Bayesian Hierarchical Spatial Modeling of COVID-19 Cases in Bangladesh

Md. Rezaul Karim[1] · **Sefat-E-Barket**[1]

Abstract

This research aimed to investigate the spatial autocorrelation and heterogeneity throughout Bangladesh's 64 districts. Moran I and Geary C are used to measure spatial autocorrelation. Different conventional models, such as Poisson-Gamma and Poisson-Lognormal, and spatial models, such as Conditional Autoregressive (CAR) Model, Convolution Model, and modified CAR Model, have been employed to detect the spatial heterogeneity. Bayesian hierarchical methods via Gibbs sampling are used to implement these models. The best model is selected using the Deviance Information Criterion. Results revealed Dhaka has the highest relative risk due to the city's high population density and growth rate. This study identifies which district has the highest relative risk and which districts adjacent to that district also have a high risk, which allows for the appropriate actions to be taken by the government agencies and communities to mitigate the risk effect.

Keywords COVID-19 · Bayesian hierarchical models · Spatial dependency · Conditional Autoregressive model · Convolution model · Modified CAR model

1 Introduction

COVID-19 is today considered a significant threat to the economy, social life, education, and workplace; in short, to human lives. This deadly virus is an RNA virus with a single strand and a thick envelope. The virus can spread through coughing droplets, direct contact, or touching contaminated surfaces with filthy hands [1]. Immigrants' unwillingness to follow social distancing and quarantine rules can be one of the primary causes of COVID-19 transmission within communities [2].

✉ Md. Rezaul Karim
   rezaul@juniv.edu

1   Department of Statistics, Jahangirnagar University, Savar, Dhaka 1342, Bangladesh

🖄 Springer
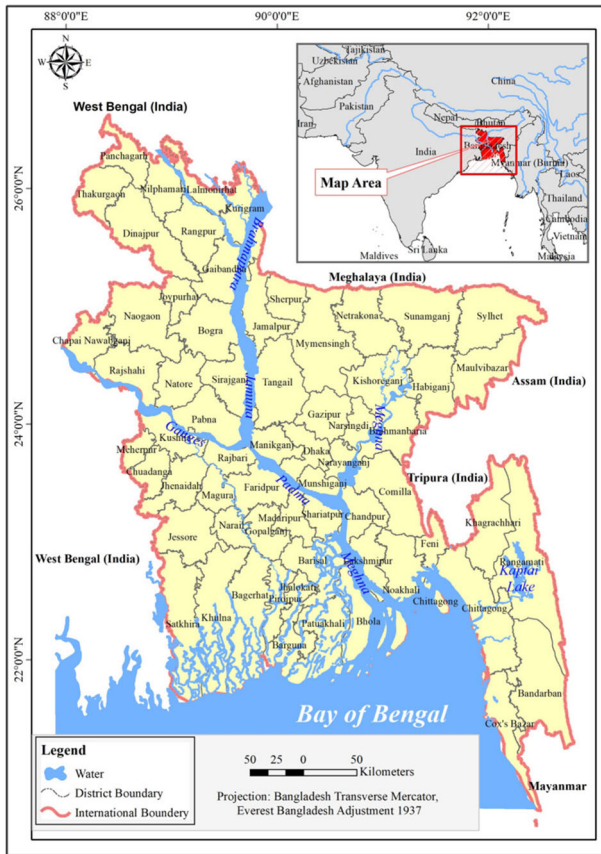
Fig. 1 The study area (District-wise map of Bangladesh) for the research

The virus was initially detected on a limited scale in November 2019. Furthermore, the following month, in December 2019, the first significant cluster was discovered in Wuhan, China [3, 4]. The World Health Organization (WHO) called the outbreak an international public health emergency on January 30, 2020 [5]. The WHO declared the epidemic pandemic on March 12 of the same year [6]. The virus has now infected about 1.19 million people in China, and about 5224 people have died. SARS-CoV-2 spread within and outside China, affecting people who had never touched animals means the virus can spread from one individual to another. According to [7], on June 7, 2022, 535,938,392 cases were identified worldwide. This infectious agent has caused 6,321,595 deaths. The first COVID-19 patient identified in Bangladesh was on March 8, 2020 [8]. The Center of Epidemiology, Disease Control and Research (IEDCR), Bangladesh's premier national institute for surveillance, outbreak investigation, and research on existing, emerging, or undiscovered infectious illnesses, hired more than 80 personnel in June 2020 to increase COVID-19 surveillance and contact tracing. This scheme was done with funding and technical support from WHO [9]. As of June 7,2022, there were 1,953,700 cases identified and 29,131 deaths [10]. Due to cultural,

political, socioeconomic, and environmental differences, it is more important than ever for people worldwide to work together to reduce the adverse effects [11].

Data science plays an increasingly essential role in finding solutions to societal and economic issues resulting from the exponential development in the current amount of data and the ongoing advancements in information technology [12]. Data science has a significant presence in business data mining [12], which enables real-time decision-making through the utilization of a mix of technologies that involve artificial intelligence (AI) and the internet of things (IoT) [13]. Various challenges have been described using data science methodologies, including crop harvesting, characterization of epidemiological outbreak patterns, commercial data mining, and e-commerce fraud [11–16]. Data science is also used in healthcare, especially since the COVID-19 pandemic started around the world at the beginning of 2020 [17]. In case to explain the transmission pattern of the COVID-19 outbreak in China, numerous data science tools have been used to analyze by undertaking retrospective and prospective investigations based on age-specific social contact-base transmission [16, 17]. The field of research known as epidemiology may be defined as the investigation of the incidence and spread of illnesses to determine the factors that cause them [18, 19].

A study using ARIMA models predicts COVID-19 will fill ICUs in Italy, France, and Spain. The number of instances will rise if the virus stays the same. Clinical and societal difficulties may be intractable, leading to catastrophe [20]. Cluster analysis was used by [14] to classify actual groups of COVID-19 datasets representing multiple states and union territories in India. This work aimed to enhance monitoring procedures and improve government policy. Previous research investigated demographic aspects important for COVID-19 transmission in Bangladesh using conventional statistical models [21], but no research has examined the spatial dependency of COVID-19 cases across Bangladesh's 64 districts. The traditional statistical methods assume that the observations are independent and identically distributed. A cluster pattern violates classical assumptions of independence and homogeneity (stationarity) and renders classical methods inefficient or inappropriate.

Bayesian inference is used for disease mapping by estimating parameters based on actual data and prior assumptions [22]. It is spatial modeling, a helpful tool for investigating the relative risk of COVID-19 [23–26]. Hierarchical Bayesian methods are often used to model the data's overdispersion and spatial correlation. Models with random effects, Poisson-Gamma and Poisson-Lognormal, are two classic solutions to the problem. These two models account for Poisson error-caused data overdispersion [27], called "uncorrelated heterogeneity" in disease mapping. Overdispersion can be caused by a spatially unstructured covariate, many zero counts, or many counts far from the mean. These models assume a gamma distribution or lognormal-distributed random effect to deal with spread-out data. Early disease mapping was dominated by them [28]. The Conditional Autoregressive (CAR) model investigates how data are related spatially [29]. Because it can use many weighting schemes, the model is widely used. It provides better solutions than unstructured alternatives. Convolution (COV) models combine unstructured random effects and structured random effects [30, 31]. Utilizing two independent sets of random effects, one of gamma and the other of normal, the combined model takes into consideration both overdispersion and clustering [32].

In this manuscript, we performed research to find the answer to the following research questions. Does the spatial patterning of the relative risk of COVID-19 give rise to the conclusion that the locations and shapes of geographic features are clustered? So, how do we measure the data's spatial dependence and spatial heterogeneity? In addition, how do we capture these by the statistical models? If there is a cluster, do governments tend to compare their policies with those of district neighbours, or do they behave independently?

Different types of spatial autocorrelation measures, including Moran I and Geary C, are used to measure spatial dependency. The Bayesian hierarchical models via Gibbs sampling are employed to assess the heterogeneity of spatial data. Complete methods and materials regarding data source and Bayesian statistical models are presented in Sect. 2. The most critical variables in the sequence of events that lead to the high relative risk of COVID-19 in Bangladesh have been found using spatial autocorrelation and well-suited model results in Sect. 3. Finally, Sect. 4 highlighted that the study's results could give government agencies useful information for taking actions that will reduce the prevalence of COVID-19.

