ACTA TECHNICA JAURINENSIS
Vol. 17, No. 3, pp. 130-142, 2024
10.14513/actatechjaur.00746

Research Article

Processing Spatial Data for Statistical Modelling and Visualization Case study: INLA model for COVID-19 in Alabama, USA

Getachew Dejene[1, *], György Terdik[2]

[1] Doctoral School of Informatics, University of Debrecen, 4028, Debrecen, Hungary
[2] Department of Information Technology, Faculty of Informatics, University of Debrecen, 4028, Debrecen, Hungary
*e-mail: engidaw.getachew@inf.unideb.hu

Submitted: 02/06/2024 Accepted: 23/08/2024 Published online: 28/08/2024

Abstract: This research emphasizes the visualization of spatial data for statistical modelling and analysis of the relative risk associated with the COVID-19 pandemic in Alabama, USA. We used Bayesian analysis and the Integrated Nested Laplace Approximation (INLA) approach on data ranging from March 11, 2020, to December 31, 2022, which included observed COVID-19 cases, the population for each of the Alabama counties, and a Geographical map of the state. The geographical distribution of COVID-19's relative risk was determined using various spatial statistical techniques, indicating high-risk locations. The study used Besag-York-Mollié (BYM) models to assess the posterior relative risk of COVID-19, and it found a statistically significant average decrease in COVID-19 case rates across the 67 counties evaluated. These findings have practical implications for evidence-based policymaking in pandemic prevention, mitigation, and preparation.

Keywords: COVID-19; Spatial Data; Disease mapping; Bayesian analysis; hot spot

I. INTRODUCTION

Spatial statistics revolves around the fundamental concept of spatial processes, which involves comprehending and modelling the variations of variables or phenomena across different spatial locations [46]. This concept is essential for capturing and analysing spatial dependencies and patterns exhibited by the variable of interest. In the context of the global COVID-19 pandemic, understanding thespread and impact of the virus is crucial for effective decision-making, resource allocation, and public health interventions [26]. Spatial data and modelling techniques provide a powerful approach to gaining insights into the dynamic nature of the pan-demic [49]. Spatial data about COVID-19 goes beyond temporal trends by considering the geographic location and spatial relationships of cases, deaths, and other relevant variables [31]. This encompasses data on COVID-19 cases, hospitalizations, deaths, testing rates, and vaccination coverage collected at various geographical resolutions, such as countries, states, counties, or smaller administrative units. By incorporating the spatial dimension, analysts can examine patterns, clusters, and disparities in the spread and impact of the virus across different regions [16]. Spatial modelling techniques enable researchers to explore and analyse spatial data, facilitating the identification of underlying patterns, assessing spatial dependencies, and predicting. These models consider spatial relationships and autocorrelation, recognizing that nearby locations are likely to exhibit similar values due to shared characteristics or proximity [29, 50]. By considering spatial effects such as the spatial spread of infections or the influence of local contextual factors, spatial models improve prediction accuracy and offer valuable insights for policymakers, healthcare professionals, and the general public [23]. COVID-19 spatial modelling encompasses a wide range of approaches [14]. One commonly used technique is spatial clustering analysis, which identifies areas with concentrated high or low COVID-19 incidence, aiding in targeted interventions and resource allocation. Other modelling approaches include graphically Weighted Regression, which accountsfor spatial heterogeneity in the relationship between Covid-19 outcomes and potential predictors, and spatial auto-regressive models such as spatial lag or spatial error models, which capture spatial dependencies among neighbouring regions [18, 30, 48]. In addition to analysing the spread of the virus,

spatial modelling can also assess the impact of interventions and policies. By integrating spatial data on containment measures, vaccination campaigns, or mobility restrictions, researchers can evaluate their effectiveness and explore spatial variations in out- comes. Spatial statistics, at its essence, entails com-prehending and modelling the variations of variables or phenomena, while also capturing and representing the outcomes or observations linked to spatial locations [28]. The data can be represented as measurements conducted at spatial units within a fixed spatial domain. This domain can be either a continuous surface or a countable collectionof spatial units, such as census tracts or ZIP codes [6]. Areal data is generated through the division ofa fixed geographic region into smaller sub-regions, which act as units for aggregating diverse outcomesor events [7]. This type of data finds applications across various fields, including the assessment of cancer cases in different counties [32], the docu- mentation of road accidents in various provinces [38], and the measurement of the proportion of individuals living below the poverty line in census tracts [40]. Researchers can analyse and understand patterns and trends within specific sub-regions by utilizing areal data, enabling insights and informed decision-making in these respective domains.

In this study, we focus on examining the spatial pattern of Standardized Incidence Rate of COVID-19 observed across the 67 Alabama Counties rather than individual points. The variable of interest rep-resents a suitable summary, such as the number of case rates within each respective area. **Fig. 1** presents the Population distribution of Alabama in 67 counties in, the USA.

1. Related works

The COVID-19 pandemic demanded the quick development and implementation of new statistical approaches for modelling and predicting viral propagation [1]. Bayesian analysis and the INLA method have emerged as useful techniques for dealing with complicated spatial and spatio-temporal data [36, 44, 52].

Bayesian analysis and the INLA method have proven essential in tackling the complex spatial and spatio-temporal data associated with COVID-19. These techniques enable the incorporation of prior information and hierarchical structures, offering robust frameworks to handle the uncertain- ties and variability in epidemiological data. For instance, hierarchical Bayesian models have been applied to COVID-19 case data in Bangladesh to account for spatial autocorrelation, thus identifying clusters of high prevalence and providing more accurate risk assessments [22].

