American Journal of Applied Mathematics and Statistics, 2024, Vol. 12, No. 3, 55-65
Available online at http://pubs.sciepub.com/ajams/12/3/3
Published by Science and Education Publishing
DOI:10.12691/ajams-12-3-3

Bayesian Spatial Modelling of Tuberculosis Incidence in Meru County, Kenya Using Conditional Autoregressive(CAR) and Poisson Lognormal(PLN) Models

Kithaka Gilbert Mucheri*, Peter Kinyua Gachoki, Kilai Mutua

Department of Pure and Applied Sciences, Kirinyaga University
*Corresponding author: giltexinfor@gmail.com

Received June 28, 2024; Revised July 30, 2024; Accepted August 05, 2024

Abstract Establishing the patterns of a disease or disease mapping is very important in disease control and prevention. The level of accuracy that is achieved at this stage determines the effectiveness of control measures to be developed. Disease mapping has been widely done using the frequentist approach which is limited in that it does not consider prior probability distribution of a phenomenon. This limitation leads to lower levels of accuracy and validity. This study proposed a Bayesian Approach for mapping tuberculosis incidence in Meru County, Kenya. Correlational research design was utilized to determine association between TB cases and geographical locations where the cases were positively identified. Secondary data from the Meru County Health Records was used for this study. Spatial autocorrelation was performed to determine patterns of TB incidence. The study applied Conditional Autoregressive (CAR) model and Poisson Lognormal (PLN) model under the Bayesian Approach to model TB incidence in order to determine spatial temporal trends. Parameter estimation for the models was done using GIBBs Sampling under Markov Chain Monte-Carlo (MCMC). The two models (PLN and CAR) were compared using Deviance Information Criteria (DIC) to determine the one that had a better fit. Morans's I statistic was -0.3150 ($p>0.05$) meaning that there was no spatial autocorrelation for TB incidence in Meru County. Model results further indicated that there was no spacial dependence for TB incidence in Meru County. Deviance Information Criterion (DIC) values obtained were 0.22541 for CAR model and 0.56723 for PLN model meaning that CAR model had outperformed the PLN model. The study concluded that CAR model is more effective for disease mapping since it incorporates information from neighboring regions directly into the model to increase accuracy of estimates. Therefore, the study recommended use of Bayesian modelling for disease mapping as it incorporates prior information to stabilize the parameter estimates.

Keywords: *tuberculosis, bayesian spatial models, spatial dependency, conditional autoregressive model, poisson lognormal model, poisson gamma model*

Cite This Article: Kithaka Gilbert Mucheri, Peter Kinyua Gachoki, and Kilai Mutua, "Bayesian Spatial Modelling of Tuberculosis Incidence in Meru County, Kenya Using Conditional Autoregressive(CAR) and Poisson Lognormal(PLN) Models." *American Journal of Applied Mathematics and Statistics*, vol. 12, no. 3 (2024): 55-65. doi: 10.12691/ajams-12-3-3.

1. Introduction

Tuberculosis (TB) remains a significant public health challenge globally especially in regions like Sub-Saharan Africa. According to a research by [1], high prevalence in the aforementioned region is occasioned by absence of preventive measures for TB as a result of poor economic situation. In US and other developed countries, Tuberculosis and other respiratory illnesses are taken very seriously. Whenever a case has been identified, the patients are isolated and treated to prevent further spread of the infection. In Kenya, TB is a major issue with particular prevalence in specific counties. According to a report by Center for Health Solutions (2020), the top ten counties with respect to TB cases in the year 2019 were Nairobi (12,425), Mombasa (4,225), Kiambu (3,702), Nakuru (3,636), Meru(3,420), Kisumu (2,933), Turkana (2,250), Machakos (2,223), Kakamega (2,154), Homabay (2,143) (Warren and Mwangi, 2017). One strategy that can be used to address the situation is identification of TB hotspots in order to develop tailor-made interventions. This can be best achieved through application of spatial modeling.

Spatial modelling is a form of disaggregation in which an area is divided into a number of similar units typically grid squares or polygons [2]. When performing spatial modelling, a set of spatially organized data is analyzed to establish statistical patterns for the sampled area [3].

Bayesian modelling refers to a statistical approach where prior knowledge is incorporated into a the model [4]. Bayesian spatial modelling is the application of Bayesian approach on spatially organized data to develop statistical inferences that are geographically represented [5]. Bayesian modelling involves combination of prior distribution and likelihood estimate to come up with a posterior distribution [6]. Bayesian modelling has been used in many researches especially in the medical field where high precision is required. Bayesian models are used in sensitive researches especially in the area of biomedical statistics where an error in the analysis and presentation of results can lead to wrong decisions and policies which in turn will pose a threat to public health [7].

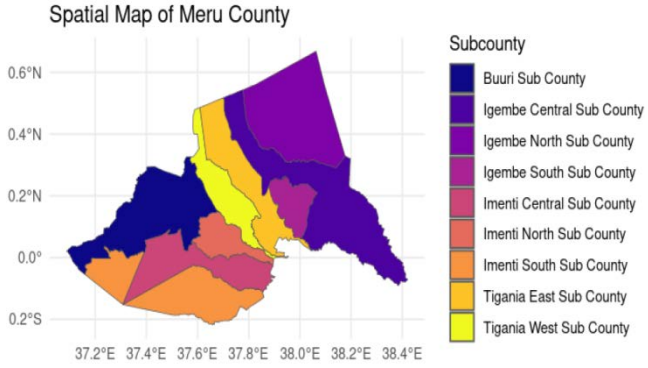

Figure 1. Study area for the research

Disease mapping is the statistical process that aims at achieving spatial correlations for a specified disease. The overall objective of disease mapping is to determine the geographical risk distribution for a disease and to make inferences on whether there exists areas of elevated or lowered risks within a specified geographic unit [8]. Mapping of disease was previously done using disease rates and relative risk data. The results obtained lacked stability meaning that there was likelihood of variations [5]. One of the models that have been applied in disease mapping is logistic regression. This is a statistical model that models the log-odds of an event in form of linear combination of one or more events [9]. It is a process of modelling probability of a discrete outcome given an input variable. Logistic regression can be binary, multinomial, or ordinal if the datasets for which its being applied are binary, nominal or ordinal levels respectively [10]. One major limitation of Logistic Regression as applied in disease mapping is that it fails to predict a continuous outcome. Secondly, logistic regression assumes linearity between predictor variable and response variable which may not always be the case with disease parameters. Finally, logistic regression may not be applicable where the sample size is small as this increases sampling error and can result to biased results [11].

General Linear Model (GLM) refers to application of multiple linear regressions for a continuous response variable given continuous or categorical predictors. It includes multiple linear regression, as well as ANOVA and ANCOVA In disease mapping, GLM has been used to determine if there exists significant association between a response variable and predictor variables. Predictor variables in the case of GLM disease mapping are geographically categorized so as to represent the spatial points [12]. The major limitation of GLM as applied in disease mapping is that it may not be able to handle very large or high-dimensional data sets, as they require more computational resources and may suffer from overfitting or multicollinearity.