2 Methods and Materials

2.1 Spatial Data Source and Description

The Directorate General of Health Service (DGHS) (https://dghs.gov.bd/) was the source of the information that was utilized in this article. From June 5, 2021 to May 14, 2022, data on the number of affected cases were collected from this source. Other two variables, namely annual growth rate per district and per district population, are obtained using data from the Bangladesh Bureau of Statistics (BBS)'s census of 2011, which was carried out in 2011 (https://tinyurl.com/3dcspsfp). The projected population of 2021 of all 64 districts has been taken from (https://tinyurl.com/3j4ff8r4). The growth rate of 2022 is then calculated for the projected population of 2021 using the geometric model. Calculation of growth rate uses the following formula :

$$r = \sqrt[n]{P/P_0} - 1$$

where, P_0 is the population of 2021, P is the projected population of 2022, n is the number of intercensal period and r is the growth rate. Also, the formula of the prevalence rate is shown below:

$$\text{Prevalence Rate} = \frac{\text{Affected No. of Cases}}{\text{Annual Growth Rate } \times \text{ District Population}} \times 100,000$$

It is necessary to determine the prevalence rate for each district individually. After that, the results estimate how many individuals are affected out of a total population of 100,000 in each district.

2.2 Distribution of Response Variable

The Poisson distribution is a discrete probability distribution. This means that the value of the variable can only be a whole number, like 0, 1, 2, 3, etc. It can't be a fraction or a decimal [33]. The probability mass function (pmf) is given below:

$$\Pr(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}; \quad y = 0, 1, 2, \ldots$$

where, e is Euler's number ($e = 2.71828\ldots$), y is the number of occurrences, $y!$ is the factorial of y, λ is the rate of occurrence.

2.3 Spatial Autocorrelation

Getis [34] states that one of the essential parts of spatial analysis is the idea of spatial autocorrelation. The use of spatial autocorrelation helps to determine whether or not there is systematic spatial variation by instantaneously taking into account the feature districts and associated values [35]. This correlation introduces a divergence from the independent observation assumption that is used in conventional models [36]. The spatial autocorrelation method analyzes the spatial patterns of individual entities, determining whether they are clustered, dispersed, or random [37].

2.3.1 Moran I Autocorrelation

Moran's I is a correlation coefficient that measures a data set's overall spatial autocorrelation. In other words, it assesses how similar one object is to others around it. If items are attracted to each other, it indicates that the observations are not independent. Given a set of features and a related attribute, it determines whether the pattern expressed is clustered, dispersed, or random. It has a wonderful utility of comparing the value of a variable at any one point with other locations. Whether or if one site is autocorrelated with others. Moran's I statistic is not limited to being less than one. Moran's I values range roughly from +1 to −1, with an expected value of $-1/(n-1)$ [38]. The formula for Moran's I statistic which is similar to the Pearson's coefficient [39] in Equation 1 is as follows:

$$I = \frac{n \sum_i \sum_j \omega_{ij} Z_i Z_j}{(n-1) \sum_i \sum_j \omega_{ij}} \tag{1}$$

where n is the number of districts; ω_{ij} indicates the quantification of the spatial weight between two districts i and j; z-scores are the transformation of the variables in which we are interested; in the numerator, products of two different z-scores in adjacent districts are being summed. Since the weights are row-standardized $\sum \omega_{ij} = 1$, the initial phase of the spatial autocorrelation study is to generate a spatial weight matrix including information on the neighborhood structure for each site. Adjacency is the

administrative districts closely adjacent to the district, including the district itself. Administrative districts that are not adjacent to one another are given no weight [40].

Two districts with larger scores will come up with positive components in the numerator, contributing to a positive spatial autocorrelation. In contrast, negative spatial autocorrelation can be found if two districts emerge with lower scores [41]. However, the p-value is what determines whether or not the clustering is significant. If the absolute values of z-scores are high, then the clustering will be intense; however, the significance of the clustering will be determined by the p-value. p-values less than significance level (0.05) and very high z-values points that null hypothesis can not be accepted. This suggests that there is clustering [42]. On the other hand, high p-values and low z-values imply the null hypothesis's acceptance. Consequently, Moran's I value of -1 shows perfect scattering, whereas zero suggests the spatial pattern of randomness, and $+1$ proclaims a clustering pattern and perfect spatial autocorrelation.

The Global Moran's I is an inferential statistic, which means that the investigation findings are always understood in the context of its null hypothesis. Let,

H_0 : 64 districts are randomly distributed i.e. there is no spatial dependency
between the neighboring districts (No spatial clustering exists);

H_1 : 64 districts are positively autocorrelated (Spatial clustering exists).

2.3.2 Local Moran I

Various approaches to local spatial autocorrelation have been developed over the last few decades [43–45]. Local Moran's I is one of the most well-known indicators that quantify the degree of similarity between two districts and their neighbors. Researchers compute the local Moran's I to find clusters and geographic outliers locally [46, 47]. According to Anselin [46], spatial statistics can find autocorrelation of specified orders in the studied area. He develops LISA (Local Indicators of Spatial Autocorrelation), which indicates on a map, for each observation, how much similar values are clustered near that observation [44]. To calculate local Moran I, the formula is defined as follows

$$I_i = p_i \sum_j \omega_{ij} p_j \qquad (2)$$

where, p_i is the variation between i's district relative risk and the mean; p_j is the weight of neighboring areas in the statistic, normalized for the number of neighbors.

2.3.3 Geary C

Geary's C is a measure of spatial autocorrelation, which can also be thought of as an attempt to assess whether or not neighboring observations of the same occurrence are associated with one another [47]. The correlation in spatial autocorrelation is multi-dimensional and works in both directions, making it a more complicated concept than simple autocorrelation. The formula used for Geary C calculation, defined by [47] in

this study, is expressed in Equation 3

$$C = \frac{(n-1)\sum_i \sum_j \omega_{ij}(x_i - x_j)^2}{2\omega \sum_i (x_i - \bar{x})^2} \tag{3}$$

where, x is the variable of interest; $\bar{x}$ is the mean of x; ω is the sum of all ω_{ij}.

The value of C can fall within the range [0,2]. If the obtained statistic is $0 \leq C < 1$, then it is possible that there is a positive autocorrelation among the districts. $C \geq 1$ indicates of having little spatial autocorrelation. If $1 \leq C < 2$, then one can deduce that there is negative autocorrelation between the districts as a whole.

2.4 Spatial Regression Models

Spatial regression is a component of regression models that incorporate spatial position. The presence of a dependent relationship among a set of observations, known as spatial dependence, indicates that the model follows an autoregressive process [48, 49].

2.4.1 Poisson-Gamma Model

A negative binomial model can readily be used to model additional variation as an alternative to the Poisson model. Consider that a negative binomial distribution can be seen as a mixed model with gamma random-effects for each area which is alternatively known as the Poisson-gamma model [50]. This model assumes that the number of affected cases within each district is independent and follows a Poisson distribution with mean $e_i\theta_i$ i.e., $y_i \sim \text{Poisson}(e_i\theta_i)$, with the assumption that

$$\lambda_i = e_i\theta_i \quad ; \quad i = 1, 2, 3, \ldots, 64$$

is constant within each district. The parameter of interest in the model is the relative risk (θ_i), and to account for unobserved heterogeneity, it is assumed that θ_i follows a gamma prior distribution with parameters a and b, and when combined with a Poisson likelihood, gives a gamma posterior. Then, the relative risk has a gamma posterior, that is

$$\theta_i \sim \text{Gamma}(a + y_i, b + e_i).$$

The Poisson-Gamma model assumes that the observations are independent. When most spatial data are correlated, it does not take into account the spatial correlation between risk in nearby areas; it does not also allow an easy adjustment for spatial covariates. For this reason, PLN, CAR, and Convolution models were considered.

2.4.2 Poisson-Lognormal Model

The Poisson-lognormal (PLN) model is an alternative that can be considered in place of the Poisson-gamma model. It connects the relative risk, denoted by θ_i, to a linear

predictor that includes a normally distributed random effects component, denoted by v_i [50]. The log-normal model for the relative risk is defined as:

$$y_i \sim \text{Poisson}(e_i \theta_i)$$

with

$$\log(\theta_i) = \alpha + v_i; \quad i = 1, 2, 3, \ldots, 64$$

where, $v_i \sim \text{N}(0, \sigma_v^2)$, is the district-specific random effects, capturing extra Poisson variability in the log-relative risk of COVID-19 in area i, $i = 1, 2, \ldots, 64$ and α is the overall level of the relative risk. In the Poisson-Gamma model, we consider $\theta_i \sim \text{Gamma}(a, b)$ whereas we consider $e^{v_i} \sim \text{Lognormal}(0, \sigma_v^2)$ with precision $\tau_v^2 = 1/\sigma_v^2$ where, σ_v^2 follows gamma prior distribution.