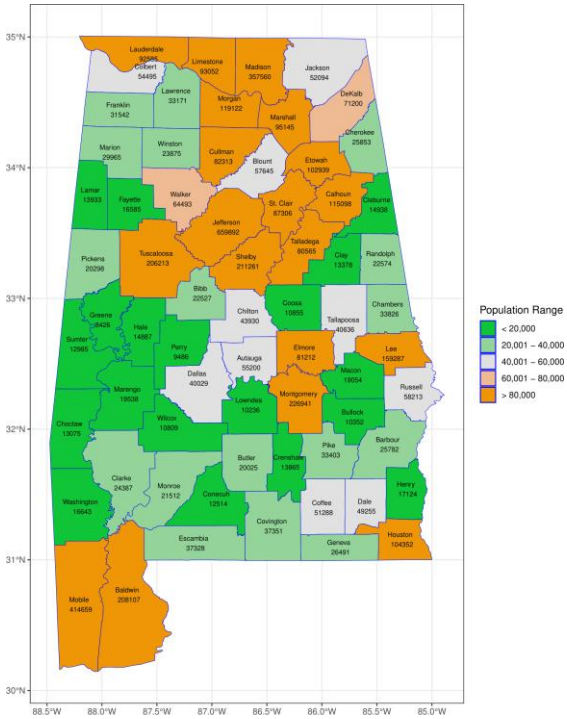


Figure 1. *Population distribution of Alabama per county*

Similarly, in Europe, spatio-temporal Bayesian models have been utilized to study the spread and control measures' impact across Spain, Italy, and Germany, demonstrating how these models can inform public health decisions by capturing the temporal dynamics and spatial heterogeneity of the pandemic [20]. Furthermore, addressing data reporting issues, hierarchical Bayesian models correct misreporting in the U.S., enhancing the reliability of spatial risk estimates [10]. As far as the authors are aware, this research represents the pioneering effort in utilizing visualization techniques for spatial data in the con- text of statistical modelling and analysis of the relative risk linked to the COVID-19 pandemic specifically in Alabama, USA, using Bayesian analysis coupled with the INLA method.

2. Spatial Data

Spatial data can be described as outcomes or observations of a random process that is associated with specific spatial locations [5, 46]:

$$Y \equiv \{y(s), s \in D\} \qquad (1)$$

where Y represents a set of measurements conductedat the spatial units' $s \in D$. The subset D in R^d (d = 2 in this context) establishes the spatial domain. Using the characteristics of the domain D, spatial data can be categorized as areal (or lattice) data, geo statistical data, or point patterns data.

Areal or lattice data is a type of spatial data that emerges when a specific geographic region, known as a fixed domain, is divided into a finite number

of sub-regions [3]. These sub-regions serve as units for aggregating various outcomes or events. Areal data find applications in a range of fields and can be found in diverse contexts. Understanding the data's structure is crucial because specific analytical methods are better suited for certain data types.Being aware of the data's characteristics helps in selecting the most suitable analytical approaches.

In the case of data referring to areas, the location ofeach object needs to satisfy an agreed convention [19]. If the areas are irregular shapes, then one options is to select a representative point such as thearea or population-weighted centroid, and then use the same procedure as for a point object to provide s_i. Alternatively, each area can be labelled, and alookup table can be provided to match rows of the data matrix to the corresponding areas on the map.

In disease mapping, the fundamental situation involves utilizing spatial data specifically related to distinct, non-overlapping n sub-regions [12]. A few examples of areal data include the count of cancer cases in different counties, the number of road accident reported in various provinces, and the proportion of people living below the poverty line in census tracts [34], etc. In each case, the fixed domain,such as a county, province, or census tract, is divided into smaller sub-regions, enabling the aggregation of relevant information within those subregions. In general, data related to a specific area are often observed and recorded within spatially aggregated domains, such as administrative geographies like postcodes, counties, or districts.

Rather than focusing on specific locations, Our study encompasses all 67 counties in Alabama, utilizing a dataset obtained from the official Kag- gle repository source. The dataset contains daily-updated information on reported cases and deaths in the United States, documented at both state and county levels. The dataset consists of two primary CSV files: 'covid_us_county.csv,' containing columns such as fips, county, state, latitude, long, date, cases, state-code, and deaths; and 'us_county.csv,' featuring columns like fips, county, state, state-code, male, female, median-age, population, female-percentage, latitude, and long. Addition- ally, the collection includes US county shape files for geospatial plots in formats like 'us_county.shp,' 'dbf,' 'prj,' and 'shx'.

However, the earliest reported incidents in the original dataset traced back to January 22, 2020. For the scope of our study, which centred on the state of Alabama, we rigorously filtered the information, yielding a total of 1,024 records. During the preprocessing and cleaning step, certain entries were eliminated, notably those with duplicate geo-identifier fields and unsigned county values. Our investigation is limited to the period between March 11, 2020, and December 31, 2022, encapsulating the period with the first nonzero values. The refined dataset will be employed in our analysis, emphasizing key fields such as county (the English name for the county), longitude, latitude (geographic co- ordinates of the region's centroid), cases (number of confirmed COVID-19 cases), population (population of the county), and geometry (polygon de- scribing the geographical area).

3. Data Processing and plot Generation for COVID-19 Dataset Analysis

In the data processing and categorization phase of our COVID-19 dataset analysis, we utilized the R programming language, leveraging key packages like "Simple Features in R" (sf), more over a set of packages called "tidyverse" that share a high-level design philosophy and low-level grammar and data structures [54]. The "sf" package facilitated spatial data manipulation, notably through the `st_sf()` function, creating spatial data frames with seamless geographic integration. Functions like `st_read()` and `st_write()` ensured efficient reading and writing of spatial data in various formats. Concurrently, the "tidyverse" package, encompassing vitalR packages, streamlined general data manipulation tasks. Functions from "dplyr" within the tidyverse,such as filter (), mutate (), and summarize (), played acrucial role in non-spatial data processing, ensuring a consistent and efficient methodology. The technical integration of "sf" and "tidyverse" contributed to a well-structured approach in handling both spatialand non-spatial aspects of the COVID-19 dataset.

Understanding the organization of COVID-19 data frames is crucial for accurately analysing and interpreting pandemic-related information. **Fig. 2** illustrates the fundamental structure of the COVID-19 data frame and the method for man- aging Alabama county shapefiles.