Cluster modelling has been applied in disease mapping. This approach starts by developing data points in terms of clusters. The clusters are based on the categories that the researcher wishes to differentiate or associate [3]. Clustering can be classified as hard when each data point is linked to one cluster and soft when the output is the probability or likelihood of a data point to belong to each cluster [13]. The major limitation of cluster modelling is that it assumes a spherical shape with similar sizes and densities for each cluster which is not usually the case in a real world situation. The results obtained are thus not practically applicable [14]. The other limitation of cluster modelling is that it is sensitive to outliers and noise which can easily distort the results [15].

Decision trees have been applied in disease mapping. Decision trees are hierarchical structures that use tree-like models of decisions and their possible consequences [16]. Decision trees are used to model chance event outcomes using algorithms that contain conditional control statements. Main algorithms used in decision tree modelling are Iteration Dichotomizer 3(ID3), Classification and regression trees (CART), Chi-Square and reduction in variance. One limitation of decision tree modelling is that the probabilities obtained are just estimates which are subject to error. A small change in data can result into a major change in the structure of the decision tree. This will present different results from what is likely to happen in a real life situation [17]. Secondly, decision trees are only applicable where quantitative data is involved. It ignores qualitative aspects of a dataset which are very useful in disease mapping.

Frequentist models like GLM, logistic regression, clustering and decision trees do not obey the likelihood principle. Likelihood principle is the proposition that in a statistical model, all evidence contained in a sample relevant to the model parameters is also contained in the likelihood function [18]. Likelihood principle can be looked at as the concept of replication. The frequentist approach fixes the parameter of interest, and replicates the data, whereas the Bayesian approach fixes the data, and replicates the parameter of interest [19]. The element of prior knowledge is what differentiates Bayesian modelling from frequentist modelling. In frequentist approach, inferences are made with no regard to prior knowledge [20]. Whereas Bayesian approach requires explicitly stating the priors, frequentist approach leaves the priors unspecified and the model is considered an incomplete model from a Bayesian viewpoint.

Bayesian approach obeys the likelihood principle. The likelihood principle holds that if two distinct sample designs yield proportional likelihood function for a parameter X, then all inferences about X should be identical from these two designs [21]. Bayesian approach defines probability as a measure of belief or uncertainty and the frequentist approach defines probability as resulting from long run frequency of occurrence for the specified events [22]. Bayesian approach is an improvement to the frequentist approach because instead

of only relying on the point estimates to test hypotheses, prior beliefs are developed then they are systematically updated using Bayes theorem to come up with posterior distributions [23]. In Bayesian approach, hypothesis testing involves a rigorous process of comparing posterior probabilities of different hypotheses given the data [24]. In frequentist approach the researcher comes up with a null and alternative hypothesis then calculates p-values or constructs confidence intervals to either accept or reject the hypotheses.

Application of Bayesian approach in disease mapping was a major achievement as it created better precision for prediction of spatial patterns for diseases (Allen, 2023). Empirical Bayes (EB) was integrated into disease mapping in order to create a smoothing effect on the results [25]. Bayesian disease mapping is the application of Bayesian approach in disease mapping (Oxford, 2023). It involves creation of hierarchically formulated Bayesian Generalized Linear Mixed Effects Models (GLMM) to estimate spatial dependence within a specified geographical area [26]. When doing Bayesian disease mapping for a specified area, Relative Risk (RR) is treated as a random variable and is used to specify a prior distribution. The prior distribution is then adjusted using model results to arrive at the posterior distribution [18]. Specified models are used to determine spatial autocorrelation depending on the nature of data collected and the expected results [27].

Poisson Log-Normal (PLN) model can be used for disease mapping. This model was derived from the Poisson-Gamma model also referred to as Negative Binomial model [28]. Poisson-Gamma model is used in situation where there is overspread count data [29]. The probability of the mean parameter in Poisson-Gamma model follows a Gamma distribution with a shape parameter and rate parameter [30]. The Poisson lognormal model assumes that the intensity parameter of a Poisson process has a lognormal distribution in a sample of observations [31]. This model can yield a highly skewed discrete distribution. Lognormal model is a versatile framework for the joint analysis of cases and incidences [32]. This model will be very effective in the joint analysis of concentration of TB incidence because it takes into account the multiplicative noise that arises from a large number of statistically independent fluctuations which give rise to a normal distribution [33].

Conditional Autoregressive (CAR) model is an effective model for disease mapping. This model according is used to predict future values based on past events [2]. It is applied to model a spatial process in which the variance of the value at each point is the same and the expected value at each point is same as the observed value at the neighboring points [34]. Conditional Autoregressive Model is used to simulate joint distribution of random vectors based on univariate conditional expectation [35]. Conditional Autoregressive Model is based on the assumption that the data is stationary, that there is a linear relationship between variables and their lagged values and lastly that the error term is a whitenoise process [12]. The advantage of using CAR model is that one can determine the degree of randomness using autocorrelation function. Another benefit of this model is that it helps to predict recurring patterns. Finally, CAR model makes it possible to predict outcomes with less information using self-variable series.

Bayesian approach has not been strongly adopted in medical research [36]. In order to understand the patterns of diseases for effective control and prevention, it is important to get accurate and on-point results. This study was aimed at applying models for TB mapping in Meru County Kenya. Meru is one of the top 5 counties in Kenya that have shown high number of TB cases. The study provided actionable insights for targeted interventions, resource allocation, and policy recommendations. To achieve the goal of the study, two models Conditional Autoregressive (CAR) and Poisson Lognormal(PLN) were fitted. The two models were compared to determine the one that had a better fit.

2. Materials and Methods

2.1. Spatial Data Source and Description

This study utilized secondary data collected from the 9 Sub-Counties in Meru County. These are; Buuri, Igembe North, Igembe South, Igembe Central, Imenti North, Imenti South, Imenti Central, Tigania East and Tigania West. The researcher specifically collected data from the Sub-County TB Coordinators. The dependent variable for the study was number of TB cases and the independent variables were year when the cases were recorded and the location (Sub-County) where the cases were identified. The data was collected for a 10-year period starting from 2014 to 2023. The collected data was sorted and recorded in excel worksheet.

2.2. Response Variable

Response variable for this study was positively identified TB cases in Meru county for the period 2014 - 2023. It assumed a Poisson distribution as follows

$$P(Y_i \mid \lambda_t) = \frac{\lambda_t^{yt} e^{-yt}}{y_t!}; y = 1, 2, 3, \ldots \quad (1)$$

Where, *e* represents Eulers Number (2.71828), *y* represents number of occurences and λ represents the rate of occurence. An important statistic that was used to assess TB situation within the Sub-counties is prevalence. Prevalence is the proportion of a population with a condition at a specified period of time [37]. Prevalence for this study was calculated using the formula: *Prevalence(P) = (Cases/ Population) * 10000*

2.3. Spatial Autocorrelation

To measure spatial autocorrelation, this study utilized the Morans I test. This is a test that determines the overall spatial autocorrelation by comparing the values of each observation with the values of neighboring observations [38]. Moran I has a local equation and a global equation as shown below. Equation 3.1 is the global equation for Moran I.