2.4.3 Conditional Autoregressive Model

In this model, the district-specific random effect component takes into account the effects that vary in a structured manner in space, i.e., the correlated heterogeneity. The model was introduced by [28] in an empirical Bayes setting and developed by [29] in a fully Bayes implementation. The model is defined as follows:

$$y_i \sim \text{Poisson}(e_i \theta_i)$$

with

$$\log(\theta_i) = \alpha + u_i; \quad i = 1, 2, 3, \ldots, 64$$

where, α is an overall level of the relative risk, correlated heterogeneity denotes by u_i, which means the values of the district-specific random effects, u_j in "neighboring areas". The model uses a spatial correlation structure to estimate the risk in any area which depends on neighbouring areas [50]. It is presumable that the correlated heterogeneity terms will behave in accordance with an intrinsic CAR model, such as the one presented by [51], the random impact caused by the CAR follows a normal distribution, and its mean and variance are weighted following the averages and variances of the adjacent areas i.e.

$$[u_i \mid u_j, i \neq j, \tau_u^2] \sim N(\bar{u}_i, \sigma_i^2)$$
$$\bar{u}_i = \frac{1}{\sum_j \omega_{ij}} \sum_j u_j \omega_{ij} \quad \sigma_i^2 = \frac{\sigma_u^2}{\sum_j \omega_{ij}}$$

where, u_i is smoothed towards the mean rate in the set of neighbouring areas; mean $\bar{u}_i$ which means it is the average of the spatial random effects of these neighbors and variance parameter σ_u^2 with precision $\tau_u^2 = 1/\sigma_u^2$. Here, σ_u^2 follows gamma prior distribution.

2.4.4 Convolution Model

Convolution models do not, however just include a random effect to correct for overdispersion; rather, they also include a random-effects term that controls for spatial autocorrelation. This is because convolution models take into consideration both [50]. In this model, district-specific random effects are decomposed into a component that takes into account the effects that varies in a structured manner in space, i.e., the correlated heterogeneity defined by u_i and a component v_i that models the effects that vary in an unstructured way between areas i.e., the uncorrelated heterogeneity. Like the CAR model, this model was equally introduced by Clayton and Kaldor [28] in an empirical Bayes setting and developed by Besag et al. [29] in a fully Bayes implementation. The model is defined as:

$$y_i \sim \text{Poisson}(e_i\theta_i)$$
$$\log(\theta_i) = \alpha + u_i + v_i \quad ; \quad i = 1, 2, 3, \ldots, 64$$

The model uses a spatial correlation structure to estimate the risk in any area which depends on neighbouring areas. This is assumed to be normally distributed i.e.

$$[u_i \mid u_j, i \neq j, \tau_u^2] \sim N(\overline{u}_i, \sigma_i^2)$$
$$\overline{u}_i = \frac{1}{\sum_j \omega_{ij}} \sum_j u_j \omega_{ij}, \; \sigma_i^2 = \frac{\sigma_u^2}{\sum_j \omega_{ij}}.$$

2.4.5 Modified CAR Model

The Poisson model, in particular, is convenient and sophisticated from a mathematical perspective, but the extension is required due to the model's restrictive nature. Firstly, the model does not accurately describe data variation, and secondly, hierarchies are frequently accounted for by including random effects that are assumed to be normally distributed. The modified CAR model takes into account both overdispersion and clustering by employing two distinct sets of random effects, one of gamma and the other of normal [32]. This model is also named as "Combined model" explained in Neyens et al. [50]. This is what the model is defined to be :

$$y_i \sim \text{Poisson}(e_i\theta_i)$$
$$\log(\theta_i) = \log(g_i) + \alpha + u_i$$

where g_i terms, which are assumed to follow a gamma distribution, are used to model uncorrelated heterogeneity

$$g_i \sim \text{Gamma}(a, b)$$

whereas the modeling of correlated heterogeneity is accomplished through the accumulation of CAR random effects u_i.

$$[\quad u_i \quad | \quad u_j, i \neq j, \tau_u{}^2] \sim N(\overline{u}_i, \sigma_i{}^2)$$
$$\overline{u}_i = \frac{1}{\sum_j \omega_{ij}} \sum_j u_j \omega_{ij}, \; \sigma_i{}^2 = \frac{\sigma_u{}^2}{\sum_j \omega_{ij}}.$$

In contrast to the convolution model, the modified CAR model models uncorrelated heterogeneity with a gamma distribution instead of a lognormal distribution [50]. The research conducted by [32] demonstrates that the gamma distribution can accurately model extra-variance. They provide a more detailed theory for multiple data types, which is useful for the combined model.

When overdispersion random effects are present alongside normal ones, [32] modified conjugacy to account for them. The goal of this property is to ensure that strong conjugacy holds even in the presence of random effects that follow a normal distribution. That is to say; we will only take conjugacy into account if the random effect ui follows a normal distribution. Hence, the Poisson and gamma distributions are conjugate. The posterior distribution is defined as

$$\theta_i | u_i, y_i \sim \text{Gamma}(a + y_i, b + e_i \kappa_i)$$

with $\kappa_i = exp(\alpha + u_i)$. As a result, the conditional mean of θ_i is $(a + y_i)/(b + e_i \kappa_i)$, and this can be rewritten as a weighted average of the prior mean, which is a/b.

2.5 Deviance Information Criterion

For the purpose of model comparison, the deviance information criterion (DIC) and a related measure, p_D, which counts the number of model parameters that are most important [52]. How to define the effective number of parameters in a Bayesian framework, particularly for complicated models, is a crucial subject. A DIC difference greater than 10 eliminates the model with the higher DIC, while a DIC difference less than 5 does not indicate a statistically significant result. Since DIC depends on MCMC output, it's sensitive to sampling fluctuations [53].

To demonstrate that DIC is additive using models and priors that are independent of one another, let the vector of parameters be θ associated with y. And $f(y|\theta)$ and $f(y)$ denote the conditional and marginal distributions of y. Then,

$$DIC = \overline{D} + p_D$$

where, D is the posterior expected value of the *deviance* function, posterior deviance is defined as :

$$p_D = \overline{D} - D(\overline{\theta})$$
$$\overline{\theta} = E[\theta|y]$$
$$\overline{D} = E[D(\theta)|y]$$

are the posterior means of θ and the Bayesian deviance

$$D(\theta) = -2\ln\{f(y|\theta)\} + 2\ln\{f(y)\} \tag{4}$$

Suppose, y and θ be partitioned as $(y_1, \ldots, y_k)$ for K collision categories and $(\theta_1, \ldots, \theta_k)$. Defining $DIC_k = \overline{D}_k + p_k$, $p_k = \overline{D}_k - \overline{D}_k(\overline{\theta}_k)$, $\overline{D}_k = E[D_k(\theta_k)|y_k]$, $\overline{\theta}_k = E[\theta_k|y_k]$, $D_k(\theta_k) = -2\ln\{f(y_k|\theta_k)\} + 2\ln\{f(y_k)\}$.
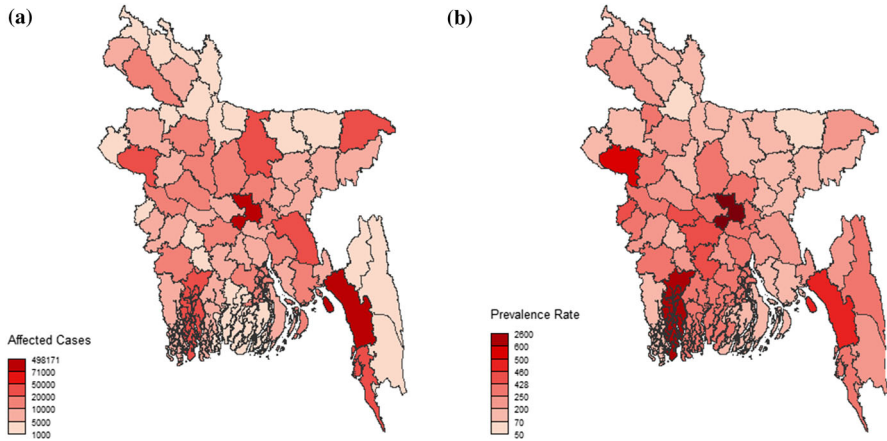
Under priors and independent models, it is found $f(y|\theta) = \prod_{k=1}^{k} f(y_k|\theta_k)$ and $f(y) = \prod_{k=1}^{k} f(y_k)$. These multiplicative conditional and marginal distributions of y contribute additively to the Bayesian deviation Equation 4, resulting in y's extreme value $DIC = \sum_{k=1}^{K} DIC_k$ [51]. A small $\overline{D}$ corresponds to a well-fitted model. If DIC differences were borderline, less complex models with lower p_D were used [50].

