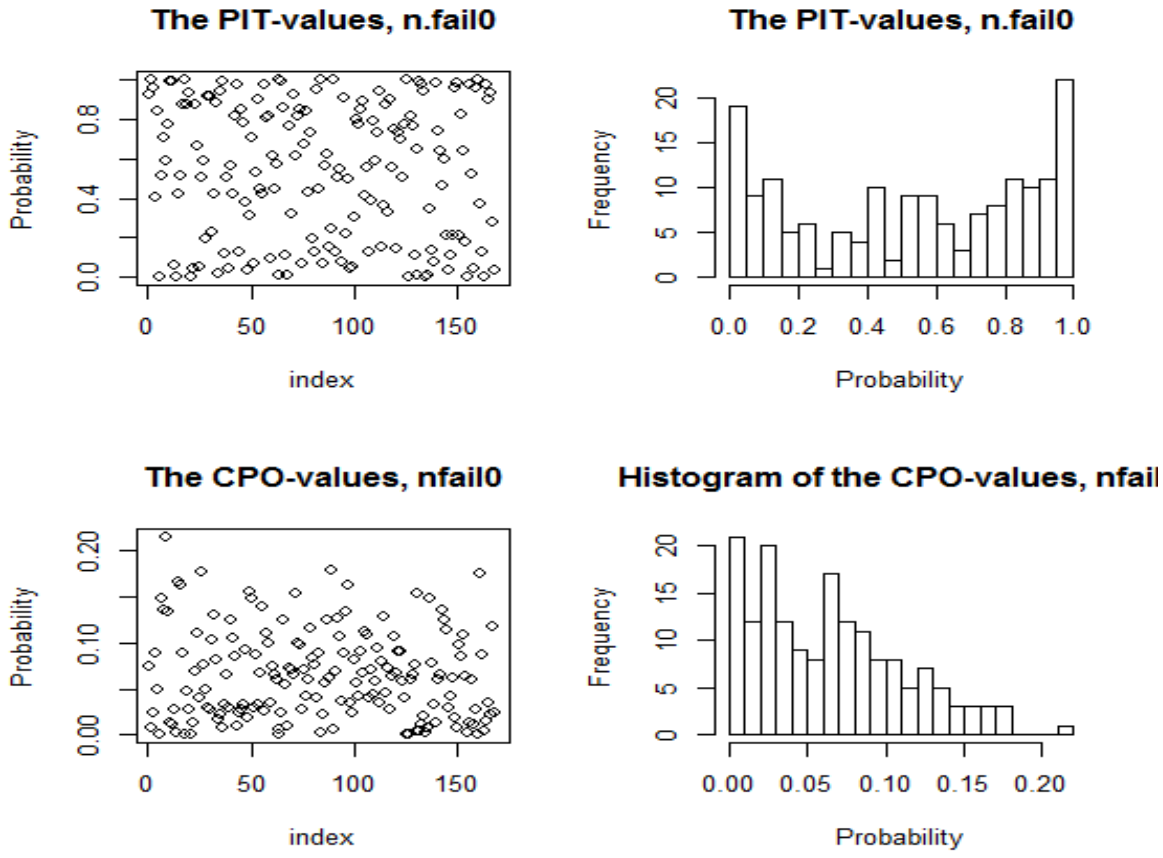
4.1.3. Model Checking

The graphical and numerical presentation of model checking methods has been tested so that to see the underlying assumption and distributional properties of the data.

In order to check whether the numerical problems may occur in the predictive measure when the CPO and PIT indexes are computed, the researcher used in-built R-INLA function which contains 0 or 1 value for each observation in that 1 is an indicator for the failure vector. Thus, for this study since the sum of failure values in CPO from the fitted model was found to be zero, the researcher has evidence to get concluded that there was no numerical problems occurred in predictive measure; since no failure has been detected.

The graphical presentation of Fig 4.2 was the scatter plot and histogram of cross-validated CPO and PIT with the possible indexes and probability. The purpose of plotting CPO was to see the surprising observation in which the extreme value is an indicator of the problem. The graph indicated that most of the observations have relatively the same distribution with very little deviated value which actually not expected to affect the model. Besides, the PIT is the measure of outliers. Hence, the idea behind having this plot was to see whether the predictive distribution matches the actual data, which is possible if the histogram shows uniform distribution. For this data, the histogram indicated that the PIT is almost uniformly distributed with very few deviated residual value and can get reasonable that the predictive distribution matches the actual data. Therefore, based on the plots of both CPO and PIT, the authors have evidence to say that predictive distribution was not significantly affected by surprising observation and extreme outliers (Gianluca, 2013).

Fig 4.2: The graphical presentation of CPO and PIT value for not fail values



4.1.4. Model Comparisons

In order to select the model which was relatively best fit the data, the researcher has intended to compare the model in two phases. The candidate models were:

Model 1: LGM with covariates of fixed effects only and default priors

Model 2: LGM including covariates of both fixed and random effects with default priors

Model 3: LGM including covariates of both fixed and random effects with PC priors

Thus, at the first stage of this model comparison, model 1 and model 2 have been compared in order to see whether the random effect has a significant effect or not. Then, to get valued on the robustness of the priors, model 2 and model 3 have been compared which in fact is to see the actual changes on the model as the priors on the parameters were changed. All the models were compared with the standard model comparison techniques, WAIC and other supportive criteria.

Table 4.6 is the summary results of WAIC, the effective number of parameters and number of equivalent replicates for the aforementioned three candidate models with the different number of parameters and/or under different priors.