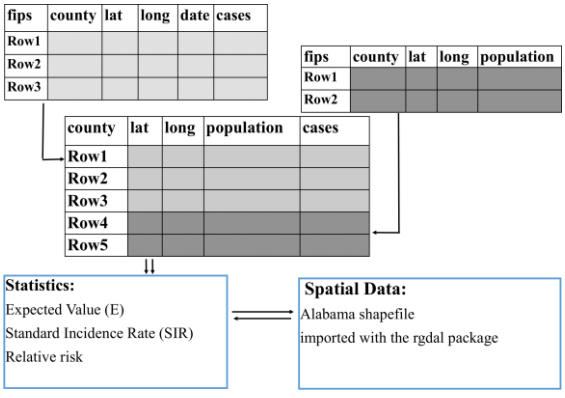


Figure 2. *Structure of Covid-19 Data frame and Alabama county shapefile process.*

Additionally, it highlights the key statistical functions utilized for data analysis. In this process, we merge these two tables using a shared key aligned with our researchobjectives. This merging approach

allows us to integrate pertinent data from both sources, facilitating a more comprehensive investigation and interpretation of COVID-19 trends at the county level. By amalgamating data from the COVID-19 data frame with Alabama county shapefiles, we are equipped to get insights into the geographical distribution of cases, demographics, and other relevant factors influencing the pandemic's trajectory.

To enhance the presentation of selected data such as COVID-19 cases, expected cases, SIR, and relative risk, we apply a statistical technique:

$$gr[1] = min + \frac{median - min}{2.5} \quad (2)$$

$$gr[2] = min + 2\left(\frac{median - min}{2.5}\right) \quad (3)$$

$$gr[3] = max - 2\left(\frac{max - median}{2.5}\right) \quad (4)$$

$$gr[4] = max - \frac{max - median}{2.5} \quad (5)$$

following the consolidation of all data frames. This technique involves defining groups (denoted as gr) using quartile splits, incorporating statistical measures like minimum (min), maximum (max), and median.

The vector "gr" plays a pivotal role in capturing these quartile breakdowns, facilitating more de-tailed data segmentation and analysis. Quartile borders are determined systematically as follows: - The lower boundary of the first quartile, gr[1], is computed by adding a fraction of the range between thelowest and median to the minimum value. The up- per boundary of the first quartile, gr[2], expands proportionally on that range. Similarly, the lower boundary of the third quartile, gr[3], is calculated by subtracting twice the proportion of the range be-tween the maximum and median from the highest value. In contrast, the upper boundary, gr[4], is determined by deducting the fraction of that range. The resulting rounded values of gr (rounded tothree decimal places) lead to the segmentation of the dataset into five distinct groups: "Very Low," "Low," "Medium," "High," and "Very High," basedon the distribution of "confirmed cases" versus "expected cases". This approach ensures a fair and meaningful classification, contributing to a comprehensive understanding of the variables' distribution within the dataset.

II. METHODOLOGY

The dataset, sourced from Kaggle, underwent initial filtration to focus specifically on Alabama counties ties. Subsequently, relevant information about population, COVID-19 cases, and counties was systematically extracted and organized. To enhance data interpretability, temporal considerations were integrated, limiting the dataset to the period from March 11, 2020, to December 31, 2022. This inspection will endorse the statistical analysis method that will be useful in summarizing the information in the data set.

In **Table 1** statistical summary of the data is presented. From 2020-03-11 to 2022-12-31, for 147 weeks an average of 23417 people were infected with COVID-19 in different counties of Alabama, USA. According to the data, the maximum number of cases is 225876 and the minimum is 2196 registered in counties Jefferson and Greene accordingly. It is worth mentioning that these latter statis-tics also depend on the counties' population. These counts are influenced by both the size and demographic makeup of the populations residing in each area. The relative COVID-19 cases $r_i = Y_i/Pop_i$ provide more specific information according to the counties where Pop_i is the population of county i. Now let us assume that the COVID-19 has uniformly spread throughout the state. Then the number of cases in the state is proportional to the population of the state with the ratio

$$\rho = \sum_{i=1}^{67} Y_i / \sum_{i=1}^{67} pop_i \quad (6)$$

i.e., the rate ρ is calculated by dividing the total number of cases by the state's total population. We have ρ = 0.32%. To address the influence of the counties, the expected numbers of dis- ease risk E1, ..., E67 are determined through indirect standardization. The expected counts Ei for each county i, where i = 1, ..., 67 is computed as:

$$E_i = \rho \times pop_i \quad (7)$$

We normalize the relative cases such that we divide it with ρ, i.e. $r_i/\rho = Y_i/(\rho Pop_i)$, this latter quantity is Y_i/E_i and is called Standard Incidence Rate (SIR) [44]. We concentrate on modelling and investigating the SIR of COVID-19. The SIR is a straight- forward metric used to assess disease risk in specific areas [33, 34]. It is calculated as the ratio between the number of observed cases Yi and the number of expected cases Ei in the i^{th} area,

$$SIR_i = Y_i / E_i \quad (8)$$

Thus, an area with an $SIR_i > 0.893$ corresponds to a high-risk area as there are more cases observed than expected. On the other hand, an area with an

Table 1. *Summary for COVID-19 cases data per county in Alabama State*

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*	*Var*
2194	5634	10357	23417	28833	225876	225876