Where *n* is the number of spatial units, *W* is the spatial weight matrix defining the spatial relationships between

units, W_{ij} is the spatial weight between *i* and *j* , and *x* is the mean of the variable across all locations. The range of values for global Moran I will be between -1 and 1 with -1 representing perfect dispersion (negative spatial autocorrelation) and 1 representing perfect clustering (positive spatial autocorrelation). When *I = 0it* means that there is no spatial autocorrelation (Randomness)

$$I = \frac{n}{W}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) \quad (2)$$

Equation 3.2 shows the local equation for Moran I for a specific location *i*:

$$I_i = \frac{(x_i-\bar{x})}{S^2}\sum_{j=1}^{N} w_{ij}(x_j-\bar{x}) \quad (3)$$

Where; *N* is the number of spatial units, *xi* and *xj* represent values of variable of interest at locations *i* and *j* respectively. *Wij* represents the spatial weight between location *i* and *j* indicating the spatial relationship or proximity between them. *S2* is the variance of the variable of interest. The range of values for global Moran I will be between -1 and 1 with -1 representing perfect dispersion (negative spatial autocorrelation) and 1 representing perfect clustering (positive spatial autocorrelation). When *I = 0* it means that there is no spatial autocorrelation (Randomness)

2.4. Spatial Models

2.4.1. Conditional Autoregressive(CAR) Model

The CAR model was used to account for spatial dependence by considering the neighboring locations of each observation. It assumed that relative risk at a specific location was influenced by the values of relative risk at nearby locations. The model was developed on the assertion that each location is connected to its neighboring locations. The strength of this connection determines how much influence neighboring locations have on each other. A parameter *ui* was used to account for Subcounty specific random effects. The model was defined as:

$$y_i \sim Poisson(e_i\theta_i) \quad (4)$$

The model was fitted to determine log relative risk using equation 3.4.

$$log(\theta_i) = \alpha + u_i; i = 1, 2, 3, ..., 9 \quad (5)$$

In the model, Ui represented the correlated heterogeneity. Subcounty random effect in neighboring areas Uj was also factored in the model where i represented the reference Subcounty and j represented neighboring Subcounty. Each Subcounty had a unique set of neighbors. a represented the overall relative risk. Ui was smoothed towards the mean rate in the set of neighboring areas as follows:

$$\bar{u}_i = \frac{1}{\sum_j w_{ij}}\sum_j u_j w_{ij}\sigma_u^2 = \frac{\sigma_u^2}{\sum_j w_{ij}} \quad (6)$$

The model assumed that

$$[u_j, i \neq j, \tau_u^2] \sim N(\bar{u}, \sigma^2) \quad (7)$$

Parameter Estimation Using Conditional Autoregressive Model

MCMC was implemented through WinBUGS software which is based on Bayesian Analysis Using Gibbs Sampling (BUGS) [39]. The CAR model was built using two separate chains with each chain starting from an arbitrary initial value. The chains were used to sample the values for posterior parameters. The sampled values were used to update the parameters in subsequent iterations until convergence was achieved (Lawson, 2018). From these distributions, Gibbs Sampler estimated the joint posterior distribution of the parameters.When carrying out the model, 10000 iterations were taken with 2000 of them excluded as burn-in samples. In order to reduce effect of autocorrelation and to improve convergence, values of 5 were used to test convergence of the estimators. The results of the model included a DIC value which described the goodness of fit for the model. Dynamic trace plots were used to check the good mixing for the model. Autocorrelation plots were extracted to determine whether the model was in anyway affected by autocorrelation. Kernel density plots were used to establish normality of the model.

2.4.2. Poisson Lognormal (PLN) Model

The Poisson lognormal model combined elements of the Poisson distribution and the lognormal distribution. The model assumed that the observed counts (Y) follow a Poisson distribution. However, instead of modeling the counts directly, the model focused on the underlying rate (λ) on the log scale using a lognormal distribution as in equation 8.

$$Log(\lambda_{it}) = \beta_0 + \beta_1 X_i + u_i + \epsilon_{it} \quad (8)$$

Where log rate $\log(\lambda_{it})$ is the log of expected rate (λ) modelled as a linear combination of of predictors β0 representing the baseline log rate, U; accounts for spatial variation and ϵ_{it} indicating the random error and X_i representing the covariates associated with spatial location In this study, PLN was used to connect relative risk θ_i to a linear predictor that included normally distributed random effects that were denoted as v_i .The lognormal model for relative risk was;

$$y_i \sim Poisson(e_i\theta_i) \quad (9)$$

The model that was fitted for log relative risk was:

$$log(\theta_i) = \alpha + v_i; i = 1, 2, 3, ..., 9 \quad (10)$$

In the model, $v_i \sim N(0, \sigma_v^2)$ represented the subcounty random effects and *a* represented the overall relative risk. $\theta_i \sim Gamma(a, b)$ Prior distributions were $e^{v_i} \sim Lognormal(0, \sigma_v^2)$. The precision parameter for the model was $\tau_v^2 = \frac{1}{\sigma_v^2}$. σ_v^2 was assumed to be normally distributed.

Parameter Estimation Using Poisson Lognormal (PLN) Model

Under Poisson Lognormal(PLN) model, Gibbs Sampling used to iteratively sample from the conditional distributions of each parameter given the data and other parameters [39]. MCMC was implemented through WinBUGS software which is based on Bayesian Analysis Using Gibbs Sampling (BUGS). The model was developed using two separate chains with each chain starting from an arbitrary initial value. The chains were used to calculate the values for posterior estimators in the Bayesian Hierarchical model. When carrying out the model, 10,000 iterations were taken with 2000 of them excluded as burn-in samples. In order to reduce effect of autocorrelation and to improve convergence, values of 5 were used to test convergence of the estimators. The results of the model included a DIC value which described the goodness of fit for the model. Dynamic trace plots were used to check the good mixing for the model. Autocorrelation plots were extracted to determine whether the model was in anyway affected by autocorrelation. Kernel density plots were used to establish normality of the model.

2.5. Model Comparison

Deviance Information Criteria (DIC) was used to determine which model had a better fit for the data. DIC was used to measure the difference between observed data and the fitted values. The equation below was applied:

$$D = -2\sum_{i=1}^{n}(y_i log(\mu_i) - \mu_i + log(y_i!)) \quad (11)$$

Where y_i represents the observed counts (number of TB cases), μ_i represents the fitted values from the model (mean number of cases) and *n* is the number of observations (Sub-Counties). For the purpose of model comparison using deviance approach, lower *D* value indicates better fit (Leon et al, 2019).

2.6. Results and Discussion

2.6.1. Descriptive Statistics

Table 1 shows TB cases that were positively identified and the population of the respective Subcounties. The data for this study included positively identified TB cases in Meru County for the period 2014 - 2024. The cases were summed for each of the subcounties. The cases recorded were between 3700 and 4500 for all the Sub-counties. The subcounty with highest number of cases was Imenti South and the lowest was Imenti Central. Igembe Central subcounty had the highest population and Imenti Central had the least.