At the first stage of model comparison, model 2 which is the model with covariates of both fixed and random effects under the assumption of default priors, have less WAIC (1105.25) as compared to model 1, WAIC (1300.79) which in fact includes only fixed effect with same priors. Thus, by the operational definition that the smaller the WAIC, the better the model fit, the researcher can prioritize model 2 as the relative better fit of the data under study. It also supports that including districts as the random effect has advantages in order to handle variation in incidence rates of TB across districts.

Once the model with both fixed and random effects under default priors was selected, the researcher was able to compare the same model under different priors which help to more ascertain the robustness of the priors. We did this because it helps to avoid the problem of model

fit due to bad priors and also used for further investigation as for whether the recent informative PC priors was efficient than the R-INAL inbuilt default priors or not.

Thus, based on the results of Table 4.6 below, the WAIC for model 3 which was 1104.27 was relatively smaller than that of model 2 which was 1105.25. However, different literature said that the models of the same parameters were considered to be significantly different, if their WAIC were at least 3-5 differences ((Ntirampeba *et al*., 2018; Spiegelhalter *et al*., 2002).

As the rule of model difference based on the WAIC's value difference is rule of thumb, other technical methods of model comparison have been used to see the clear difference between the models. Hence, the concept of the effective number of parameters and number of equivalent replicates were applicable here. Since the expected number of effective parameters is basically the number of independent parameters included in the model, the smaller is the better the model. This is because, at any stages of model comparison, the intention of the researcher was to come with the best model of the smaller parameter. Besides, the number of equivalent replicates is the result of sample size per effective number of parameters in the model and thus the smaller is an indication of poor fit.

But, the difference in both effective number of parameters and number of equivalent replicates for model 2 and 3 seems not such significant. However, since there were no clear-cut rules that judge to decide on the size of difference on such comparison techniques, the researcher has been forced to compare the models without valued the size of the difference in those model comparison methods.

Therefore, since the effective number of parameters of model 3 (24.74) was slightly smaller than that of model 2 (25.08), we can say that the candidate model with PC priors was relatively better in fitting the data well. Moreover, the number of equivalent replicates of model 3 was very slightly greater than that of model 2 and this also still has some information to decide that model with PC prior was comparatively better.

On the other regards, we extended our evidence from the perspective of standard errors by considering the results of standard deviation for the precision of the district (random effect) based on table 3 and 5 above. Recalling the direct proportion between standard error and standard deviation of the same sample size, we can say that the greater the standard deviation, the larger the standard error is. Hence, the standard deviation for the precision of districts in the model with default priors was 4.489 and that of the model with PC priors was 3.911, and considering the truth that the smaller the standard error, the efficient the model was, we still can support model with PC priors was better in fitting the data. Additionally, since the credible interval for the precision of districts of the model with PC priors (4.526, 19.73) was narrower than that of the model with default priors (5.771, 23.24), this may also assist the researcher to conclude that model with PC priors was relatively better in fitting the data.

Generally, considering the collective evidence detailed above and since PC priors are informative priors, we finally selected the LGM of Poisson distributional assumption of TB cases including covariates of both fixed and random effects with PC priors as the best model.

Table 4.6: Results of WAIC, effective number of parameters and number of equivalent replicates for the three candidate models

Candidate models	WAIC	Effective no. of parameters	No. of equivalent replicates
Model 1	1300.79	7.016	23.95
Model 2	1105.25	25.08	6.699
Model 3	1104.27	24.74	6.791

4.2. Discussions

The descriptive results of the study indicated that the number of males with TB cases (53%) was greater than the number of females with the same cases. These results were in line with WHO reports of 2017, which also presented as the number of males with TB cases was greater than females worldwide. Similarly, the number of TB counties of middle-aged people was greater than the two extreme categories of the ages and this also matches with the truth existed throughout the world (WHO, 2017) and different studies from Ethiopia were also persisted with this results (Gugssa *et al*., 2017; Hamusse, 2017). Besides, the descriptive summary clearly showed that the counts of the cases were varied from one district to the other and empowers the researcher to bear in account whether the number of population of each district may also have effect for the variability of the cases across the districts.

In order to assimilate the variation in the population size across districts with the corresponding TB cases, the offset variable was included in the correction factor. The offset in a sense means that the expected counts of TB cases in each district and especially used to correct the number of events (TB cases). The offset is the special type of variable that was widely applicable when the observation was assumed to have Poisson distribution with the known slope of 1 that helps to adjust the problem due to variation in population size from one district to the other.

Some of the researchers have been considered this adjustment under different dataset (Kipruto *et al*., 2015; Blangiardo *et al*., 2013). But, many studies that included geographical variation as random effect had missed these potential terms offset which used to weighted (corrected) the effect of miss many numbers of events and population size (Iddrisu *et al*., 2016; Musenge *et al*., 2013). Thus, our study has been filling the gap with the miss used of the offset variable.

At the first stage of the model fit, the LGM with Poisson distributional assumption of the observations has been fitted with covariates of fixed effects only under R-INLA inbuilt default priors. The variables age, HIV-co-infection, and population density were found to be significant. In order to check the effectiveness of simplified Laplace approximation method that applied in this model, the researcher considered the value of KLD in which the minimum the KLD is the less difference between the standard Gaussian and the Simplified Laplace Approximation. In our

case, since the value of KLD corresponding to all the variables was zero, the SLA was reasonably well in approximating the value which expected from standard Gaussian (Gianluca, 2013).

The LGM of Poisson distributional assumption of the observation which includes both fixed and random effects with default priors revealed that all the covariates have significant effects on the incidence rates of TB. The efficiency and relevance of the model were supported by the work of different researchers which in fact applied for the different dataset (Bivand *et al*., 2015; Martins *et al*., 2013; Rue *et al*., 2009). Moreover, the significance of the variables in this study was persistent with the finding of different researchers (Sreenivasulu *et al*., 2018; Roza *et al*., 2012, Couceiro *et al*., 2011). Since KLD result was found to be zero, the underused SLA has well approximated the standard Gaussian and no need to go for further intensive approximation methods like full Laplace approximation (Gianluca, 2013).