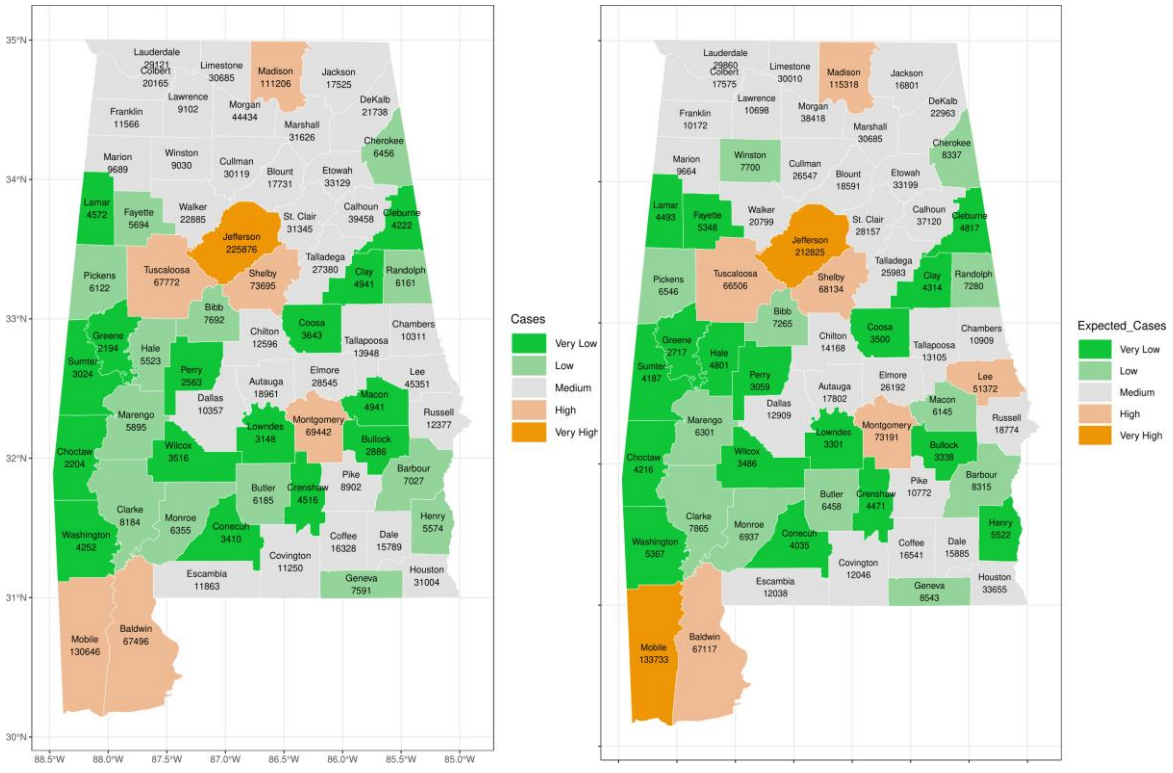
$SIR_i < 0.708$ corresponds to a low-risk area. Figure 4 and 5. provides a visual representation of the SIR map. However, the SIR becomes an unreliable measure of disease risk, particularly when dealing with rare diseases or small populations at risk, resulting in small values for the expected number of cases E_i. Due to this instability, researchers often opt for an alternative approach to estimate risk by utilizing model-based methods [11, 13].

To visually represent each group, a custom colour palette was defined. Utilizing the ggplot2 pack- age, we crafted an informative choropleth map, employing distinct colours to distinguish and delineate geographical locations based on their respective value of case groups.

Improved interpretabilitywas achieved by adding labels to the map using the `geom_sf_text` function. This methodology facilitated the creation of insightful plots, as exemplifiedby **Fig. 3** and **Fig. 4**, offer a comprehensive dataset representation.

Fig. 3a and **3b** provide insightful visualizationsof the pandemic's impact across diverse counties in Alabama. Each county is represented by a shaded area, with colours indicating the disparity between actual confirmed actual cases and expected cases. The accompanying legend serves as a reference, establishing a connection between colours and specific categories of cases and expected cases. Counties shaded in green, like Choctaw, Washington, Wilcox, and Clay, indicate a "Very Low" number of expected cases compared to the actual confirmed cases. Conversely, yellow areas, covering counties such as Pickens, Randolph, and Geneva, signify a "Low" number of actual cases compared to the expected ones. White regions, encompassing counties like Chilton, Lee, Coffee, and others, signal a "Medium" difference between observed and expected cases. Counties shaded in light blue, such asMadison, Tuscaloosa, Shelby, and others, demonstrate a "High" difference, while areas shaded in pink, including counties like Jefferson and Mobile, are associated with "Very High" disparities in the numbers of actual confirmed and expected cases.

The visual presentation depicted in **Fig. 4** offers a comprehensive overview, facilitating the identification of regions exhibiting heightened SIR on the left **Fig. 4a** and, juxtaposed with the posterior relative risk on the right **Fig. 4b**, about COVID-19. This aids in grasping the spatial spread of disease prevalence and plays a pivotal role in pinpointing hotspots or areas of particular concern. Specifically, nine counties: -Colbert, Franklin, Morgan, Winston, Cullman, Walker, St. Clair, Clay, and Hale have been classified as hotspots due to their SIR values exceeding 0.893 on the SIR map. Similarly, al- most identical results were achieved using the same classification threshold value, except for two counties Choctaw and Russe which were classified as low-risk counties based on SIR but as high-risk hotspots in the posterior risk maps.



(a) *(b)*

Figure 3. *Map of Alabama: Covid-19 Confirmedcases (a) and Expected cases (b) per Counties.*

(a) *(b)*

Figure 4. *Visualization of SIR (a) and postier risk ξ (b) for cases per Counties, Alabama.*

This highlights the effectiveness of utilizing both SIR and posterior relative risk in tandem for dis- ease surveillance and hotspot identification. The SIR values provide insight into the standardized incidence rates, enabling comparisons across regions while considering population differences. Conversely the posterior relative risk offers a nuanced understanding of the risk of disease occurrence in one area compared to a reference group, allowing for a more localized assessment of risk. In this case, the discrepancies between SIR and relative risk classifications for Choctaw and Russe counties underscore the importance of considering multiple metrics for a comprehensive understanding of disease distribution and risk assessment.

1. Model

In epidemiology, disease mapping has a long history, and one of its main objectives is to look into the spatial distribution of disease burden [15, 27]. At the county level, the BYM model [2, 4, 39] was utilized to investigate the geographical

distribution of the SIR of COVID-19 connected to Relative risk It is a widely used spatial model that acknowledges the spatial correlation of data and recognizes that neighbouring areas exhibit greater similarity than distant areas [6, 10, 44, 49].

This model incorporates a spatial random effect that smooths the data based on a neighborhood structure. Additionally, it includes an unstructured exchangeable component that captures uncorrelated noise [34,51]. To visualize and understand the relationships between variables and COVID-19 relative risk, we consider COVID-19 dataset with spatially referenced dataset with spatially referenced observations Yi at location i, and let S be the set of all locations in Alabama state. The Poisson distribution is commonly employed as a standard model for count data [21, 47]. It serves as the foundation for many of the count models utilized by analysts today within a hierarchical Bayesian framework [9, 35]. The Bayesian spatiotemporal model is critical for assessing disease propagation and identifying places with high incidence rates across time and space [25]. This model, which incorporates the susceptibility infection recovery paradigm, enables a complete examination of illness trends within populations. It successfully considers a variety of factors that influence illness prevalence, including physical geographical components such as temperature, rainfall, and air pollution, as well as socioeconomic elements such as economic indicators, healthcare accessibility, and demographic characteristics. The applications of Bayesian spatiotemporal models are numerous and important [53]. Firstly, in dis- ease surveillance, these models provide real-time risk assessments and dynamically monitor disease spread. Secondly, they