2.6 Computational Procedure

RStudio version 4.2.0 uses `moran.test` and `geary.test` (available in `spdep` package) to measure spatial autocorrelation. Before computing these two statistics, `poly2nb` compiles a list of districts that share adjacent boundaries. `nb2listw` adds weights to a neighbor's list. By using `moran.test` p-value is calculated analytically, not by MC. This isn't always significant. A function `moran.mc` can test significance using MC simulation. Local Moran I provides I value, variance, p-value, predicted I, and variation for each district using `localmoran` function.

RStudio's `maptools` package was used to visualize affected cases. For reading shape files `readShapePoly` function is used. Two functions `moran.test` and `geary.test` are used to measure spatial autocorrelation. Before computing two statistics, `poly2nb` function compiles a list of districts that are neighbors based on their adjacent boundaries, meaning they share one or more boundary points. `nb2listw` function adds spatial weights to an existing neighbors list. p-value of `moran.test` is calculated analytically, not by MC. This doesn't always indicate importance. `moran.mc` can test significance using MC simulation. Using `localmoran` function, local Moran I provides its own I value, variance, p-value, predicted I, and variation of I for each district. In this instance, GeoDa with version 1.20.0.10 is put to use in order to track down significant areas of relative risk via a LISA cluster map that employs 999 simulations at a significance level of 10%.

WinBUGS, a statistical software for Bayesian analysis using Markov Chain Monte Carlo (MCMC), is used to perform Bayesian models and spatial data analysis. This software is based on the BUGS (Bayesian inference Using Gibbs Sampling). and it also offers a goodness-of-fit measure called the deviance information criteria, which can be used to compare models [54]. For each model, two separate chains starting from different arbitrary initial values were used to calculate the realized value of posterior estimators in the Bayesian hierarchical model. The dynamic trace plots were used to check the good mixing of two chains with 100000 iterations taken in which 20000 were excluded as a burn-in sample using WinBUGS. In case to improve convergence and reduce the effect of autocorrelation, thin values of 5 were used for testing the convergence of the estimator in spatial modeling.

(a) **(b)**

Affected Cases

498171
71000
50000
20000
10000
5000
1000

Prevalence Rate

2600
600
500
460
428
250
200
70
50

**Fig. 2** **a** District-wise affected number of cases; and **b** District-wise prevalence rate

3 Data Analysis

From 2021 to 2022, Dhaka district, Bangladesh's capital city, had the highest number of cases with 498,171 (see, Fig. 2). The districts surrounding Dhaka have fewer reported incidences. Maintaining adequate safety precautions in a small, densely populated city is impossible. Two districts, Chattogram ranks second and Khulna third for infected cases, correspondingly. Ports, business advantages, improved communication, education, and other amenities drive Chattogram's population growth. As it is the second-largest city, the number of affected cases is also higher. The least number of COVID-19 cases are in Lalmonirhat. Fewer people are affected when there are fewer people in an area. Few people got sick with a virus in the hilly parts of Bandarban. Mountain dwellers can tolerate low oxygen levels and have a virus-free environment, say, researchers. Dry mountain air, high levels of UV radiation, and low barometric pressure combine to create an inhospitable habitat; these conditions, taken together, lower the survival rate of airborne viruses. Those who live in the mountains may benefit from it [54].

Figure 2 demonstrates that Dhaka district has the highest prevalence rate. COVID-19 prevalence varies within districts. Rajshahi district, on the west side of the country, has a lower prevalence rate than Dhaka. The disease is highly prevalent in Khulna district, which is also situated in the south. Although Khulna has a population five times smaller than that of Chattogram, Chattogram has more people affected by the COVID-19 virus than Khulna. Due to the population of Khulna being theoretically disproportionate to relative risk, the calculated relative risk for this city is greater than that of Chattogram. In Faridpur, Gopalganj, and Rajbari districts, prevalence rates are lower than in Dhaka. As the virus evolved and underwent mutations, an increasing number of people contracted the disease and perished. Dhaka has the most people affected by the outbreak, and its dense population makes it vulnerable. Lower prevalence rates have been observed in Bangladesh's northern (Sunamganj) and northeastern (Habiganj) districts. The northern district of Gaibandha has the lowest prevalence.

Table 1 Moran *I* & Geary *C* Statistic are calculated under randomization

Summary statistics	Moran *I*	Geary *C*
Statistic	0.0846	0.8786
Expectation	−0.0161	1.0000
Variance	0.0019	0.0286
Standard deviate	2.2866	0.7172

Density Plot of Permutation Outcomes

Relative Risk
Monte-Carlo simulation of Moran I
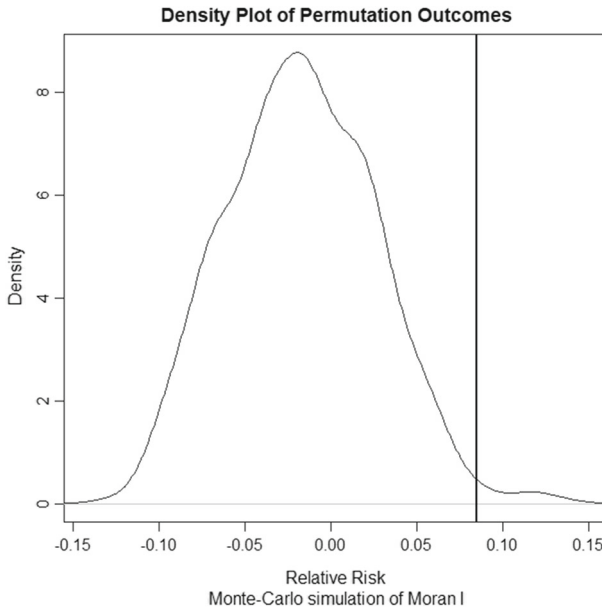
Fig. 3 Density Plot of Global Moran *I*

The result for Moran *I* of COVID-19 relative risk is 0.0846, which indicates positive spatial autocorrelation between districts (Table 1). The obtained *p*-value of 0.0111, less than 0.05, and the corresponding *z*-value of 2.54 also show that the null hypothesis should be rejected. Using MC simulation of 599 global Moran *I* depicts the same *p*-value of 0.0111 at a significance level of 5%. In both cases, the null hypothesis should not be accepted. A further demonstration of how likely the observed test statistic is is provided by a density plot (Fig. 3) of the Monte Carlo permutation outcomes. Moreover, the obtained value of the Geary *C* statistic is 0.8786, which falls within the interval [0,1), indicating the existence of positive autocorrelation between districts. Both of the spatial autocorrelation analysis procedures indicate the existence of clusters between the district of COVID-19 relative risk.

Taking into account the effect of spatial lag and the spatial weights of the districts next to each other, the LISA cluster map in Fig. 4. shows the important districts with weighted spatial homogeneity at a 90% confidence level. This popular choropleth map sorts places with a significant local Moran statistic value from Equation 2 by type of spatial correlation. A bright red color indicates a spatial cluster that is High-High, while
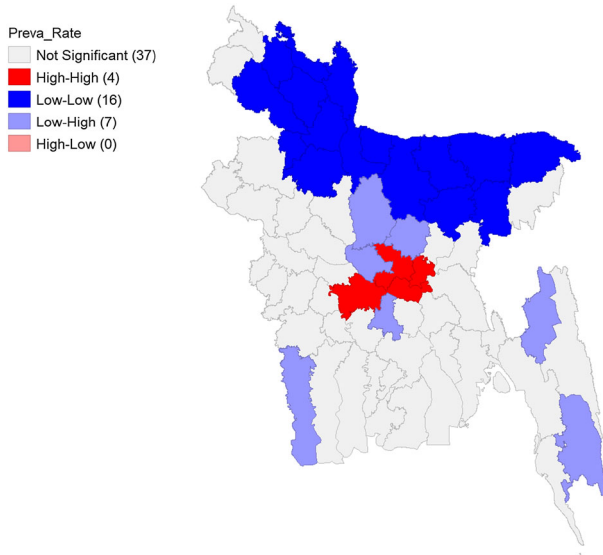
Fig. 4 Spatial Clustering (local Moran's *I*) of COVID-19 Relative Risk