The other model called LGM of Poisson distributional assumption of observation with both the fixed and random effects under PC priors was applied. The same to model with default priors, all the covariates were found to be significant and the KLD values were also zero; meaning that SLA approximately had the same results with standard Gaussian. The developer of PC priors (Simpson *et al*., 2014) has been checked the effectiveness of the priors with simulated data and other few European researchers including Professor Havard Rue who is famous and developer of R-INLA program (Rue *et al*., 2009) had also exercised with the same simulated data. Thus, since the prior was developed in a very recent time and is informative, the authors were intended to apply for this actual data.

In order to make the model comparison, the researcher preferred the WAIC model comparison technique because of theoretical reasoning and inclusive advantages of the method detailed under sub-topic of 3.6.2 above and in the literature (Gelman *et al*., 2013). For clarity purpose, the three candidate models were compared in two phases; at the first stage, model 1 and 2, and at the second stage model 2 and 3 were compared so that to selected the best model which fitted the data well.

The results of WAIC indicated that model 2 which was the LGM of Poisson distributional assumption with both fixed and random effects under default priors was better than model 1 which was the same to model 2 except it includes only fixed covariates. Then, to check further for the effects of the priors, model 2 was compared with model 3 which was similar to model 2 except priors' assignments. Thus, since the WAIC of model 3 was smaller than that of model 2, the LGM of Poisson distributional assumption with both fixed and random effects considering PC priors was selected as the relative best model to fit the incidence rates of TB cases in Jimma zone. The advantage of comparison of models with different priors were certified under previous studies (Simpson *et al*., 2014; Schrödle, 2011; Simpson *et al*., 2017).

The random effect in the study was found to be significant and varied across the districts. This is an indicator that including districts as random effect here is advantageous so that to identify the district(s) with the highest TB cases. With this study, therefore, Seka chokorsa district was found to be the most severed districts. Previous studies also are also consistent with this result that TB cases were varied across the geographical regions (Kipruto *et al*., 2015; Iddrisu *et al*., 2016).

The CPO and PIT were used for model checking. Before further go for graphical model checking, the researcher intended to check whether the usual numerical problem occurred during the computation of CPO. Thus, since the sum of the number of failure in CPO was zero, no failure was detected and meaning that no numerical problem has occurred. The histogram and scatter plot of PIT indicated that the predictive residual based values were almost uniformly distributed with very few deviated outlier and we can get reasonable that the predictive distribution matches the actual data. Besides, the same graphs of CPO also indicated that most of the observed predictive values have the same distributional shape with the tolerance of surprising observation. Therefore, based on the plots of both CPO and PIT, the predictive values seem not significantly affected by surprising observation and extreme outliers (Gianluca, 2013; Rue *et al*., 2011; Martino, 2008).

CHAPTER FIVE

5. CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusions

For better attainment of the model fit, three different candidate LGMs namely: Poisson distributional assumption of TB cases of fixed effects only with default priors, Poisson distributional assumption of TB cases of both fixed and random effects with default priors and Poisson distributional assumption of TB cases with both fixed and random effect under PC priors have been fitted. With all the candidate models, the fixed covariates were sex, age, HIV co-infection, population density and districts as the random effect. Thus, based upon the WAIC and other supportive model comparison technique, the LGM of Poisson distributional assumption of TB cases which includes both fixed and random effects with PC priors has been selected as the best model that fits the data well. All the covariates under the best model are found to be significant.

The Kullback-Leibler divergence which is the difference between the standard Gaussian and Simplified Laplace approximation (full Laplace approximation) is found to be zero corresponding to all the covariates of the best model and is an indicator that the posterior marginal is well approximated by the Normal distribution. Thus, having the computational advantages of SLA and its better approximation in the data, the researcher preferred not to use the more computationally intensive full Laplace approximation.

Finally, the model check has been assessed by using CPO and PIT methods. The graph of CPO indicated that the predictive distribution seems not significantly affected by surprising observation; meaning that the predictive distribution was consistent with the actual observation. Whereas the scatterplot and histogram of PIT indicated that the predictive distribution was less affected by outliers and the model fit was reasonably well.

5.2. Recommendations

Based on the findings of this study, the researcher recommended the following points for researchers, Jimma health office and individuals interested in any sub-work of this study.

1. All the covariates in this study are significant factors of TB cases. Thus, Jimma zone health office and other health sectors should have to focus on controlling TB cases with special focus to districts that have a high severity of the disease.
2. The posterior marginal of this study is totally determined with the methods of INLA which actually is very fast and has a less computational burden. However, the issue related to the better approximation of INLA over MCMC in determining the posterior marginal is not addressed here. Thus, the researchers should strongly recommend so that to compare the methods.
3. This thesis was limited to few variables recorded at the health office. Thus, researchers are recommended to include clinical diagnostic related variables.
4. Interested researchers are recommended to extend this work by including fully spatial covariates in order to map and identify the hot-spot areas. This is very flexible and widely applicable in INLA.

5.3. Future Works

INLA was designed for LGM that Gaussian distribution is assigned to all parameters. But, for the models other than LGM (for those have non-Gaussian distribution of priors), the application of INLA is still under-development. Thus, the authors and other researchers should have to do for further development of the methodological formulation and R-INLA codes (syntax). This methods will enhance the researchers so that to flex with non-Gaussian distribution of the parameters.

REFERENCES

Ali H, Zeynudin A, Mekonnen A, Abera S, Ali S. (2017) Smear positive pulmonary tuberculosis (PTB) prevalence amongst patients at Agaro teaching health center, South West Ethiopia.

