SciencePG
Science Publishing Group

Analysis of TB Case Counts in Southwest Ethiopia Using Bayesian Hierarchical Approach of the Latent Gaussian Model

Endale Alemayehu[1, *], Reta Habtamu[2], Akalu Banbeta[2]

[1]Department of Statistics, College of Natural Sciences, Ambo University, Ambo, Ethiopia

[2]Department of Statistics, College of Natural Sciences, Jimma University, Jimma, Ethiopia

Email address:
endalestat@gmail.com (E. Alemayehu), habtamuretab@yahoo.com (R. Habtamu), akalubanbeta.stat@gmail.com (A. Banbeta)
*Corresponding author

Abstract: 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. Africa has shared around 25% of the incidence and specifically in Ethiopia around 82 thousand was caught by Tuberculosis. Methods: The study has been conducted in, south west Ethiopia, Jimma zone of entire districts and the data is basically secondary which is obtained from Jimma zone health office. The counts of Tuberculosis case counts 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 of the parameters of interest. Results: The Latent Gaussian Model (LGM) 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: 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.

Keywords: Tuberculosis, Bayesian Approach, LGM, INLA

1. Background

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 according to WHO [64-66].

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 [65], the most estimated number of TB cases is 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 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 [2, 65].

According to the [65, 66] report, Africa is not among the regions registered to have a declined in TB mortality rates. 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% [65].

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 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 [13, 14, 65].

1.1. Statements of the Problem

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 was gradually increased till 2016 [2, 13]

According to different studies, the Bayesian approach have empowered over the frequentist based on analysis using TB cases and simulation studies [17, 29, 47]. But, the Bayesian methods in those studies were based on MCMC sampling technique which has the burdensome of time-consuming, convergence problem and Monte Carlo error. This implies that analysis based on the MCMC technique could essentially affect the conclusions of researches.

In this research, we apply the latent Gaussian model with INLA methods under the framework Bayesian hierarchical paradigm. Based on this method, this 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.

1.2. Objective of the Study

1.2.1. General Objective

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

1.2.2. Specific Objectives

The research includes the following specific objectives:
1. To identify the predictors of TB cases and see the variation in the distribution of the cases across districts.
2. To compare the R-INLA's inbuilt default priors with the informative penalized complexity priors for robustness of the priors.
3. To fit the latent Gaussian model with INLA methods under the framework of Bayesian hierarchical paradigm.

1.3. 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 is an additional about the trend of TB prevalence to the previous studies. Furthermore, this research can assist for those who are interested to work in this area for the future. The result of this study will also be expected to be helpful for policy makers in TB concern agendas and strategies.

2. Methods

2.1. 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. 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.

2.2. 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 from September 2016 to August 2017 have been considered.

2.3. Variables of the Study

2.3.1. 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.

2.3.2. 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.

2.4. 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.

2.4.1. 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

[43]. Formally, the Latent Gaussian Model (LGM) can be written as:

$$\frac{y}{x}, \theta_i \sim \prod p\left(\frac{y_i}{\eta}, \theta_2\right) \text{ Likelihood} \qquad (1)$$

$$x/\theta_1 \sim p(x/\theta_1) = N(0, Q) \text{ Latent Field} \qquad (2)$$

$$\theta = [\theta_1, \theta_2]^{\mathrm{T}} \sim p(\theta) \text{ Hyper-priors} \qquad (3)$$

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

$$y \sim \prod_{i}^{N} p(y_i/\lambda_i)$$

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

2.4.2. 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 [50]. 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 [49]

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 (2). Generally, the marginal posterior distributions for each element of the parameter vector can be formulated as:

$$\pi\left(\frac{xi}{y}\right) = \int_\theta \pi\left(\frac{xi}{\theta}, y\right) \pi\left(\frac{\theta}{y}\right) d\theta \qquad (4)$$

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

$$\pi(\theta_j/y) = \int \pi(\theta/y) d\theta_{-j} \qquad (5)$$

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 $\pi(x_i/y)$ [6, 50].

2.4.3. Prior 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 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 [5, 6, 36, 50]. According to the study by [50] (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. 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) [6, 50].

The other and very applicable recent priors are the one called Penalized Complexity prior (PC) priors. It was developed by [52] 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 [52].

2.4.4. 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. Summary statistic typically used is the posterior mean, which, for a hypothetical continuous parameter of interest θ, is:

$$E(\theta/y) = \int_{\theta \in \vartheta} \theta P(\theta) d\theta \qquad (6)$$

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.

3. 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.

3.1. Descriptive Data Analysis

Without considering the effect of sex and ages, Nono bench district accounted minimum (2%) TB cases, whereas Sekachokorsa 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.

The two extreme age classes (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%.