a bright blue color indicates a spatial cluster that is Low-Low. A light blue color indicates a spatial outlier that is Low-High, while a light red color indicates a spatial outlier that is High-Low. Dhaka, Munshiganj, Narayanganj, and Faridpur are the four districts that are shown to form a statistically significant High-High spatial cluster. It shows that the relative risk of COVID-19 is high in these districts, and it also shows that the relative risk is high in the adjacent districts. The districts that meet the criteria for statistical significance and are located in the Low-Low spatial cluster are as follows: Nilphamari, Lalmonirhat, Rangpur, Kurigram, Gaibandha, Dinajpur, Bogura, Joypurhat, Jamalpur, Sherpur, Mymensingh, Netrokona, Sunamganj, Kishoreganj, Habiganj, and Sylhet. It would appear that there is a low relative risk of COVID-19 in these districts, and it would also appear that there is a low relative risk in the districts adjacent to them. Even though Chattogram has a population five times larger than Khulna, Khulna has a lower number of people whom COVID-19 has impacted than Chattogram. The calculated relative risk for Khulna is more significant than that of Chattogram because the population in Khulna is theoretically disproportionate to relative risk. In addition, the districts of Tangail, Gazipur, Manikganj, Khagrachari, Bandarban, Madaripur, and Satkhira are categorized as Low-High spatial outliers. These seven districts give off the impression of having a low relative risk, whereas the districts that surround them typically show a high relative risk.

Table 2 displays the summary statistics of posterior estimators including the 95% credible intervals. Whereas, a credible interval implies that the true parameter would lie within the lower limit and upper limit and we can be 95% confident about that. Figures 6, 9, 12, 15, and 18 (in "Appendix") that the data has a normal distribution for the overall mean. On the other hand, the presence of variance and precision is suggestive of a chi-square distribution.

Table 2 Summary statistics of Poisson-Gamma, Poisson-Lognormal, CAR, convolution and modified CAR models

Posterior estimator	Poisson-Gamma	Poisson-Lognormal	CAR	Convolution	Modified CAR
$\hat{a}$	2.45	–	–	–	–
$\hat{b}$	3.53	–	–	–	–
Mean	0.697	0.5614	0.5632	0.5667	1.317
Variance	0.205	–	–	–	–
$\hat{\alpha}$	–	−0.577	−0.5741	−0.57	0.2605
$\hat{\tau}_v^2$	–	3.36	–	3.693	–
$\hat{\sigma}_v^2$	–	0.307	–	0.2957	–
$\hat{\sigma}_u^2$	–	–	1.173	0.3984	0.01133
$\hat{\tau}_u^2$	–	–	0.8795	419.9	1112.0
95% Credible interval	(0.5960,0.8178)	(0.4954,0.6470)	(0.5615,0.5650)	(0.5073,0.6316)	(0.9062,1.869)
$\overline{D}$	769.7340	769.8500	773.0380	769.6580	775.927
$\hat{D}$	704.8430	704.8000	704.8830	704.8740	761.765
p_D	64.8910	64.9700	68.1550	64.7840	14.162
DIC	834.6250	834.8200	841.1930	834.4400	790.089

According to [55] autocorrelation plot in any figure can "indicate dimensions of the posterior distribution that are mixing slowly, where slow mixing is often associated with high posterior correlations between parameters". Figures 7, 13 (in "Appendix") demonstrates estimators are mixing well and autocorrelation is rapidly disappearing before each case is considered. Hence, no autocorrelation is present here. In contrast, Figs. 10, 16, 19 (in "Appendix") exhibits poor mixing for α and that autocorrelation is not significantly decreasing before each case is evaluated. Therefore, substantial autocorrelation exists for α in this case.

The two main features desired in the trace plots are stationarity and well mixing. For the path to be considered stationary, it must remain inside the posterior distribution.To be more explicit, all of the traces congregate around an extremely consistent central trend. Figures 5, 8, 11, 14, and 17 (in "Appendix") show stable stationarity. The second characteristic of a chain is called "good mixing" which means that each sample in each parameter is not strongly related to the sample that came before it. As the trace moves across the posterior distribution without getting tangled in any one place, each path can be seen to move in a zigzag pattern. The second trait is evident in experimental trace plots. Red and blue chains used both features.

Table 2 shows DIC values for four different models, two of which are non-spatial and two of which are spatial. Rules say that the model with the lowest DIC value provides a superior fit. The modified CAR model has the lowest DIC value compared to the other models, which are almost identical. Even the $\overline{D}$ and p_D values are of a relatively low magnitude. So, the modified CAR model significantly outperformed most other models.

4 Discussion and Conclusions

This research aims to ascertain the degree to which COVID-19 cases differ in their spatial distribution across 64 districts. Four Bayesian hierarchical models, both spatial and non-spatial, were used to verify the heterogeneity of the spatial data. Spatial models help explain geographical differences. These models show that the model's fit is not the same everywhere. The examination of spatial autocorrelation performed at the district level gives data regarding districts' embeddedness and spatial dependency, which conveys that the districts are significantly clustered. Since the number of affected cases is proportional to the relative risk, it is reasonable to expect that there will be fewer affected cases if the relative risk is low. Dhaka was found to have the highest relative risk compared to other districts in Bangladesh. Additionally, there is evidence that Khulna has a high risk, but one that is lower than that of capital. The results of this study show that the risk is also higher in districts with many people, like Chattogram. Overpopulation is a cause of a higher risk, along with fast transmission, lack of safety, and not taking precautions.

This research looks at the overall situation of COVID-19 in each district. The government should spread information and make safety materials available. Keep the cost of preventive measures at a reasonable level. A vaccination campaign must be started to make antibodies in the body. If the disease can be stopped from spreading

in areas with many people, then the number of people who are adversely affected can also be reduced.

In this research, we focused on one response variable without considering other covariates except for the district population's density. The results will be more generalized if we use other related risk factors in the model. Further research will be performed using a Bayesian hierarchical spatial model with other related covariates. The results of this investigation will have led to the discovery of further significant information.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Approval We have conducted ourselves with integrity, fidelity, and honesty. We have not intentionally engaged in or participated in any form of malicious harm to another person or animal.

Consent to Participate Not Applicable.

Consent for Publication Not Applicable.