Asemahagn, M. A., & Amsalu, G. (2018). Determinants of Sputum Smear Positivity among Tuberculosis Suspected Patients in Bahir Dar City , Northwest Ethiopia, 111–117.

Baio, G. (2013). An introduction to INLA with a comparison to JAGS Laplace ' s liberation army (?), (May).

Biruk, M., Yimam, B., Abrha, H., Biruk, S., & Amdie, F. Z. (2016). Treatment Outcomes of Tuberculosis and Associated Factors in an Ethiopian University Hospital. *Advances in Public Health*, *2016*, 1–9. https://doi.org/10.1155/2016/8504629

Bivand, R. S., Gomes-Rubio, V., & Rue, H. (2015). Spatial Data Analysis with R - INLA with Some Extensions. *Journal of Statistical Software*, *63*(20), 1–31. https://doi.org/http://dx.doi.org/10.18637/jss.v063.i20

Blangiardo, M., & Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. https://doi.org/10.1002/9781118950203

Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and Spatio-temporal models with R-INLA. *Spatial and Spatio-Temporal Epidemiology*, *4*, 33–49. https://doi.org/10.1016/j.sste.2012.12.001

Bolin, D. (2015). Lecture 1 : Introduction Practical information.

Chalovich, J. M., & Eisenberg, E. (2013). NIH Public Access. *Magn Reson Imaging*, *31*(3), 477–479. https://doi.org/10.1016/j.immuni.2010.12.017.Two-stage

Corbett, E. L., Watt, C. J., Walker, N., Maher, D., Williams, B. G., Raviglione, M. C., & Dye, C. (2003). The Growing Burden of Tuberculosis. *Archives of Internal Medicine*, *163*(9), 1009. https://doi.org/10.1001/archinte.163.9.1009

Czado, C., Gneiting, T., & Held, L. (2009). Predictive model assessment for count data. *Biometrics*, *65*(4), 1254–1261. https://doi.org/10.1111/j.1541-0420.2009.01191.x

Datiko, D. G., & Lindtjørn, B. (2009). Health extension workers improve tuberculosis case detection and treatment success in southern Ethiopia: A community randomized trial. *PLoS ONE*, *4*(5), 1–7. https://doi.org/10.1371/journal.pone.0005443

Deribew, A., Abebe, G., Apers, L., Abdissa, A., Deribe, F., Woldemichael, K., … Colebunders,

R. (2012). Prevalence of pulmonary TB and spoligotype pattern of Mycobacterium tuberculosis among TB suspects in a rural community in Southwest Ethiopia. *BMC Infectious Diseases*, *12*(1), 54. https://doi.org/10.1186/1471-2334-12-54

Ethiopian Federal Ministry of Health. (2015). Health Sector Transformation Plan (2015/16-2019/20), *20*(May), 1–118.

Ferkingstad, E., Held, L., & Rue, H. (2017). Fast and accurate Bayesian model criticism and conflict diagnostics using R-INLA, 1–22. https://doi.org/10.1002/sta4.163

Geirsson, O. P., & Hrafnkelsson, B. (2014). Computationally efficient spatial modeling of annual maximum 24 hour precipitation . An application to data, (May).

Gelman, A., & Hill, J. (2009). Data Analysis Using Regression and Multilevel/Hierarchical Models. *Journal of Statistical Software*, *30*(April), 1–5.

Gelman, A., & Hwang, J. (2013). Understanding predictive information criteria for Bayesian models ∗.

Getahun, H., Gunneberg, C., Granich, R., & Nunn, P. (2010). HIV Infection–Associated Tuberculosis: The Epidemiology and the Response. *Clinical Infectious Diseases*, *50*(s3), S201–S207. https://doi.org/10.1086/651492

Girum, T., Muktar, E., Lentiro, K., Wondiye, H., & Shewangizaw, M. (2018). Epidemiology of multidrug-resistant tuberculosis (MDR-TB) in Ethiopia: a systematic review and meta-analysis of the prevalence, determinants and treatment outcome. *Tropical Diseases, Travel Medicine and Vaccines*, *4*(1), 5. https://doi.org/10.1186/s40794-018-0065-5

Gugssa Boru, C., Shimels, T., & Bilal, A. I. (2017). Factors contributing to non-adherence with treatment among TB patients in Sodo Woreda, Gurage Zone, Southern Ethiopia: A qualitative study. *Journal of Infection and Public Health*, *10*(5), 527–533. https://doi.org/10.1016/j.jiph.2016.11.018

Hamusse, S. D., Teshome, D., Hussen, M. S., Demissie, M., & Lindtjørn, B. (2016). Primary and secondary anti-tuberculosis drug resistance in Hitossa District of Arsi Zone, Oromia Regional State, Central Ethiopia. *BMC Public Health*, *16*(1), 1–10. https://doi.org/10.1186/s12889-016-3210-y

Harling, G., & Castro, M. C. (2014). Health & Place A spatial analysis of social and economic determinants of tuberculosis in Brazil. *Health & Place*, *25*, 56–67. https://doi.org/10.1016/j.healthplace.2013.10.008

Hayward, S., Harding, R. M., Mcshane, H., & Tanner, R. (2018). Factors influencing the higher incidence of tuberculosis among migrants and ethnic minorities in the UK [version 1 ; referees : 1 approved , 1 approved with reservations] Referee Status :, (0). https://doi.org/10.12688/f1000research.14476.1

Heunis, J. C., Kigozi, N. G., Chikobvu, P., Botha, S., & Rensburg, H. C. J. D. Van. (2017). Risk factors for mortality in TB patients : a 10-year electronic record review in a South African province. *BMC Public Health*, 1–7. https://doi.org/10.1186/s12889-016-3972-2

Hosseini, F., Eidsvik, J., & Mohammadzadeh, M. (2011). Approximate Bayesian inference in spatial GLMM with skew normal latent variables. *Computational Statistics and Data Analysis*, *55*(4), 1791–1806. https://doi.org/10.1016/j.csda.2010.11.011

Iddrisu, A., & Amoako, Y. A. (2016). Spatial Modeling and Mapping of Tuberculosis Using Bayesian Hierarchical Approaches. *Open Journal of Statistics*, *6*(June), 482–513. https://doi.org/10.4236/ojs.2016.63043

Jacirema, M., Gonçalves, F., Carlos, A., Leon, P. De, Lúcia, M., & Penna, F. (2009). A multilevel analysis of tuberculosis- associated factors, *11*(6), 918–930.