Table l. TB cases per Subcounty in Meru County

subcounty	Total_Cases	Population
Buuri	3996.00	157360.00
Igembe Central	4028.00	221412.00
Igembe North	4369.00	169317.00
Igembe South	4011.00	161646.00
Imenti Central	3711.00	134666.00
Imenti North	4024.00	177567.00
Imenti South	4413.00	206506.00
Tigania East	4322.00	177279.00
Tigania West	4291.00	139961.00

Sub-counties. The subcounty with highest number of cases was Imenti South and the lowest was Imenti Central. Igembe Central subcounty had the highest population and Imenti Central had the least.

The implication of these results is that every Subcounty had a different TB situation. This could be as a result of available risk factors or the social-economic environment [40]. These results are in agreement with the findings of Nyamogoba and Mbuthia [41] who when mapping TB cases in Western Kenya noted that different locations had recorded varying number of TB cases depending on the risk factors that were present. Distribution of TB cases is dependent on risk factors like under-nutrition, diabetes, HIV infection, alcohol use disorders and smoking among others [42].

Figure 2 presents a geographical mapping of TB cases in all the Subcounties. Most Subcounties had more than 4000 positively identified cases within the period. There were high cases recorded in Imenti South, Imenti North and Igembe North subcounties where the county borders Isiolo and Tharaka-Nithi counties.

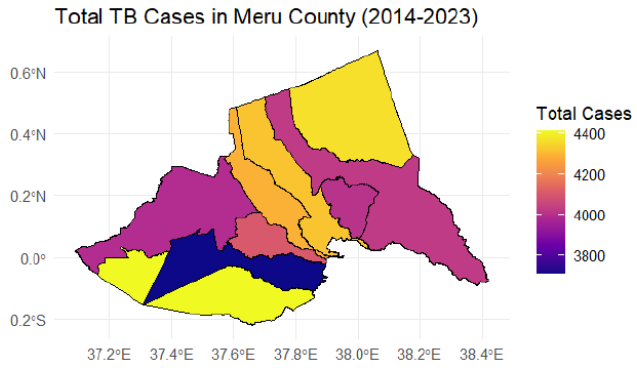

Figure 2. Mapping of TB cases within the Sub-counties

A similar finding was made by Nyamogoba and Mbuthia [41] in a paper that sought to establish gender and age distribution of TB cases in Western Kenya. It was noted that Kisumu County had the highest number cases and Bungoma County had the lowest. Areas neighboring Kisumu County had more cases compared to areas neighboring Bungoma County.

Table 2 shows the prevalence rates for all the Sub-counties. Prevalence per Sub-county ranged between 1800

and 3100. All the Sub-counties apart from Igembe Central had prevalence rates above 2000. Highest prevalence rate was recorded in Tigania West Sub-county and the lowest was Igembe central.

According to (WHO, 2021) prevalence rate is a relative measure of risk which is dependent on the number of cases and the population of an area. Prevalence rate represents the burden of a disease in a given population [43]. For this study, higher prevalence rate meant that there are higher chances of getting TB infection.

The chances of getting TB varied across the Subcounties as shown in figure 3. TB prevalence was found to be lower in the Imenti North and Imenti South Counties compared to the Igembe North and Igembe South. Imenti South and Imento North subcounties border Tharaka-Nithi County whereas Igembe North and Igembe South borders Isiolo County.

Table 2. TB Prevalence per Subcounty in Meru County

subcounty	Total_Cases	Population	Prevalence
Buuri	3996.00	157360.00	2539.40
Igembe Central	4028.00	221412.00	1819.23
Igembe North	4369.00	169317.00	2580.37
Igembe South	4011.00	161646.00	2481.35
Imenti Central	3711.00	134666.00	2755.71
Imenti North	4024.00	177567.00	2266.19
Imenti South	4413.00	206506.00	2136.98
Tigania East	4322.00	177279.00	2437.97
Tigania West	4291.00	139961.00	3065.85

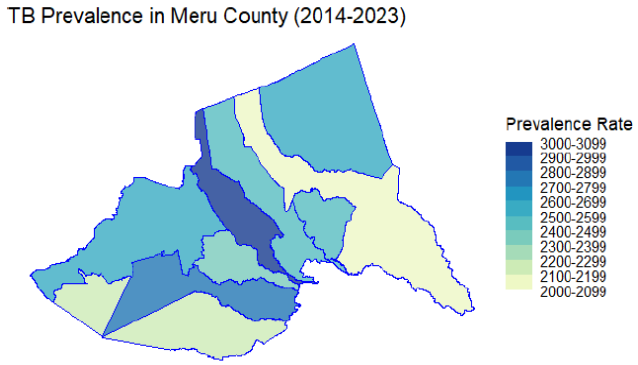

Figure 3. Mapping of TB prevalence within the Sub-counties

The results show clearly that high number of cases do not always imply a high prevalence rate as they are standardized by the population. High TB prevalence implies that the cases recorded were high compared to the population of the Subcounties. In a study by Nyamogoba and Mbuthia [41] Homabay County had the highest TB prevalence in Western Kenya despite the fact that Kisumu County had the highest number of cases.

2.6.2. Spatial Trends in TB Incidence

The first objective of this study was to determine spatial trends of TB incidence in Meru County. To achieve this objective, Global Moran's I and Local Moran's I tests were carried out to determine spatial autocorrelation.

Global Morans I

Table 3 shows results that were obtained for global Moran's I. Global Moran I statistic was found to be -0.3150. This value is less than 0 meaning that there was a negative spatial autocorrelation or dispersion.

Table 3. Results of the Global Moran I test under randomization

Statistic	Value
Moran I statistic standard deviate	-0.76845
p-value	0.7789
alternative hypothesis	greater
Moran I statistic	-0.31502734
Expectation	-0.12500000
Variance	0.06115078

The results obtained are contrary to the findings of Nyamogoba and Mbuthia [41] who pointed a positive spatial autocorrelation for TB prevalence in Western Kenya. In the study, areas with low prevalence had neighboring areas with low prevalence and areas with high prevalence had neighboring areas with high prevalence (high-high and low-low). For this study the Moran I statistic means implies that Subcounties with high prevalence had neighboring Subcounties with low prevalence and those with low prevalence had neighbors with high prevalence (high-low and low-high). The p-value which is 0.7789 (>0.05) means that local Moran I values are statistically insignificant and are likely to be as a result of random chance.

Local Moran's I

Table 4. First few rows of the Local Moran's I results

Ii	E.Ii	Var.Ii	Z.Ii	Pr(z > E(Ii))
0.007667509	-0.06270927	0.12595032	0.1983032	0.4214039
-0.547803543	-0.14315712	1.10396843	-0.3851212	0.6499262
-0.100495542	-0.04362420	0.16092436	-0.1417694	0.5563689
-0.227195978	-0.03275343	0.06788709	-0.7462730	0.7722487
-1.683953239	-0.20298091	0.62400725	-1.8747878	0.9695890
0.026296373	-0.05450753	0.06626116	0.3139083	0.3767953

The Z-scores for these observations are also close to zero, further supporting the conclusion that there is no strong evidence of clustering (high-high or low-low values) or outliers (high-low or low-high values) in the data. Specifically, while some observations have positive or negative I values, indicating potential spatial association or dissimilarity, these values are not statistically significant. Overall, the data does not demonstrate significant local spatial autocorrelation for these initial observations, implying that the distribution of values is fairly random across the study area. These findings are in agreement with what was established by JPH [44], that TB cases in Central Asia are spatially independent.