Appendix

See Figs. 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19.
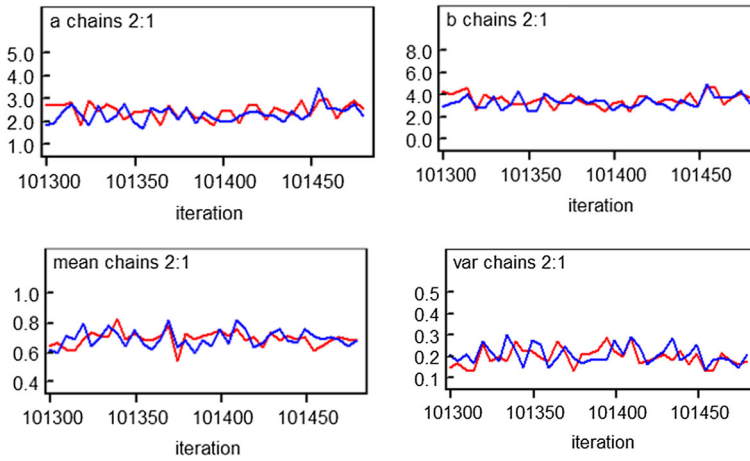
Fig. 5 Dynamic Trace plots for posterior distribution obtained by Poisson-Gamma model
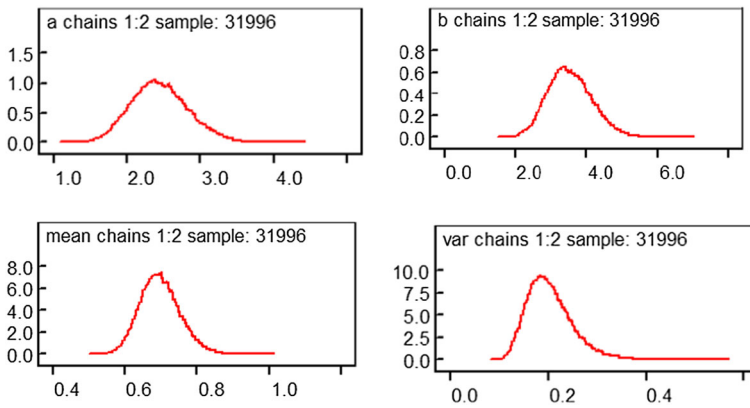
Fig. 6 Kernel Density plots for posterior distribution obtained by Poisson-Gamma model
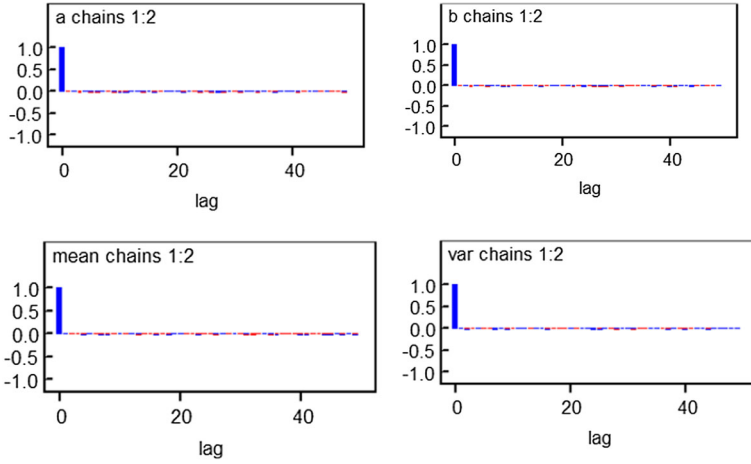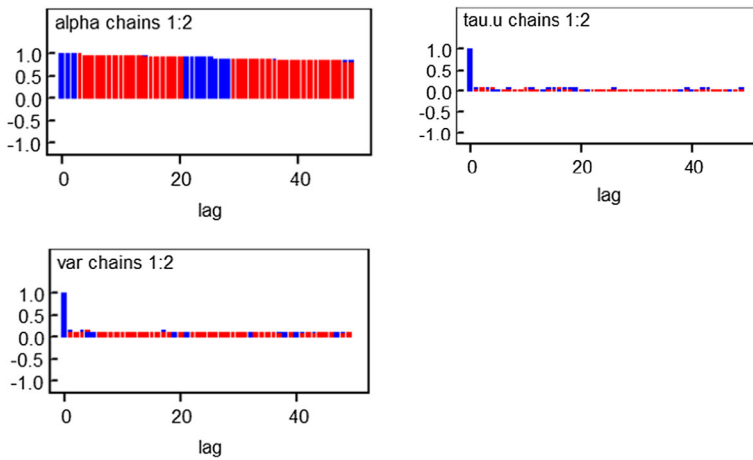
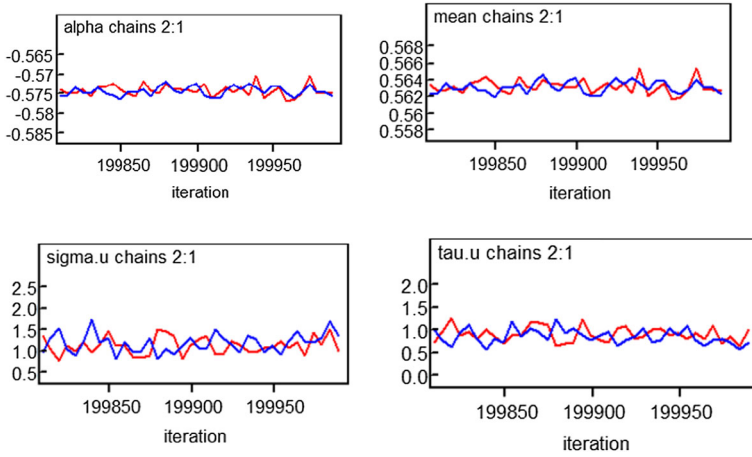**Fig. 7** Auto-correlation plots for posterior distribution obtained by Poisson-Gamma model
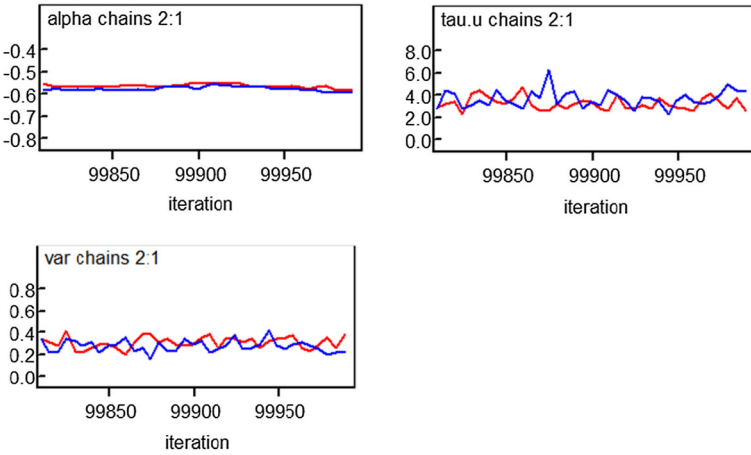
Fig. 8 Dynamic Trace plots for posterior distribution obtained by Poisson-Lognormal model
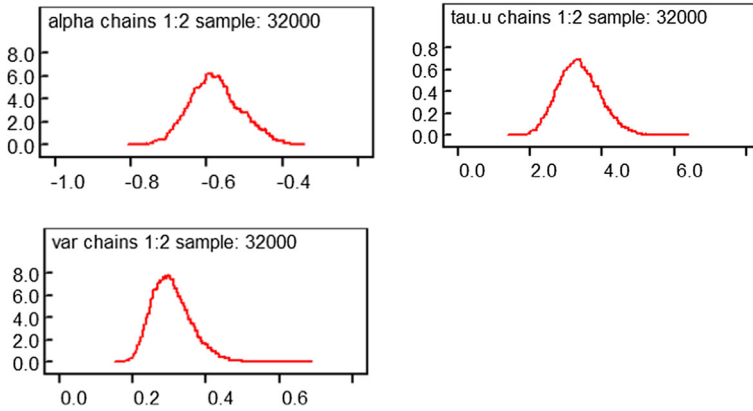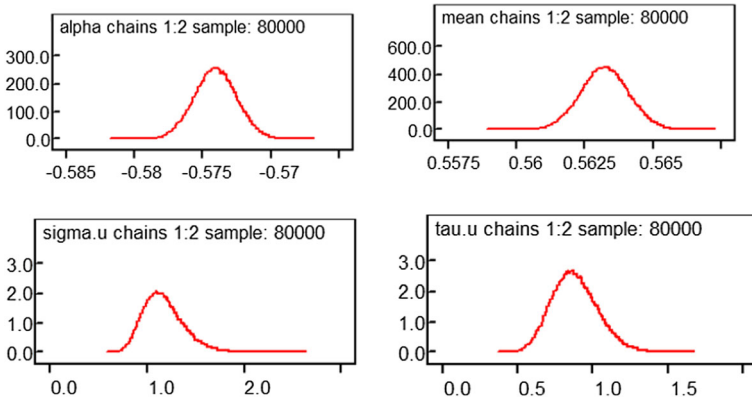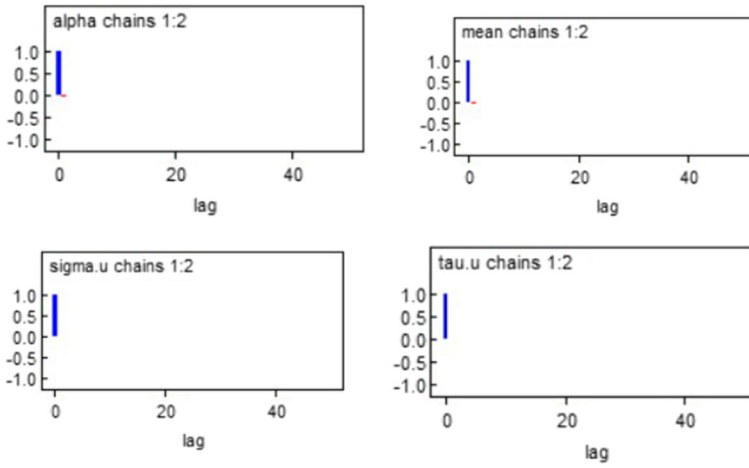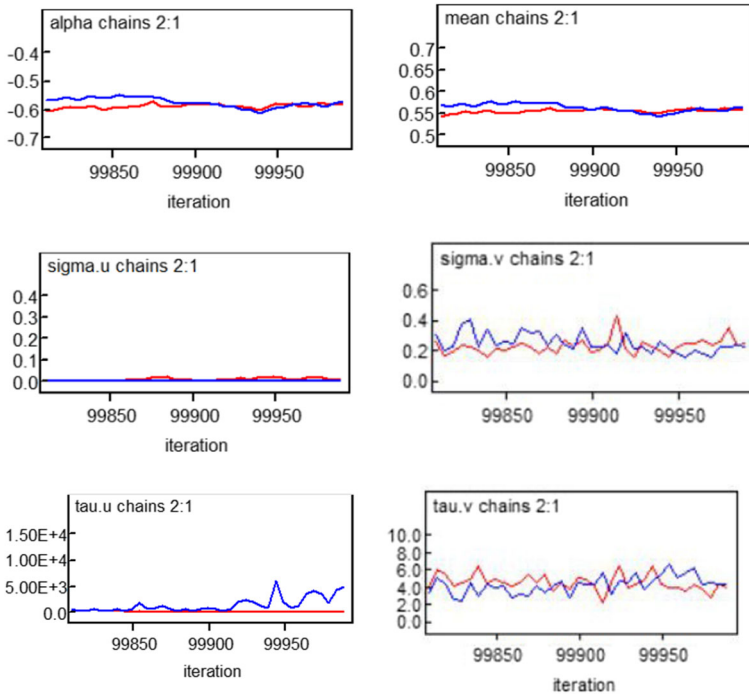
Fig. 9 Kernel Density plots for posterior distribution obtained by Poisson-Lognormal model