Jaya, M., Padjadjaran, U., & Folmer, H. (2014). Conditional Autoregressive Model on Dengue Fever Disease Mapping In Conditional Autoregressive Model on Dengue Fever Disease Mapping In Bandung City, (November). https://doi.org/10.13140/2.1.2654.4961

Kebede, A. H., Alebachew Wagaw, Z., Tsegaye, F., Lemma, E., Abebe, A., Agonafir, M., … Onozaki, I. (2014). The first population-based national tuberculosis prevalence survey in Ethiopia, 2010-2011. *International Journal of Tuberculosis and Lung Disease*, *18*(6), 635–639. https://doi.org/10.5588/ijtld.13.0417

Kipruto, H., Kipruto, H., Mung, J., Ogila, K., Adem, a, Mwalili, S., … Sang, G. (2015). Spatial Temporal Modelling of Tuberculosis in Kenya Using Small Area Estimation Spatial Temporal Modelling of Tuberculosis in Kenya Using Small Area Estimation, *4*(October), 1216–1224.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning and Behavior*, *36*(3), 210–226. https://doi.org/10.3758/LB.36.3.210

Legido-Quigley, H., Montgomery, C. M., Khan, P., Atun, R., Fakoya, A., Getahun, H., & Grant, A. D. (2013). Integrating tuberculosis and HIV services in low- and middle-income countries: A systematic review. *Tropical Medicine and International Health*, *18*(2), 199–

211. https://doi.org/10.1111/tmi.12029

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series B*, *73*, 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Lönnroth, K., Migliori, G. B., Abubakar, I., D'Ambrosio, L., De Vries, G., Diel, R., … Raviglione, M. C. (2015). Towards tuberculosis elimination: An action framework for low-incidence countries. *European Respiratory Journal*, *45*(4), 928–952. https://doi.org/10.1183/09031936.00214014

Martinez, L., Shen, Y., Mupere, E., Kizza, A., Hill, P. C., & Whalen, C. C. (2017). Transmission of Mycobacterium Tuberculosis in Households and the Community: A Systematic Review and Meta-Analysis. *American Journal of Epidemiology*, *185*(12), 1327–1339. https://doi.org/10.1093/aje/kwx025

Martino, S. (2008). Implementing Approximate Bayesian Inference using Integrated Nested Laplace Approximation : a manual for the inla program, (2007), 1–72.

Moghaddam, H. T., Moghadam, Z. E., Khademi, G., Bahreini, A., & Saeidi, M. (2016). Tuberculosis: Past, Present and Future. *Int J Pediatr*, *4*(125), 1243–1255. https://doi.org/10.22038/IJP.2016.6266

Ntirampeba, D., Neema, I., & Kazembe, L. (2018). Modelling spatio-temporal patterns of disease for spatially misaligned data: An application on measles incidence data in Namibia from 2005-2014. *PLoS ONE*, *13*(8), 1–18. https://doi.org/10.1371/journal.pone.0201700

Nzabanita, J. (2012). *Estimation in Multivariate Linear Models with Linearly Structured Covariance Matrices*.

Octavianty, Toharudin, T., & Jaya, I. G. N. M. (2017). Geographically weighted poisson regression semiparametric on modeling of the number of tuberculosis cases (Case study: Bandung city). *AIP Conference Proceedings*, *1827*. https://doi.org/10.1063/1.4979438

Ojo, O. B., Lougue, S., & Woldegerima, W. A. (2017). Bayesian generalized linear mixed modeling of Tuberculosis using informative priors, 1–14. https://doi.org/10.1371/journal.pone.0172580

Opitz, T., Opitz, T., Gaussian, L., & Inla, A. (2016). Latent Gaussian modeling and INLA : A review with focus on space-time applications To cite this version : ´ e ´ Franc Journal de la

Soci et Latent Gaussian modeling and INLA : A review with focus on space-time applications.

Pahlavanzadeh, B., Nasehi, M., & Sekhavati, E. (2016). The examination of relationship between socioeconomic factors and number of tuberculosis using quantile regression model for count data in Iran 2010-2011.

Pedro, H. S. P., Coelho, A. G. V, Mansur, I. M., Chiou, A. C., Pereira, M. I. F., Belotti, N. C. U., … Chimara, E. (2017). Epidemiological Analysis of Pulmonary Tuberculosis in Heilongjiang Province China from 2008 to 2015. *International Journal of Mycobacteriology*, *6*(3), 239–245. https://doi.org/10.4103/ijmy.ijmy

Piironen, J., & Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735. https://doi.org/10.1007/s11222-016-9649-y

Randremanana, R., Richard, V., Rakotomanana, F., Sabatier, P., & Bicout, D. (2010). Bayesian mapping of pulmonary tuberculosis in Antananarivo,Madagascar. *BMC Infectious Diseases*, *10*(1), 21. https://doi.org/10.1186/1471-2334-10-21

Repository, Z. O. (2011). Statistical analysis of spatio-temporal veterinary surveillance data : Applications of integrated nested Laplace approximations Statistical Analysis of Spatio-Temporal Veterinary Surveillance Data : Applications of Integrated Nested Laplace Approximations.

Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian Computing with INLA : A Review Ha. https://doi.org/10.1146/annurev-statistics-060116-054045

Rue, H., & Martino, S. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, 319–392.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approxima- tions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *71*(2001), 319–392.

Samuels, M. B., & Bell, N. K. (1990). Book Reviews Editor. *The Journal of Rural Health*, *6*(3), 328–330. https://doi.org/10.1111/j.1748-0361.1990.tb00671.x

Simpson, D. (2014). Penalising model component complexity : A principled , practical approach to constructing priors, (January 2016). https://doi.org/10.1214/16-STS576

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian Measures of Model Complexity anf Fit. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Sreenivasulu, T., Jahnavi, K., Sreenivasulu, T., & Med, J. A. (2018). A cross sectional study on factors affecting treatment outcome among TB patients, *5*(1), 175–178.

Srinivasan, R., & Venkatesan, P. (2014). Bayesian random effects model for disease mapping of relative risks, *5*(1), 23–31.

Statistical Decision Theory and Bayesian Analysis. (2015), (April).

Sturtz, S. (2008). *H. Rue, L. Held: Gaussian Markov random fields. Theory and applications*. *Metrika* (Vol. 67). https://doi.org/10.1007/s00184-007-0162-3

Sulis, G., Centis, R., Sotgiu, G., Ambrosio, L. D., Pontali, E., Spanevello, A., … Migliori, G. B. (2016). Recent developments in the diagnosis and management of tuberculosis, (June). https://doi.org/10.1038/npjpcrm.2016.78

Tonui, B., Mwalili, S., & Wanjoya, A. (2018). Spatio-Temporal Variation of HIV Infection in Kenya, 811–830. https://doi.org/10.4236/ojs.2018.85053

Van Der Linde, A. (2005). DIC in variable selection. *Statistica Neerlandica*, *59*(1), 45–56. https://doi.org/10.1111/j.1467-9574.2005.00278.x

Vehtari, A., & Gelman, A. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC ∗, (September), 1–28.

Volmink, J., & Garner, P. (2015). Directly observed therapy for treating tuberculosis. *Cochrane.Database.Syst.Rev.*, (5), DOI: 10.1002/14651858.CD003343.pub4. https://doi.org/10.1002/14651858.CD003343.pub4.www.cochranelibrary.com

WHO. (2014). Global tuberculosis report 2014 WHO Library Cataloguing-in-Publication Data. *World Health Organization*, 1–171.

WHO. (2016). Global Tuberculosis Report 2016. *Cdc 2016*, (Global TB Report 2016), 214. https://doi.org/ISBN 978 92 4 156539 4

WHO. (2017). Global Tuberculosis Report 2017. *World Health Organization*. https://doi.org/WHO/HTM/TB/2017.23

world health organization. (2015). Global Tuberculosis report 2015. *World Health Organisation*, *20*(1), 1–145. https://doi.org/10.1017/CBO9781107415324.004

Appendix 2: Some selective relevant tables and graphs.

Table 4.7: Posterior marginal distributions of standard deviation for random effect under default priors

Posterior distribution	Mean	St.dev	Media	95% CI
St.dev for districts	0.325	0.062	0.280	(0.225, 0.470)

Table 4.8: Empty model with R-INLA default priors

	Mean	St.dev	0.025quant	0.5quant	0.975quant	mode	kld
(Intercept)	2.6998	0.089	.52	2.7002	2.875	2.7009	1e-04
Prec. District	6.989	2.245	3.374	6.724	12.12	6.199	-
WAIC=2077.63							

Table 4.9: Individual model fit of all districts as random effect with PC priors

Districts	Mean	sd	0.025quant	0.5quant	0.975quant	mode	Kld
Agaro	-0.27	0.12	-0.52	-0.27	-0.04	-0.27	0
Mana	-0.05	0.12	-0.28	-0.05	-0.02	-0.05	0
Mencho	0.02	0.11	-0.19	-0.02	0.001	-0.02	0
Nono Benj	-0.56	0.16	-0.88	-0.55	-0.26	-0.54	0
Omo beyem	-0.24	0.12	-0.47	-0.24	0.00	-0.24	0
Omonada	0.17	0.11	0.05	0.17	0.39	0.17	0
Kersa	0.04	0.11	0.18	0.04	0.26	0.04	0
Seka chokorsa	0.82	0.10	0.63	0.82	1.01	0.81	0
Setema	-0.04	0.12	-0.27	-0.04	0.19	-0.03	0
shebe sombo	-0.03	0.12	-0.26	-0.03	-0.019	-0.03	0
Sigmo	0.01	0.12	-0.24	0.01	0.22	0.01	0
Boter tolay	-0.31	0.15	-0.60	-0.30	-0.03	-0.30	0
Sokoru	0.22	0.11	0.00	0.22	0.44	0.21	0
Tiro atefa	-0.15	0.12	-0.39	-0.15	-0.08	-0.15	0
chora botol	0.07	0.11	-0.14	0.07	0.29	0.07	0

Dedo wereda	-0.11	0.11	-0.34	-0.11	-0.01	-0.11	0
Gera	0.08	0.11	0.14	0.08	0.29	0.08	0
Gomma	0.52	0.11	0.31	0.51	0.73	0.51	0
Gumma	0.05	0.14	0.22	0.05	0.31	0.06	0
Limmu kossa	0.04	0.11	0.17	0.04	0.26	0.04	0
Limmu seka	-0.27	0.14	-0.56	-0.27	-0.12	-0.26	0