Figure 4 shows how p-values and Z-scores varied across locations. For all Subcounties in Meru County, z-scores were less that 1.65 and p-values were above 0.05.

This confirms the earlier finding that there is no evidence of spatial autocorrelation and that the observed spatial patterns are not statistically significant. Similar results for spatial independence were obtained by JPH (2013).

Figure 5 shows the relationship between total TB cases and local Moran's I. The values indicate an inverse relationship between Local Moran's I and total TB cases. This implies that as the TB cases increase, Local Moran's I decreases and when cases decrease local Moral I increases.

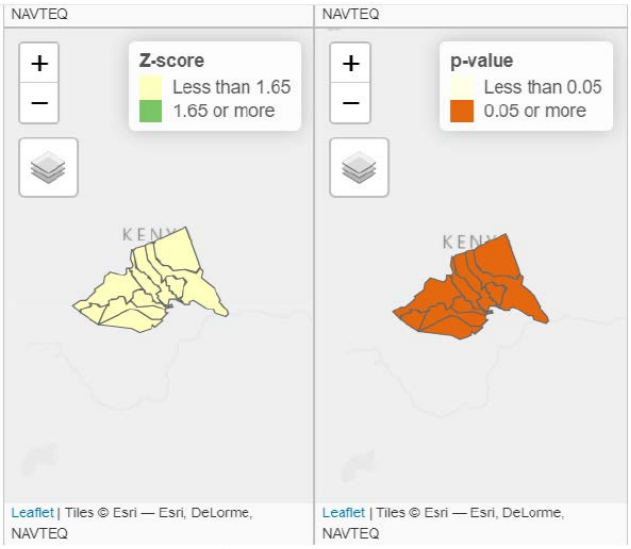

Figure 4. Variation of p-values and Z-scores across Subcounties

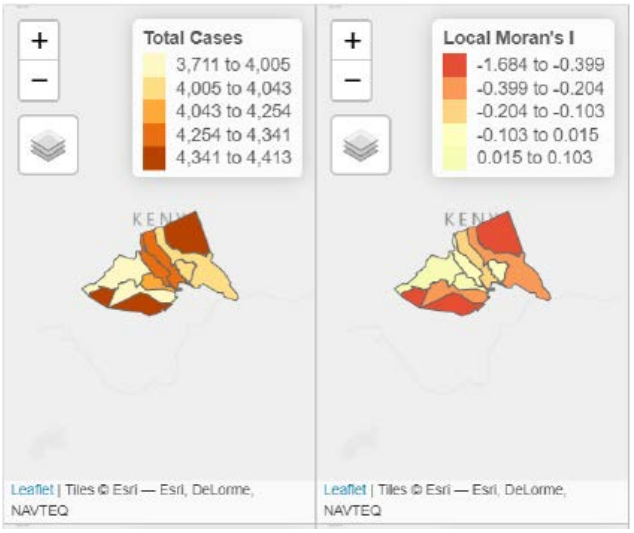

Figure 5. Relationship between total TB cases and Local Moran's I

The interpretation here is that as TB cases increase, there is a relative dispersion and as cases decrease there is a relative clustering. A similar finding was made by Ridzon and Mayanja [1] who found that spatial areas with lower number of HIV cases had neighboring areas with low cases and areas with higher number of cases had neighboring areas with low number of cases.

Figure 6 shows spatial autocorrelation clusters within Meru County. All Subcounties were classified as Low-High meaning that there is no strong evidence of clustering (high-high or low-low values). This type of classification means that Subcounties with low number of cases have neighboring Subcounties with high number of cases. This can be further described as one-way dispersion.

In the case of JPH [44], the nature of dispersion was described as high-low and low-high (two-way) which is different from this study which only recorded low-high and not high-low. From the addition evidence from low Z-scores and high p-values (>0.05), the imperfect dispersion can be attributed to chance as it has been stated before.

Figure 7 shows the nature of spatial autocorrelation that was established for the Subcounties. The results presented imply that Subcounties in Meru County had no Spatial Autocorrelation with respect to TB prevalence. Every Subcounty was dependent on itself and not in any way affected by neighboring Subcounties.

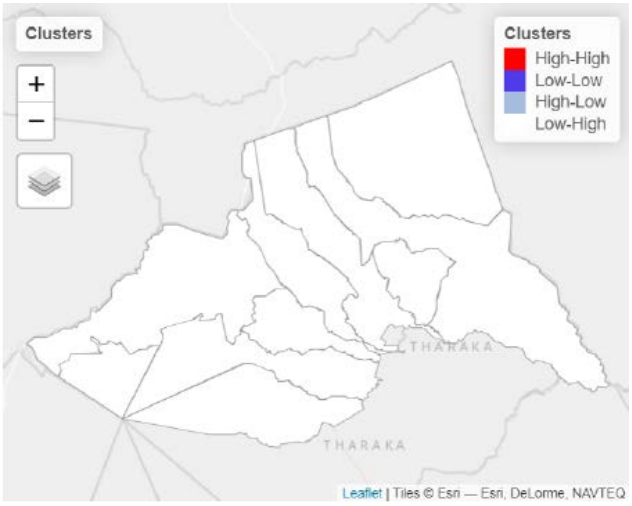

Figure 6. Spatial Autocorrelation Clusters

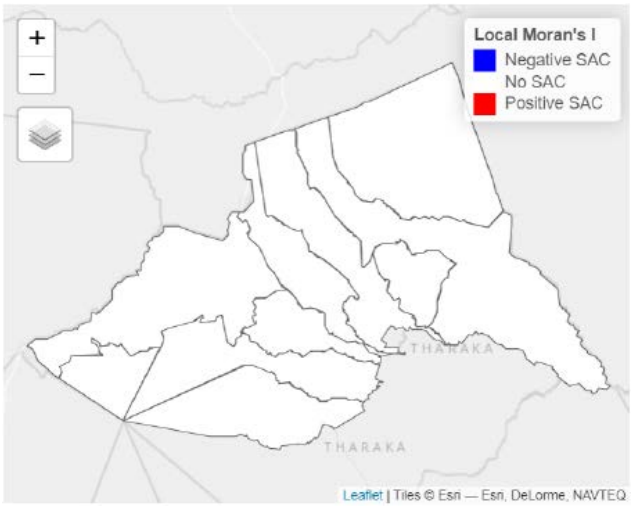

Figure 7. Summary of Spatial Autocorrelation

In such a situation, the results of a specific spatial point cannot be used to predict or control other spatial points. A similar finding was made by Warren and Mwangi [17], who recommended that each cluster should be handled independently as there was no spatial association.