The trend seen in the figure 1 below (in the figure legend) is also indicated the distribution of TB cases across all districts of Jimma zone.
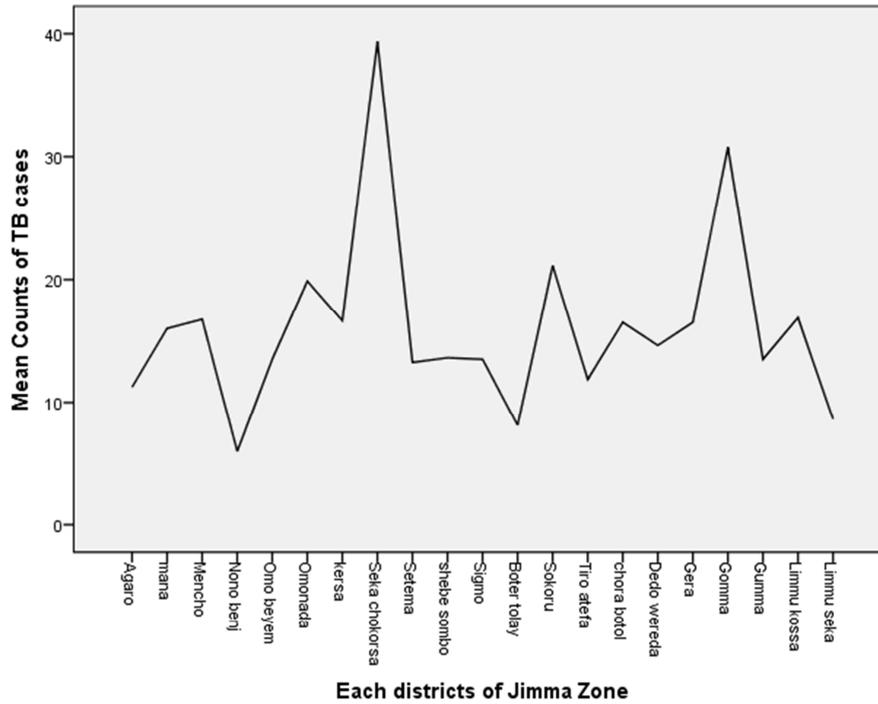

Figure 1. *Counts of all forms of TB cases in each district of Jimma zone.*

3.2. Model Based Data Analysis

The summary data in table 1 (in the table legend) 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.

Table 1. *Summary results of LGM for a model including both fixed and random effects with Penalized Complexity priors.*

Fixed effects	Post. Estimate	St. dev	Media	95% CI	Mode	kld
Results of Fixed Effects						
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 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.

The summarized result of table 3 below (in the table legend) 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 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. 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.

3.3. Model Comparison

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 2 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-INLA inbuilt default priors or not.

Thus, based on the results of Table 2 below (in the table legend), 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 [39, 53].

Table 2. 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

Table 3. 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)

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 tables 1 and 2 (in the table legend). 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.

4. Discussion

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 [65] and different studies from Ethiopia were also persisted with this results [21, 22]. 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 [6, 31]. 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 [27]. 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 [26].

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 [5, 36, 50]. Moreover, the significance of the variables in this study was persistent with the finding of different researchers [51, 55]. Since KLD result was, found to be zero, the under used SLA has well approximated the standard Gaussian and no need to go for further intensive approximation methods like full Laplace approximation [26].

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 [52] 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 [50] 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. 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 advantages of comparison of models with different priors were certified under previous studies [52, 57].

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, Sekachokorsa 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 [27, 31].

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 [37, 50].

5. 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. 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. 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.

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.

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.

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. Thus, the researchers should strongly recommend so that to compare the methods.

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.

Declaration

During conducting the study, the investigators have included the following declarations.

Abbreviations

AFB	Acid-Fast Bacilli
CPO	Conditional Predictive Ordinate
GLMM	Generalized Linear Mixed Model
GMRF	Gaussian Markov Random Field
HIV	Human Immunodeficiency Virus
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
SLA	Simple Laplace Approximation
TB	Tuberculosis
WAIC	Watanabe Akaike Information Criteria
WHO	World Health Organization

Ethics Approval and Consent to Participate

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.

Availability of Data and Material

The data and material are provided with the manuscript. There are no restrictions on the availability of data and R-code, and the investigators are willing to provide the code as well.

Competing Interests

The authors declare that they have no competing interest

Funding

Authors' Contributions

This study was designed and compiled by EA (MSc in Biostatistics) as the principal investigator. The development of the basic research questions, identifying the problems and selecting appropriate statistical models have been done by him. Edition of the overall progress of the work and compiling the R-command was supported by AB (PhD scholar) and RH (MSc in Mathematical and Statistical Modeling).

Acknowledgements

References

[1]   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.

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

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

[4]   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.

[5]   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.

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

[7]   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.

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

[9]   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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

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

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

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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

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

[51] 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.

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

[53] 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.

[54] 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.

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

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

[57] 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.

[58] 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.

[59] 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.

[60] 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.

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

[62] 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.

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

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

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

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