Fig 4.3: Marginal distribution of the fixed effects in final model

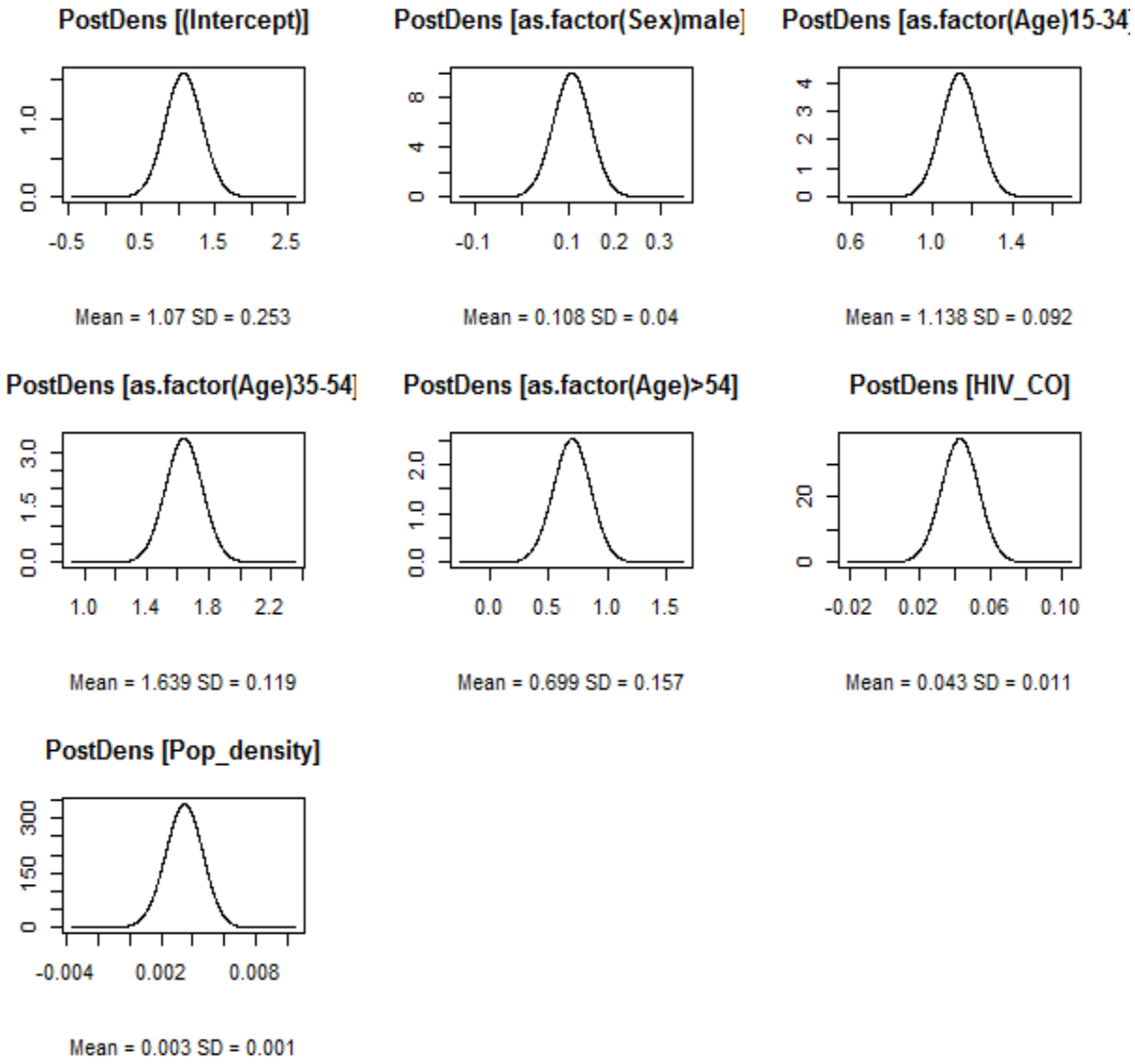


Fig 4.4: Histogram of posterior density for precision of district

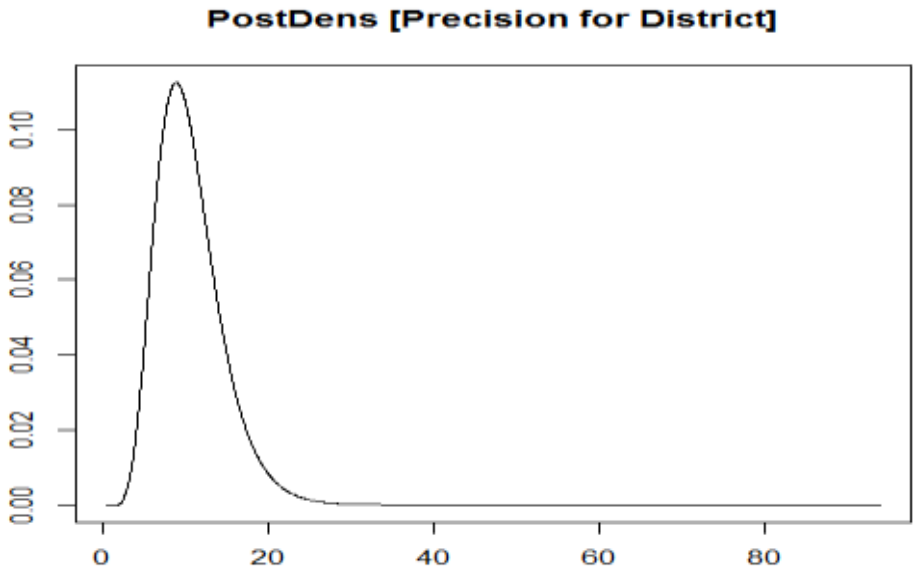


Fig 4.5: Posterior marginal distribution of st.dev for the random effects with PC priors