2.6.3. Model Fitting

This section presents the results that were obtained from the fitted models. The CAR and PLN models were fitted as shown in the following equations

$$y_i \sim Poisson(e_i\theta i) \quad (12)$$

$$log(\theta_i) = \alpha + v_i; i = 1, 2, 3, ..., 9 \quad (13)$$

$$log(\theta_i) = \alpha + u_i; i = 1, 2, 3, ..., 9 \quad (14)$$

equation 12 shows the distribution of y_i that was assumed for both models. Equations 13 and 14 were applied for the PLN and CAR model respectively. Parameter a represents relative risk in both models and v_i and u_i represent random effects for the PLN and CAR model respectively. Precision parameter for the PLN

model was expressed as $\tau_v^2 = \frac{1}{\sigma_v^2}$ and precision parameter for CAR model was expressed as $\tau_u^2 = \frac{1}{\sigma_u^2}$ Table 5 shows the results for posterior estimators for the CAR and PLN models that were obtained from GIBBS Sampler.

Table 5. Comparison of PLN and CAR model parameters

	PLN		CAR	
	Estimate	95%	Estimate	95%
Mean	40.97		38.8	
Variance				
α	4.56×10^{-4}	$(3.32 \times 10^{-4}, 5.57 \times 10^{-4})$	0.01287	(0.01039,0.01535)
τ_v^2	37.96	(31.1,44.82)		
τ_u^2			35.65	(30.46,40.84)
σ_v^2	24.17			
σ_u^2			19.39	

Posterior estimators for the models were a which represented the overall level or relative risk, τ_v^2 which repre-sented the precision parameter for the PLN model, τ_u^2 which the precision parameter for the CAR model, σ_v^2 which was variance of v_i , σ_u^2 which was variance of u_i and the means $\bar{v}$ for PLN and $\bar{v}$ for CAR model.

Figure 8 shows trace plots for PLN model. Dynamic trace plots were used to visualize the level of mixing and stationarity for the samples within the models. Trace plots for PLN model indicated that there was good mixing which implied that each sample parameter was not strongly related to the sample that came before it.

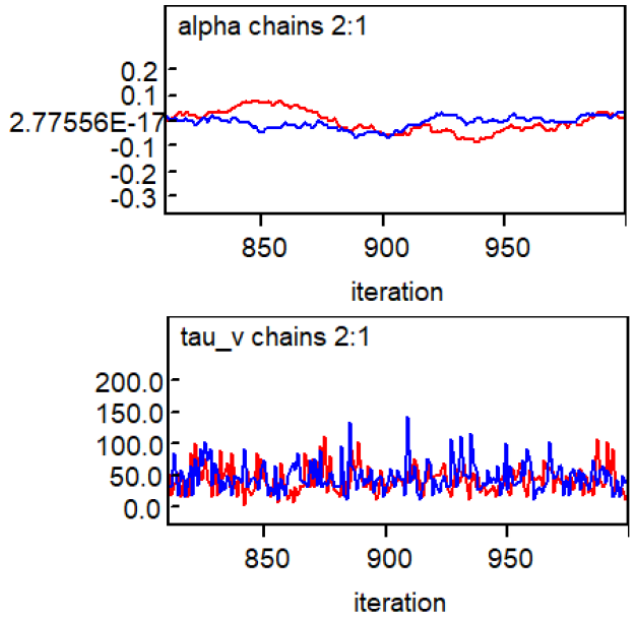

Figure 8. Dynamic Trace Plots for PLN Model

The trace plots show that each path was within the posterior distribution and consistent within a central trend. This finding of consistent stationarity and good mixing was similar to the finding of Sun and Wang [2] who established that the model paths were within the distribution from the first iteration to the last. The mixing for the alpha chains was smoother than that of tau_v chains. This indicated that relative risk was more consistent within the model compared to random effects.

Figure 9 presents the dynamic trace plot for CAR model. The plots indicated good stationarity and perfect mixing. This implies that the the chain has reached its equilibrium distribution and is sampling from the target posterior.

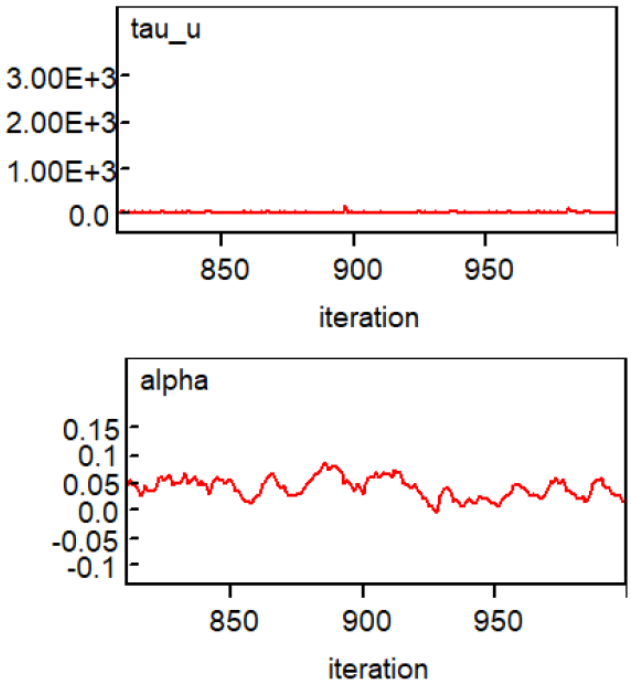

Figure 9. Dynamic Trace Plots for CAR Model

The CAR model plot shows a smoother curve compared to PLN plot. A smooth trace plot suggests that the MCMC chain is exploring this parameter space very efficiently. The smoothness indicates good mixing, meaning the chain is moving around the parameter space without getting stuck and is covering the range of values in the posterior distribution consistently. This behavior is ideal as it ensures that the estimates are reliable and representative of the true posterior distribution.

Figure 10 represents the autocorrelation plots for the posterior distribution in PLN models. These plots show that the dimensions of posterior distributions were mixing slowly. This was indicative of a high posterior correlation between parameters.

The autocorrelation was disappearing within the parameters, meaning that there were no random effects between locations. This strengthened the earlier finding of no spatial autocorrelation. The results indicated that the spatial dependence initially suspected was not statistically significant, confirming that the observed distribution of TB cases was not influenced by the proximity of different subcounties. Consequently, the spatial component did not contribute to the variance in TB prevalence, affirming that the spread of TB cases was independent of geographical clustering. This lack of spatial autocorrelation suggested that other factors, potentially socio-economic or environmental, might be driving the distribution of TB cases in Meru County.

Figure 11 shows autocorrelation plots for the CAR model. The plots show that autocorrelation was disappearing within the parameters. This suggests that random effects between locations were minimal or non-existent.

The diminishing autocorrelation observed in the plots reinforces the previous finding of no spatial autocorrelation, indicating that the distribution of TB

cases is not influenced by geographical proximity. The absence of significant spatial dependence suggests that other non-spatial factors are likely playing a more substantial role in driving the TB prevalence across the subcounties in Meru County.

Figure 12 presents the Kernel density plots for PLN model. The plots indicated that there was a relatively normal distribution for overall mean within the model parameters.