Fig. 10 Auto-correlation plots for posterior distribution obtained by Poisson-Lognormal model

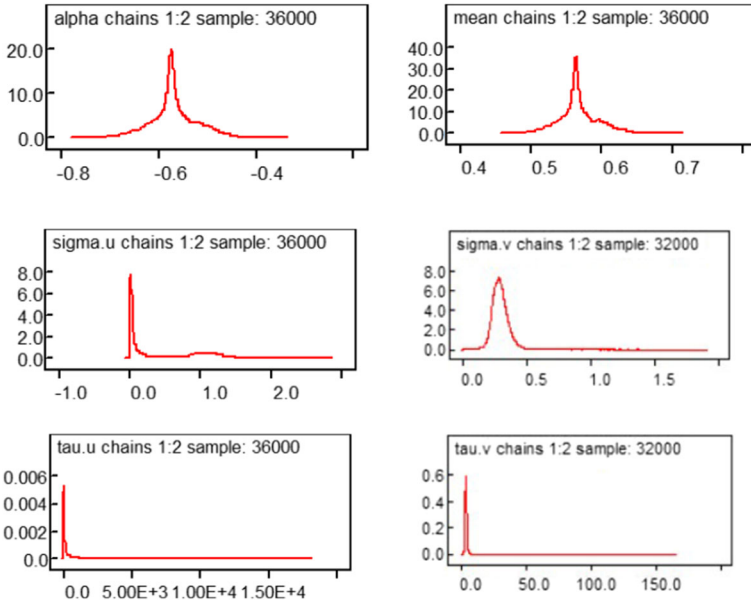**Fig. 11** Dynamic Trace plots for posterior distribution obtained by CAR model

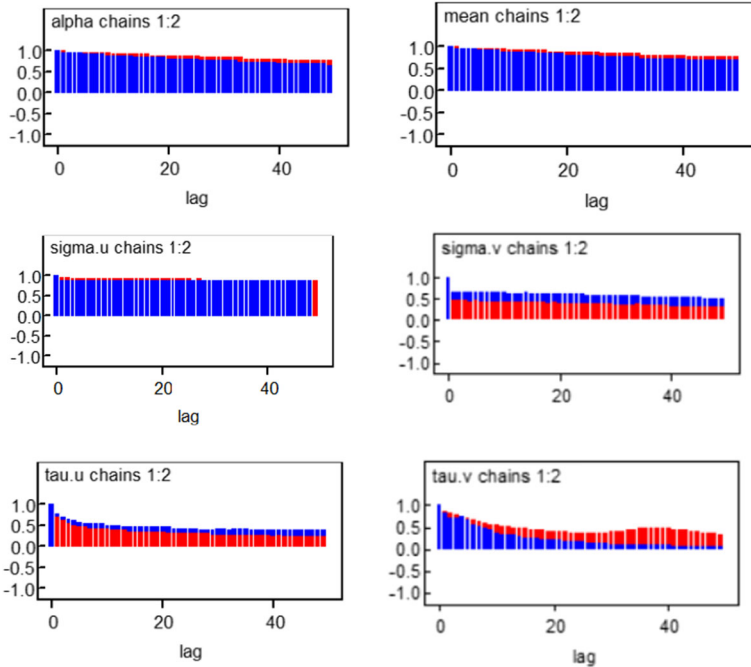**Fig. 12** Kernel Density plots for posterior distribution obtained by CAR model

Fig. 13 Auto-correlation plots for posterior distribution obtained by CAR model

Fig. 14 Dynamic Trace plots for posterior distribution obtained by COV model

Fig. 15 Kernel Density plots for posterior distribution obtained by C.O.V. model

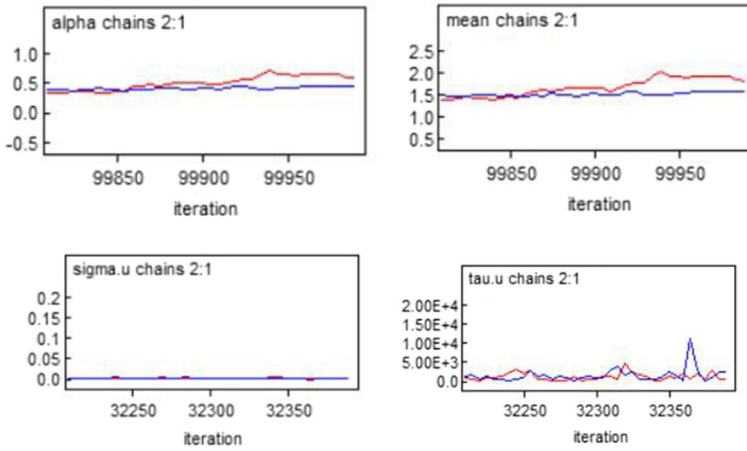**Fig. 16** Autocorrelation plots for posterior distribution obtained by COV model

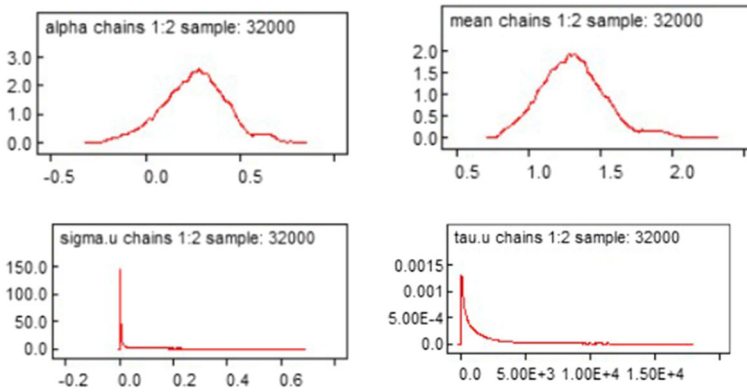**Fig. 17** Dynamic Trace plots for posterior distribution obtained by Modified CAR model

Fig. 18 Kernel Density plots for posterior distribution obtained by Modified CAR model
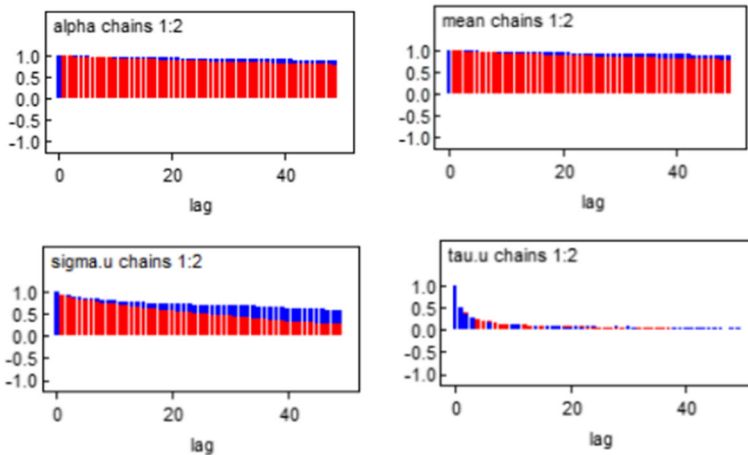
Fig. 19 Autocorrelation plots for posterior distribution obtained by Modified CAR model