assist in epidemic forecasting by simulating spatiotemporal disease patterns, enabling accurate predictions and timely interventions. Numerous case studies have validated the effectiveness of these models [42, 43, 45]. The BYM model assumes that the observed counts follow a Poisson distribution:

$$Y_i \sim \text{poisson}\,(\lambda i) \tag{9}$$

where λ_i is the expected rate at location i. The key feature of the BYM model is the decomposition of the expected rates $\lambda_i = \rho_i E_i$ where ρ_i corresponds to the relative risk in area i. Here E_i denotes the expectation of the number of cases for each area and acts as an offset to the Poisson model. In this case, the linear predictor is defined on the logarithmic scale

$$\eta_i = \log(\rho_i) = a_0 + u_i + v_i \tag{10}$$

such that $\rho_i = exp\ (\alpha + u_i + v_i)$. In this equation α represents the average rate across all areas, u_i is the spatially structured residual, and v_i is an unstructured exchangeable component that is modeled as independent and identically distributed normal variables with zero mean and variance σ^2. In the BYMmodel, the spatially structured residual, u_i, of (6) is modeled using the intrinsic conditional on neighbors u_{-i} autoregressive (iCAR) specification

$$u_i \mid u_{-i} \backsim Normal\left(\mu_i + \sum_{j=1}^{n} r_{ij}(u_i - u_j), s_i^2\right) \tag{11}$$

In the context of equation (6), μi represents the mean value for area i, and s2 = σ 2/Ni corresponds to the variance within the same area. The variance depends on the number of neighbours Ni that an area has, meaning that if an area has a larger number of neighbours, its variance will be smaller [37]. This variance structure acknowledges that when there is a strong spatial correlation, areas with more neighbours contain more information in the data regarding the value of their random effect. The variance parameter σ 2 controls the amount of variation be- tween the spatially structured random effects. The value of ri j represents the spatial proximity betweenareas and can be computed as:

$$r_{ij} = \begin{cases} 1/N_i, & if\ areas\ i\ and\ j\ are\ neighbours \\ 0, & otherwise \end{cases} \tag{12}$$

It is important to note that rii is set to 0. County specific relative risks of cases are estimated as ζi = eα+ ui + vi. The above formulae of the BYM model are used for mapping COVID-19 incidences and are implemented using R-INLA see [24, 36, 45].

2. Integrated Nested Laplace Approximation

The INLA is a powerful computational approach designed for approximate Bayesian inference in complex hierarchical models, particularly when dealing with latent Gaussian models [41]. Thecore idea behind INLA is to use a combination of nested Laplace approximations to efficiently compute posterior distributions without resorting to traditional, often computationally expensive, MarkovChain Monte Carlo (MCMC) methods [17]. INLA has gained popularity due to its ability to handle high-dimensional problems and provide accurate approximations quickly [34, 41]. It is particularly well-suited for spatial and spatiotemporal models, allowing for the analysis of complex data structures in fields such as epidemiology, environmental science, and disease mapping. INLA's flexibility and efficiency have made it a valuable tool in various application areas, including risk assessment, ecology, and public health, where researchers of- ten needs to model intricate dependencies and un- certainty in data.

A practical example of INLA's application isin analysing the spread of infectious diseases like COVID-19 [36]. For instance, public health officials seeking to understand the geographic distribution of COVID-19 cases within a state like Alabama might consider factors such as population density, healthcare access, and socioeconomic conditions. By employing the INLA model, they can construct a spatial regression framework that incorporates these variables while accounting for spatial relationships between neighbouring regions.

3. Dataset Analysis

In this study, the BYM model [8] was used to explore the spatial distribution of COVID-19 risk in Alabama. We estimated the relative risk (RR) of COVID-19 incidence for each county in Alabama state and compared it to the SIR and RR, which were used as the baseline reference, and calculated95% credible intervals (CrI). The RR was significantly higher than 1 when the 95% CrI was over 1. A map of the incidence patterns or probability riskwas then generated using the RStudio 2023.06.2 version.

III. RESULTS

This study was initiated to investigate the relative risk of COVID-19 cases in Alabama counties and yields significant findings.

This outcome implies an average decrease of 4.3% in the COVID-19 cases rate across the 67 surveyed counties. The incorporation of a 95% credibility interval enhances precision, supplying a range within which we can confidently assert that the ex- potentiated intercept's true value exists. The interval, ranging from 0.927 to 0.987, corresponds to a 95% probability that the real impact lies within this bracket. To be more specific, the interval suggests a potential reduction in the COVID-19 cases rate, spanning from 7.3% to 1.3%. In essence, the combined implications of the posterior mean and credibility interval indicate a statistically substantiated average decline in the relative risk of

COVID- 19 across Alabama counties, providing valuable insights for shaping public health decisions and policies. In the risk classification process, it is crucial to prioritize areas for intervention and address heightened morbidity and mortality in future outbreaks by examining the geographic spread of extreme relative risks. To achieve this goal, we utilize a combination of data manipulation and visualization techniques to develop an informative and visually captive risk. To pinpoint regions with increased, occurrence of a specific phenomenon, we adopt criteria rooted in exceedance probability. The probability that the relative risk of area i is higher than a value c can be written as P (ρi > c). This probability can be calculated by subtracting P (ρi c) to 1 as follows:

$$P(\rho_i > c) = 1 - P(\rho_i \leq c) \tag{13}$$

To compute the probability P (ρi c) in R-INLA, use the inla.pmarginal() function with ρi's marginal distribution and c as the threshold value. The spatial exceeding probability is calculated using the posterior distribution of the relative risk. This analytical metric gives useful information about the likelihood of the calculated posterior relative risk exceeding a predefined threshold value inside a specific area.