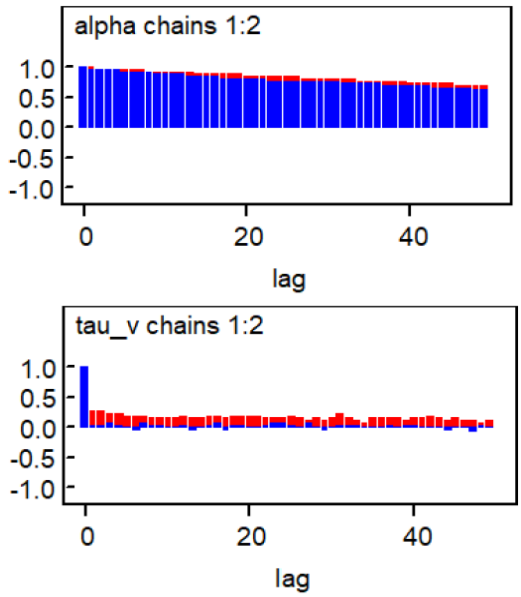

Figure 10. Autocorrelation Plots for PLN Model

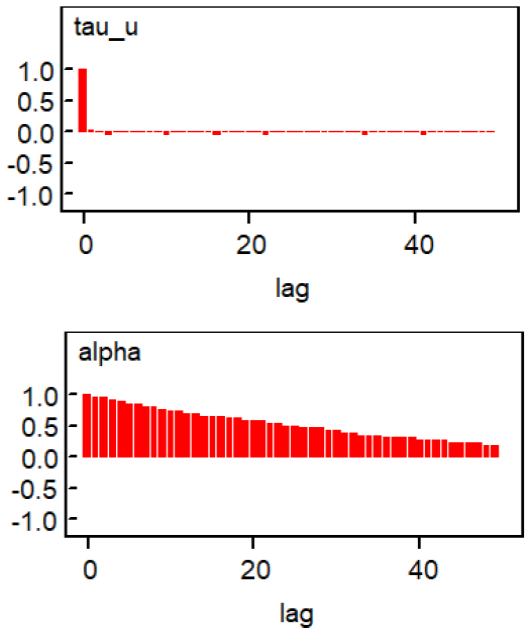

Figure 11. Autocorrelation Plots for CAR Model

The normality that was observed suggests that the model's assumptions regarding the underlying data distribution are appropriate and that the data conforms well to the expected pattern. Similar results were obtained by Karim and Barket [40] when modelling Covid-19 cases in Bangladesh using PG, PLN, CAR and Convolution models.

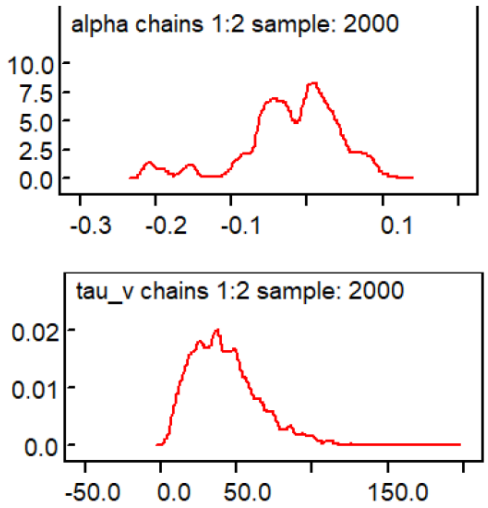

Figure 12. Kernel Density Plots for CAR Model

Figure 13 shows the Kernel density plots for CAR model. The normality observed in the kernel density plots enhances the reliability of the model's estimates and predictions. It also indicates that the variation in TB prevalence across the subcounties is well captured by the model, providing a solid basis for further analysis and interpretation.

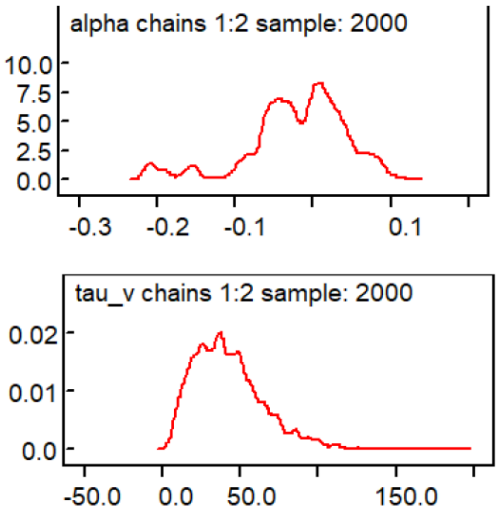

Figure 13. Kernel Density Plots for CAR Model

The alignment with a normal distribution implies that the model can be used confidently to explore the impact of different factors on TB prevalence in Meru County. The normal distribution observed in the kernel density plots also suggests that there are no significant outliers or skewness affecting the model's performance. This conformity enhances the model's robustness and supports its validity for making accurate inferences about TB prevalence across the region.

3. Model Comparison

Deviance Information Criterion (DIC) was used to compare the CAR and PLN models. Table 6 presents results of DIC test.

Table 6. DIC results for model comparison

Posterior Estimator	CAR Model	PLN Model
DIC	0.22541	0.56723

DIC for the CAR model was 0.22541 and that of PLN model was 0.56723. As Tam (2022) pointed out, a lower value of DIC indicates better fit for the model. The obtained results imply that CAR model had a better fit compared to the PLN model. A similar conclusion was made by Karim and Barket [40] who compared PG, PLN, CAR and Convolution models and found that CAR model had the best fit.

3.1. Conclusion

In conclusion, the values of Moran's I indicated that there was relatively no spatial autocorrelation between the Subcounties in Meru County. Absence of spatial dependence means that locations do not impact on each other. This is referred to as spatial independence. After fitting the models, it was established that there was good mixing for the sample parameters. Sample estimators for different samples were not related meaning that there was no autocorrelation. For both models, it was established that there was a consistent central trend for the paths within the posterior distributions. The model fitting results for the CAR and PLN models illustrated distinct strengths in the context of disease mapping for the nine subcounties. The CAR model's posterior estimators α , τ_u^2 and σ_u^2 offered a comprehensive understanding of spatial dependencies and random effects, providing nuanced insights into relative risks across subcounties. The PLN model, with its posterior estimators α , τ_v^2 and σ_u^2 similarly elucidated relative risk and random effects, albeit without the explicit spatial component of the CAR model. Dynamic trace plots for both models indicated good mixing and stationarity, suggesting that the sample parameters were consistent within their respective posterior distributions. The autocorrelation plots revealed high posterior correlation between parameters, with diminishing autocorrelation indicating the absence of random effects between locations. Kernel density plots further confirmed the normality of the overall mean distribution within model parameters. These findings underscored the robustness and reliability of both models in estimating disease risk, though the CAR model's explicit consideration of spatial relationships offered a more detailed and accurate representation of disease prevalence across Subcounties. Deviance Information Criterion (DIC) was employed to compare the fit of the CAR and PLN models. The study concluded that the CAR model provides a superior fit compared to the PLN model, further supporting the preference for the CAR model in capturing the spatial dynamics and risks associated with disease prevalence in the studied Subcounties.

References

[1] Ridzon, R. and Mayanja-Kizza, H. (2020) Tuberculosis AIDS in Africa, pp. 373–386.