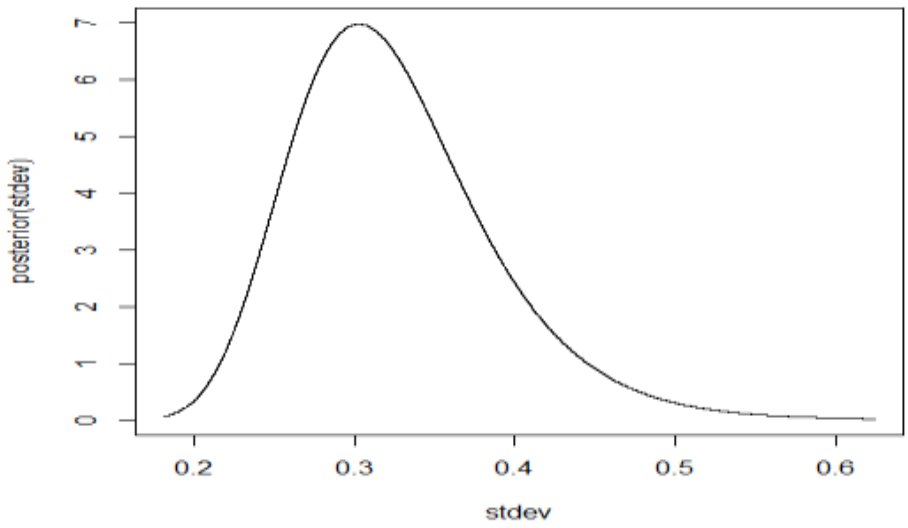


Fig 4.6: Linear predictor (above) and fitted values of linear predictor (below) with 95%CI

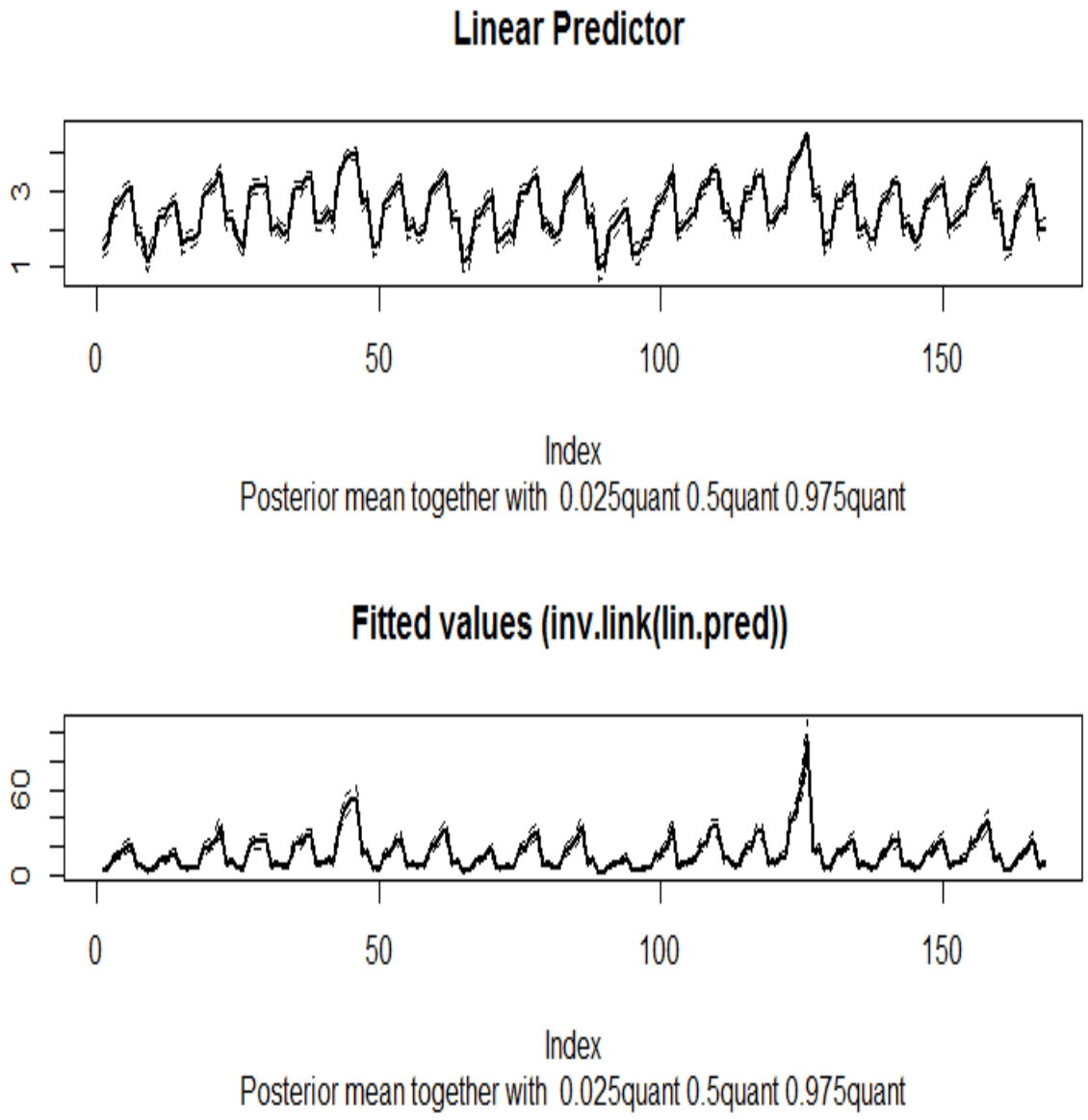


Fig 4.8: Scatterplot of the posterior mean for the predictive distributions against the observed values (left) and Histogram of the posterior predictive p-value (right).