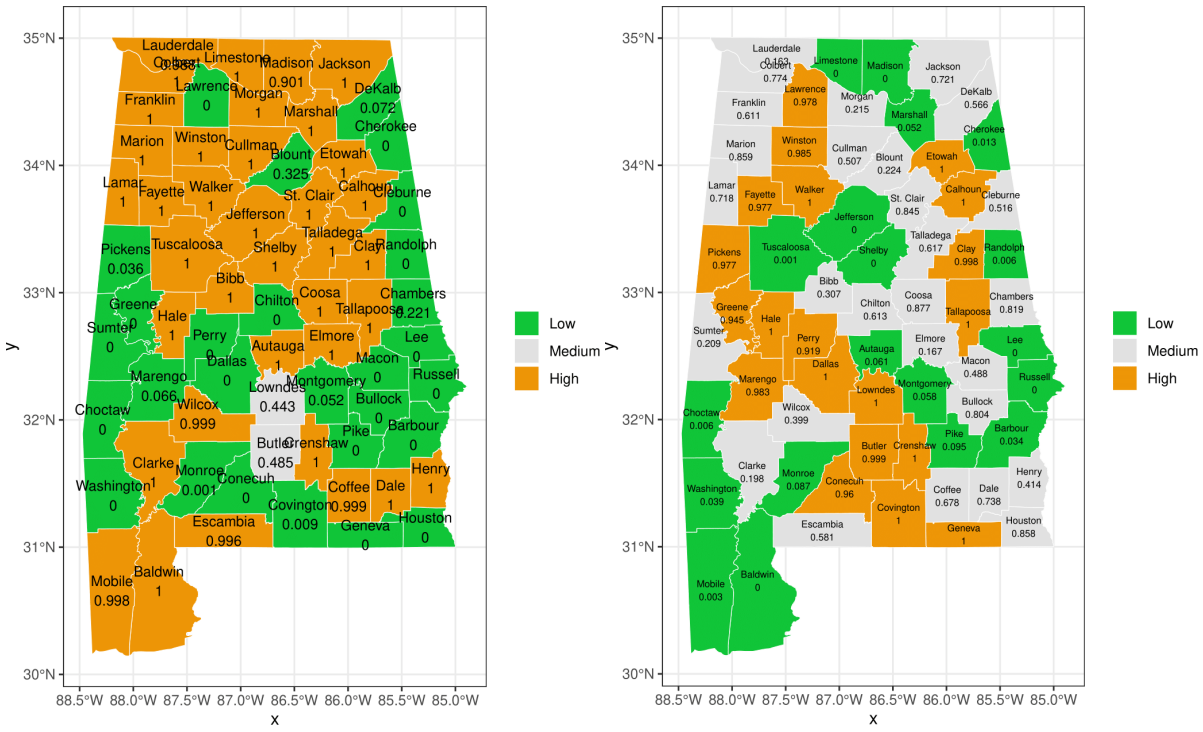
Fig. 5 illustrates the COVID-19 county SIR (Standard Incidence Rate) trends about Alabama mortality rates from March 11, 2020, to December 31, 2022. This graphic has five unique color-coded groups. Some Alabama counties, including Walker, Etowah, Hale, Lowndes, and Crenshaw, have considerably higher relative risks of COVID-19 incidence and mortality rates, above the baseline reference of > 1, placing them in the orange category. In contrast, about four counties, including Madison, Shelby, Lee, and Russell, had much reduced relative risks in both COVID-19 instances and mortality rates and so were allocated to the green group. It's worth noting that the remaining counties typically fell somewhere in between these extremes, as indicated by the plot's legendary different colour ranges. Identifying hotspots or regions of danger is crucial for evidence-based policymaking since these sites act as epicentres, playing an important role in the spread of phenomena, such as diseases or other occurrences.

Fig. 6 presents hotspots with posterior probability $P\ (e^{\rho_i} > 1\ y)$ of relative risk of COVID-19 cases versus deaths. The figure shows that the bulk of relative cases hotspots emerged in the Northern and central counties of Alabama, including Lauderdale, Madison, Jackson, Franklin, Jefferson, and Shelby, to mention a few.

Furthermore, certain northern and central counties nearby (though not neighboring) continued to be identified as hotspots for death risk, albeit in fewer numbers. These counties include Pickens, Greene, Lawrence, Etowah, Calhoun, Covington, among others. **Fig. 6** shows the discrepancy between counties where the relative risk and likelihood of death are greater than one.

(a) *(b)*

Figure 5. *Visualization of SIR (a) and posterior risk ρ (b) of deaths per Counties, Alabama.*



(a) *(b)*

Figure 6. *Visualization of posterior probability$P\ (e^{\rho_i} > 1\ \mathbf{y})$ of relative risks, per Counties, Alabama.*

Using relative risk as an illustration, **Fig. 6a** and **6b** demonstrates that the county at the center has a relative risk surpassing one (depicted in orange), yet it is not classed as a hotspot (marked by a green hue). This is because its exceedance probability is less than 0.398, hence it does not qualify as a hotspot.

IV. DISCUSSION

The fundamental goal of disease mapping is the identification of high-risk locations, which is essential for formulating effective public health strategies. The consequences of a disease mapping model inaccurately predicting cases or deaths in these high-risk regions could lead to misaligned resource allocation decisions that do not address the actual health needs. Additionally, understanding areas with distinctly low risk is crucial, not only for optimizing resource allocation but also for discerning environments that foster a reduction in health risks.

This study was centred on spatial data processing for statistical modelling and visualization, specifically utilizing the BYM model in R-INLA package for COVID-19 in the case of Alabama. We focused on estimating the relative risk of COVID-19 across 67 counties, underscoring the critical importance of accurate disease mapping in public health endeavours. The results of our analysis, particularly the posterior mean of the exponentiated intercept, revealing a substantial 4.3% decrease in the COVID-19 case rate, offer indispensable insights. The 95% credibility interval of 0.927 to 0.987 linked with our findings enhances the robust-ness of our estimations, presenting a nuanced rangeof 7.3% to 1.3%.

V. CONCLUSION

In conclusion, this study leveraged spatial data processing, Bayesian analysis, and advanced statistical modeling, specifically employing the INLA model, to investigate the relative risk of COVID-19 cases across Alabama counties. The calculated posterior mean of the exponentiated intercept α value for relative risk indicated a statistically significant average decrease of 4.3% in the COVID-19 risk rate. The incorporation of a 95% credibility interval (0.927 to 0.987) added precisionto the findings, providing a range within which thetrue value of the exponentiated intercept is likelyto exist, suggesting a potential reduction in theCOVID-19 cases rate ranging from 7.3% to 1.3%. It is worth noting that the posterior mean of the exponentiated intercept $\alpha = 1.1126$ with credibility interval (1.0387 to 1.1903). This implies that $\alpha > 1$ is significant, contrary to the cases when it is smaller than 1.