[2] Sun, Z. and Wang, H. (2020) Network imputation for spatial autoregression model with incomplete data Statistica Sinica [Preprint].

[3] Zhang, H. (2009) Statistical Clustering Analysis: An introduction Clustering Challenges in Biological Networks, pp. 101–126.

[4] Allen, P. (2023) Bayesian analysis of conditional autoregressive models. Available at: https://www.ism.ac.jp (Accessed: 10 March 2024).

[5] Lawson, A.B. (2018a) Disease cluster detection Bayesian Disease Mapping, pp. 131–162.

[6] Bazett, T. (2022) Summary: Bayesian inference Bayesian Inference [Preprint].

[7] MacNab, Y.C. (2022) Revisiting gaussian markov random fields and bayesian disease mapping Statistical Methods in Medical Research, 32(1), pp. 207–225.

[8] Chiquet, J., Mariadassou, M. and Robin, S. (2021) The poisson-lognormal model as a versatile framework for the joint analysis of Species Abundances Frontiers in Ecology and Evolution, 9.

[9] Bruce, J. (2015) Multilevel modeling with logistic regression Best Practices in Logistic Regression, pp. 434–449.

[10] Gillian, B. (2021) Interpreting logistic regression coefficients Logistic Regression: A Primer, pp. 19–50.

[11] H. (2022) Polytomous logistic regression and alternatives to logistic regression Applied Logistic Regression Analysis, pp. 92–102.

[12] Berchtold, A. (2015) General autoregressive modelling of Markov chains Genève : Université de Genève /Faculté des sciences économiques et sociales.

[13] Emannuel, J. (2018) Clustering through decision tree construction in Medical Research Nonlinear Analysis: Modelling and Control, 6(2), pp. 29–41.

[14] Lu, X. (2021) Information mandala: Statistical distance matrix with clustering [Preprint].MacNab, Y.C. (2022) 'Bayesian disease mapping: Past, present, and future', Spatial Statistics, 50, p. 100593. Ma, X., Chen, S. and Chen, F. (2016) 'Correlated random-effects bivariate poisson lognormal model to study single-vehicle and Multivehicle crashes', Journal of Transportation Engineering 142(11).

[15] Klar, N. and Darlington, G. (2018) Methods for modelling change in Cluster Randomization Trials Statistics in Medicine, 23(15), pp. 2341–2357.

[16] Sheng, X. and Thuente, D. (2011) Decision tree learning in general game playing', Artificial Intelligence and Applications / 718: Modelling, Identification, and Control [Preprint].

[17] Warren, C. and Mwangi, A. (2017) Integrating tuberculosis case finding and treatment into focused antenatal care in Kenya [Preprint].

[18] Zamzuri, Z.H. (2015) Critical elements on fitting the Bayesian multivariate Poisson lognormal model AIP Conference Proceedings [Preprint].

[19] Aitkin, M. (2022) Statistical inference I – discrete distributions', Introduction to Statistical Modelling and Inference pp. 47–90.

[20] Watts, G. (2012) Who annual report finds world at a crossroad on tuberculosis BMJ, 345(oct19 1).

[21] Zamzuri, Z.H. (2018) The spatio-temporal multivariate Poisson lognormal model AIP Conference Proceedings [Preprint].

[22] Gelman, A. and Nolan, D. (2017b) Statistical inference Oxford Scholarship Online [Preprint].

[23] Lee, P.M. (2012) Bayesian statistics: An introduction Chichester, West Sussex: Wiley.

[24] Coly, S. et al. (2019) Bayesian hierarchical models for disease mapping applied to contagious pathologies, PLOS ONE. Available at: https://journals.plos.org/plosone/article?id=10.1371, (Accessed: 05 March 2024).

[25] Martinez-Beneito, M.A. and Botella-Rocamora, P. (2019) Disease mapping from foundations Disease Mapping, pp. 105–187.

[26] Lawson, A.B. (2018b) Bayesian inference and modeling Bayesian Disease Mapping, pp. 19–36.

[27] Erkelens, J.S. (2016) Autoregressive modelling for speech coding: Estimation, interpolation and quantisation., Netherlands: Delft University Press.

[28] Ding, H. and Sze, N.N. (2022) 'Effects of road network characteristics on bicycle safety: A multivariate poisson-lognormal model', Multimodal Transportation 1(2), p. 100020.

[29] Kazemi, I. et al. (2013) 'Multivariate poisson-lognormal model for modeling related factors in crash frequency by severity',

International Journal of Environmental Health Engineering 2(1), p. 30.

[30] Engen et al. (2002) 'Analyzing spatial structure of communities using the two-dimensional poisson lognormal species abundance model', The American Naturalist 160(1), p. 60.

[31] Martinez-Beneito, M.A. and Botella-Rocamora, P. (2019) Some essential tools for the practice of bayesian disease mapping Disease Mapping, pp. 51–103.

[32] Chiquet, J., Mariadassou, M. and Robin, S. (2020) The poisson-lognormal model as a versatile framework for the joint analysis of Species Abundances [Preprint].

[33] Francis, J. (2023) Poisson-lognormal model with measurement error in covariate for small area estimation of Count Data Communications in Mathematical Biology and Neuroscience [Preprint].

[34] Zhang, H. et al. (2015) 'PLNseq: A multivariate Poisson lognormal distribution for high-throughput matched RNA-sequencing read count data', Statistics in Medicine 34(9), pp. 1577–1589.

[35] Rabier, C.E. (2014) on statistical inference for selective genotyping Journal of Statistical Planning and Inference, 147, pp. 24–52.

[36] Haining, R. and Li, G. (2020) Modelling spatial and spatial-temporal data: Future agendas? Modelling Spatial and Spatial-Temporal Data, pp. 565–576.

[37] Harding, E. (2020) Who global progress report on tuberculosis elimination The Lancet Respiratory Medicine, 8(1), p. 19.

[38] Elliot, C. (2014) Poisson - lognormal distribution Wiley StatsRef: Statistics Reference Online [Preprint].

[39] Oxford (2023) Conditional autoregressive model. Available at: https:// www. oxfordreference.com/ view/ 10.1093/ oi/ authority.20110803095631676 (Accessed: 11 March 2024).

[40] Karim, R and Barket, S 'Bayesian Hierarchical Spatial Modeling of COVID-19 Cases in Bangladesh', Annals of Data Science

[41] Nyamogoba, H. and Mbuthia, G. (2018) 'Gender-age distribution of tuberculosis among suspected tuberculosis cases in western Kenya', Medicine Science | International Medical Journal p. 1.

[42] WHO (2021) Tuberculosis World Health Organization. Available at: https://www.who.int/health-topics/tuberculosis (Accessed: 11 March 2024).

[43] World Bank (2022) Incidence of tuberculosis (per 100,000 people) - kenya, World Bank Open Data - TB situation in Kenya Available at: https:// data.worldbank.org/ indicator/ SH.TBS.INCD? locations = KE (Accessed: 11 March 2024).

[44] JPH (2013) 'Tuberculosis control in south-East Asia Region: Annual TB Report 2013', WHO South-East Asia Journal of Public Health 2(1), p. 75.