References

1. Medicine JH (2022) What Is Coronavirus?, https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus. Online; Accessed 24 February 2022
2. Khandaker Mursheda F (2020) The covid-19 pandemic: Challenges and reality of quarantine, isolation and social distancing for the returnee migrants in bangladesh, Technical report, University Library of Munich, Germany
3. Lai C-C, Shih T-P, Ko W-C, Tang H-J, Hsueh P-R (2020) Severe acute respiratory syndrome coronavirus 2 (sars-cov-2) and coronavirus disease-2019 (covid-19): The epidemic and the challenges. Int J Antimicrob Agents 55(3):105924
4. Sun P, Lu X, Xu C, Sun W, Pan B (2020) Understanding of covid-19 based on current evidence. J Med Virol 92(6):548–551
5. Sohrabi C, Alsafi Z, Oneill N, Khan M, Kerwan A, Al-Jabir A, Iosifidis C, Agha R (2020) World health organization declares global emergency: a review of the 2019 novel coronavirus (covid-19). Int J Surg 76:71–76
6. Analytica O (2020) The who's covid-19 pandemic declaration may be late, *Expert Briefings*
7. Worldometers (2022) COVID-19 Coronavirus Pandemic, https://www.worldometers.info/coronavirus/. Online; Accessed 7 June 2022
8. Karim MR, Akter BM, Haque S, Akter N et al (2021) Do temperature and humidity affect the transmission of sars-cov-2?-a flexible regression analysis. Annals of Data Science 9(1):153–173
9. GOARN (2022) Responding to COVID-19 in Bangladesh: WHO supports the government to roll-out contact tracing across the country. https://shortest.link/8y7q Accessed 19 Nov 2022
10. Worldometer (2022) Bangladesh COVID - Coronavirus Statistics - Worldometer, https://www.worldometers.info/coronavirus/country/bangladesh/. Online; Accessed 7 June 2022
11. Li J, Guo K, Viedma EH, Lee H, Liu J, Zhong N, Gomes LFAM, Filip FG, Fang S-C, Ozdemir MS, et al (2020) Culture versus policy: more global collaboration to effectively combat covid-19. Innovat 1(2):100023
12. Olson DL, Shi Y, Shi Y (2007) Introduction to business data mining, vol 10. McGraw-Hill/Irwin New York, New York
13. Tien JM (2017) Internet of things, real-time decision making, and artificial intelligence. Annal Data Sci 4(2):149–178
14. Kumar S (2020) Monitoring novel corona virus (covid-19) infections in india by cluster analysis. Annal Data Sci 7(3):417–425
15. Shi Y, Tian Y, Kou G, Peng Y, Li J (2011) Optimization based data mining: theory and applications. Springer, Berlin

16. Liu Y, Gu Z, Xia S, Shi B, Zhou X-N, Shi Y, Liu J (2020) What are the underlying transmission patterns of COVID-19 outbreak? an age-specific social contact characterization. Eclinic Med 22:100354
17. Shi Y (2022) Healthcare applications. Advances in big data analytics: theory, algorithms and practices. Springer, New York, pp 643–668
18. Smans M, Muir CS, Boyle P (1992) Atlas of cancer mortality in the european economic community. Atlas of cancer mortality in the European Economic Community
19. Ceylan Z (2020) Estimation of covid-19 prevalence in Italy, Spain, and France. Sci Total Environ 729:138817
20. Sarkar SK, Ekram KMM, Das PC (2021) Spatial modeling of covid-19 transmission in bangladesh. Spat Inf Res 29(5):715–726
21. Coly S, Garrido M, Abrial D, Yao A-F (2021) Bayesian hierarchical models for disease mapping applied to contagious pathologies. PLoS ONE 16(1):e0222898
22. Guliyev H (2020) Determining the spatial effects of covid-19 using the spatial panel data model. Spat Stat 38:100443
23. Desjardins MR, Hohl A, Delmelle EM (2020) Rapid surveillance of covid-19 in the united states using a prospective space-time scan statistic: Detecting and evaluating emerging clusters. Appl Geogr 118:102202
24. Adekunle IA, Onanuga AT, Akinola OO, Ogunbanjo OW (2020) Modelling spatial variations of coronavirus disease (covid-19) in africa. Sci Total Environ 729:138998
25. Sarwar S, Waheed R, Sarwar S, Khan A (2020) Covid-19 challenges to Pakistan: Is GIS analysis useful to draw solutions? Sci Total Environ 730:139089
26. Agresti A (2003) Categorical data analysis. John Wiley and Sons, New Jersey
27. CClayton D, Kaldor J (1987) Empirical bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43(3):671–81
28. Besag J, York J, Mollié A (1991) Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 43(1):1–20
29. Anselin L (1988) Spatial econometrics: methods and models, vol 4. Springer Science and Business Media, Berlin
30. NAC C (1993) Cressie NAC statistics for spatial data, *Probab Math Statist*
31. Molenberghs G, Verbeke G, Demétrio CG, Vieira AM (2010) A family of generalized linear models for repeated measures with normal and conjugate random effects. Stat Sci 25(3):325–347
32. Hayes A (2022) Poisson Distribution, https://www.investopedia.com/terms/p/poisson-distribution.asp. Online; Accessed 19 May 2022
33. Getis A (2008) A history of the concept of spatial autocorrelation: a geographer's perspective. Geogr Anal 40(3):297–309
34. Haining RP, Haining R (2003) Spatial data analysis: theory and practice. Cambridge University Press, Cambridge
35. Diniz-Filho JAF, Bini LM, Hawkins BA (2003) Spatial autocorrelation and red herrings in geographical ecology. Glob Ecol Biogeogr 12(1):53–64
36. Chou Y-H (1995) Spatial pattern and spatial autocorrelation. International conference on spatial information theory. Springer, Berlin, pp 365–376
37. Rosenberg MS, Sokal RR, Oden NL, DiGiovanni D (1999) Spatial autocorrelation of cancer in western europe. Eur J Epidemiol 15(1):15–22
38. Cliff A, Ord J (1973) Spatial autocorrelation. Pion limited, London
39. Tsai P-J, Lin M-L, Chu C-M, Perng C-H (2009) Spatial autocorrelation analysis of health care hotspots in taiwan in 2006. BMC Public Health 9(1):1–13
40. Rogerson PA (2001) Data reduction: factor analysis and cluster analysis. Stat Methods Geogr 2001:192–97
41. Djukpen RO (2012) Mapping the hiv/aids epidemic in nigeria using exploratory spatial data analysis. GeoJournal 77(4):555–569
42. Getis A, Ord J (1992) The analysis of spatial association by use of distance statistics. Geogr Anal 24(3):189–206
43. Anselin L (1995) Local indicators of spatial association-lisa. Geogr Anal 27(2):93–115
44. Tiefelsdorf M, Boots B (1997) A note on the extremities of local Moran's iis and their impact on global Moran's i. Geogr Anal 29(3):248–257
45. Jacquez GM, Greiling DA (2003) Local clustering in breast, lung and colorectal cancer in long island. New York. Int J Health Geogr 2(1):1–12

46. Jeffers J (1973) A basic subroutine for Geary's contiguity ratio. J R Stat Soc Ser D (Stat) 22(4):299–302
47. Kelejian HH, Robinson DP (1995) Spatial correlation: a suggested alternative to the autoregressive model. New directions in spatial econometrics. Springer, Berlin, pp 75–95
48. Kelejian HH, Prucha IR (2010) Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. J Econom 157(1):53–67
49. Neyens T, Faes C, Molenberghs G (2012) A generalized poisson-gamma model for spatially overdispersed data. Spatial and spatio-temporal epidemiology 3(3):185–194
50. Besag J, Kooperberg C (1995) On conditional and intrinsic autoregressions. Biometrika 82(4):733–746
51. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc Ser B (Stat Methodol) 64(4):583–639
52. Lesaffre E, Lawson AB (2012) Bayesian biostatistics. John Wiley and Sons, New Jersey
53. El-Basyouny K, Sayed T (2009) Collision prediction models using multivariate poisson-lognormal regression. Accid Anal Prevent 41(4):820–828
54. ANI (2020) People living on higher altitude less likely to get infected by coronavirus: Study, https://rb.gy/wktvj0. Online; Accessed 14 June 2022
55. GGriffin JE, Steel MF (2007) Bayesian stochastic frontier analysis using winbugs. J Prod Anal 27(3):163–176