These results hold substantial implications for public health decision-making, guiding policymakers in prioritizing areas for intervention based on the relative risk distribution. The classification of risk, depicted in the visual representation of COVID-19 relative risk patterns

across Alabama counties from March 11, 2020, to December 31, 2022, revealed notable disparities. Counties such as Colbert, Franklin, Morgan, Winston, Cullman, Walker,St. Clair, Clay, and Hale exhibited significantly higher relative risks, categorizing them in the orange color group. In contrast, several counties witha relative risk range below 0.398 were designated as green, indicating lower risk. Notably, counties like Lowndes and Butler fell within the moderate risk range (0.398 to 0.796), as depicted by the white hue.

This comprehensive analysis, combining statistical insights with visual representations, contributes valuable information for proactive public health measures. By identifying regions with elevated risk, authorities can strategically allocate resources, implement targeted interventions, and mitigate the im- pact of future epidemics. The integration of spatial data processing and visualization techniques enhances our understanding of the geographic distribution of relative risk, fostering informed decision- making for effective public health management.

NOMENCLATURE

ρ	The autocorrelation
BYM	Besag, York, and Mollié
E	Expected value of a random variable,
inunite	of a random variable.
GIS	Geographical Information Systems
POP	Population of county.
RR	Relative Risk.
SIR	Standardized Incidence Rate.

ACKNOWLEDGEMENT

This research by Gy. Terdik was supported by the project TKP2021-NKTA of the University of Debrecen, Hungary. Project no. TKP2021-NKTA-34 has been implemented with support from the Ministry of Culture and Innovation of Hungaryfrom the National Research, Development and Innovation Fund. financed under the TKP2021-NKTA funding scheme.

AUTHOR CONTRIBUTIONS

A. D. Getachew: Writing the manuscript, Writing thecode, and theoretical analysis.

B. Gy. Terdik: Conceptualization, writing the code, Review and editing. Supervision.

DISCLOSURE STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ORCID

D.Getachew http://orcid.org/0000-0001-7547-943x

G. Terdik http://orcid.org/0000-0002-9663-6892

REFERENCES

[1] T. Alamo D. G. Reina, P. Millán. Data-driven methods to monitor, model, forecast and control COVID-19 pandemic: Leveraging data science, epidemiology and control theory. arXiv preprint arXiv: 2006.01731 (2020). https://doi.org/10.1016/j.coi.2020.09.011

[2] M. A. S. Alhdiri N. A. Samat, Z. Mohamed. Disease mapping for stomach cancer in libya based on besag–york–mollié (bym) model. Asian Pacific Journal of Cancer Prevention: APJCP 18 (6) (2017) 1479. https://doi.org/10.1007/s11356-022-23319-6

[3] M. P. Armstrong G. Rushton, D. L. Zimmerman. Geographically masking health data to preserve confidentiality. Statistics in medicine 18 (5) (1999) pp. 497–525. https://doi.org/10.1002/(SICI)1097-0258(19990315)18:5

[4] J. Besag J. York, A. Mollié. Bayesian image restoration, with two applications in spatial statis-tics. Annals of the institute of statistical mathematics 43 (1991) pp. 1–20. https://doi.org/10.1007/BF00058655

[5] M. Blangiardo, M. Cameletti. Spatial and spatio-temporal Bayesian models with R-INLA (2015). John Wiley & Sons.

[6] M. Blangiardo M. Cameletti G. Baio, H. Rue. Spatial and spatio-temporal models with r-inla.Spatial and spatio-temporal epidemiology 4 (2013) pp. 33–49. https://doi.org/10.1016/j.sste.2012.12.001

[7] P. J. Brantingham. Crime diversity. Criminology 54 (4) (2016) pp. 553–586. https://doi.org/https://doi.org/10.1111/doi:1745- 9125.12116.

[8] M. J. Breslow, O. Badawi. Severity scoring in the critically iii: Part 2: Maximizing value from outcome prediction scoring systems. Chest 141 (2) (2012) pp. 518–527. https://doi.org/10.1378/chest.11-0331

[9] A. C. Cameron, P. K. Trivedi. Regression analysis of count data (2013). Number 53. Cam- bridge university press.

[10] J. Chen J. J. Song, J. D. Stamey. A Bayesian hierarchical spatial model to correct for misreporting in count data: application to state-levelCOVID-19 data in the United States. International Journal of Environmental

Research and PublicHealth 19 (6) (2022) 3327. https://doi.org/10.3390/ijerph19063327

[11] J. T. Chen. 11 multilevel and hierarchical models for disease mapping. Geographic health data: Fundamental techniques for analysis (2013), pages183. https://doi.org/10.1079/9781780640891.0183

[12] E. Clement. Small area estimation with application to disease mapping. International Journal ofProbability and Statistics 3 (1) (2014) pp. 15–22. https://doi.org/10.5923/j.ijps.20140301.03

[13] A. Comunian R. Gaburro, M. Giudici. Inversion of a sir-based model: A critical analysis about the application to covid-19 epidemic. Physica D:Nonlinear Phenomena 413 (2020) 132674. https://doi.org/10.1016/j.cossms.2021.100411

[14] E. Cuevas. An agent-based model to evaluate the COVID-19 transmission risks in facilities. Computers in Biology and Medicine 121 (2020) 103827. https://doi.org/10.1016/j.compbiomed.2020.103827

[15] P. Elliott, D. Wartenberg. Spatial epidemiology: current approaches and future challenges. Environmental health perspectives 112 (9) (2004) pp. 998–1006. https://doi.org/10.1093/ije/dyz047

[16] I. Franch-Pardo B. M. Napoletano F. Rosete-Verges, L. Billa. Spatial analysis and gis in thestudy of COVID-19. a review. Science of the total environment 739 (2020) 140033. https://doi.org/10.1016/j.scitotenv.2020.140033

[17] V. Gómez-Rubio, F. Palm Perales. Multivariate posterior inference for spatial models with the integrated nested Laplace approximation. Journal of the Royal Statistical Society Series C: Ap- plied Statistics 68 (1) (2019) pp. 199–215. https://doi.org/10.1111/rssc.12292

[18] G. Grekousis Z. Feng I. Marakakis Y. Lu, R. Wang. Ranking the importance of demographic, socioeconomic, and underlying health factors on us COVID-19 deaths: A geographical random forest approach. Health & Place 74 (2022) 102744. https://doi.org/10.1016/j.healthplace.2022.102744

[19] R. P. Haining. Spatial data analysis: theoryand practice (2003). Cambridge university press.

[20] A. Jalilian, J. Mateu. A hierarchical spatio temporal model to analyse relative risk variations of COVID-19: a focus on Spain, Italy and Germany. Stochastic Environmental Research and Risk Assessment 35 (2021) pp. 797–812.

[21] H. Joe, R. Zhu. Generalized poison distribution: the property of mixture of poison andcomparison with negative binomial distribution. Biometrical Journal: Journal of Mathematical Methods in Biosciences 47 (2) (2005) pp. 219–229. https://doi.org/10.1002/bimj.200410102Citations: 140

[22] M. R. Karim. Bayesian hierarchical spatial modeling of COVID-19 cases in bangladesh. An- nals of Data Science (2023) pp. 1–27. https://doi.org/10.1007/s40745-022-00381-1

[23] M. U. Kraemer S. I. Hay D. M. Pigott D. L. Smith G. W. Wint, N. Golding. Progress and chal- lenges in infectious disease cartography. Trends inparasitology 32 (1) (2016) pp. 19–29. https://doi.org/10.1016/j.pt.2015.09.006

[24] E. Krainski V. Gómez-Rubio H. Bakka A. Lenzi D. Castro-Camilo D. Simpson F. Lind- gren, H. Rue. Advanced spatial modeling with stochastic partial differential equations using R and INLA (2018). Chapman and Hall/CRC.

[25] A. Lal J. Marshall J. Benschop A. Brock S. Hales M. G. Baker, N. P. French. A Bayesian spatio-temporal framework to identify outbreaks and examine environmental and social risk fac- tors for infectious diseases monitored by routine surveillance. Spatial and spatio-temporal epidemi- ology 25 (2018) pp. 39–48.

[26] K. Lancaster T. Rhodes, M. Rosengarten. Making evidence and policy in public health emergencies: lessons from COVID-19 for adap- tive evidence-making and intervention. Evidence & policy 16 (3) (2020) pp. 477–490. https://doi.org/10.1332/174426420X15913559981103

[27] A. B. Lawson. Bayesian disease mapping: hierarchical modeling in spatial epidemiology (2018). Chapman and Hall/CRC.

[28] P. Legendre. Spatial autocorrelation: troubleor new paradigm? Ecology 74 (6) (1993) pp. 1659– 1673. https://doi.org/10.2307/1939924

[29] J. W. Lichstein T. R. Simons S. A. Shriner, K. E. Franzreb. Spatial autocorrelation and autoregressive models in ecology. Ecological monographs 72 (3) (2002) pp. 445–463. https://doi.org/10.1890/0012-9615(2002)072[0445:SAAAMI]2.0.CO;2

[30] J. Ma H. Zhu P. Li C. Liu F. Li Z. LuoM. Zhang, L. Li. Spatial patterns of the spread of COVID-19 in singapore and the influencing factors. ISPRS International Journal of Geo- Information 11 (3) (2022) 152. https://doi.org/10.3390/ijgi11030152

[31] A. Maiti Q. Zhang S. Sannigrahi S. Pra-manik S. Chakraborti A. Cerda, F. Pilla. Ex-ploring spatiotemporal effects of the driving fac-tors on COVID-19 incidences in the contigu-ous

united states. Sustainable cities and society 68 (2021) 102784. https://doi.org/10.1016/j.scs.2021.102784

[32] T. J. Mason. Atlas of Cancer Mortality for US counties, 1950-1969 (1975). US Department of Health, Education, and Welfare, Public Health Service.

[33] J. C. Miller, E. M. Volz. Incorporating disease and population structure into models of sir disease in contact networks. PLOS ONE 8 (8) (2013) e69162. https://doi.org/10.1371/journal.pone.0069162

[34] P. Moraga. Geospatial health data: Modeling and visualization with R-INLA and shiny (2019).Chapman and Hall/CRC.

[35] A. K. Muoka O. O. Ngesa, A. G. Waititu. Statistical models for count data. Science Jour-nal of Applied Mathematics and Statistics 4 (6) (2016) pp. 256–262. https://doi.org/10.11648/j.sjams.20160406.12

[36] N. Nazia Z. A. Butt M. L. Bedard W.-C. Tang H. Sehar, J. Law. Methods used in the spatial and spatiotemporal analysis of COVID-19 epidemi- ology: a systematic review. International Jour- nal of Environmental Research and Public Health 19 (14) (2022) 8267. https://doi.org/10.3390/ijerph19148267

[37] K. W. Pettis T. A. Bailey A. K. Jain, R. C. Dubes. An intrinsic dimensionality estimator from near-neighbor information. IEEE Transactions onpattern analysis and machine intelligence 1 (1979) pp. 25–37.

[38] P. Puvanachandra C. Hoe T. Özkan, T. La-junen. Burden of road traffic injuries in turkey. Traffic injury prevention 13 (1) (2012) pp. 64–75. https://doi.org/15389588.2011.633135

[39] A. Riebler S. H. Sørbye D. Simpson, H. Rue.An intuitive bayesian spatial model for disease mapping that accounts for scaling. Statistical methods in medical research 25 (4) (2016) pp. 1145–1165. https://doi.org/10.1016/j.sste.2016.10.014

[40] A. D. Roux C. I. Kiefe D. R. Jacobs Jr M. Haan S. A. Jackson F. J. Nieto C. C. Paton,R. Schulz.Area characteristics and individual-level socioeconomic position indicators in three population-based epidemiologic studies. Annals of epidemiology 11 (6) (2001) pp. 395–405. https://doi.org/10.1016/S1047-2797(01)00221-6

[41] H. Rue S. Martino, N. Chopin. Approximate Bayesian inference for latent gaussian models by using integrated nested Laplace approximations.Journal of the Royal Statistical Society Series B:Statistical Methodology 71 (2) (2009) pp. 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

[42] P. Saavedra A. Santana L. Bello J.-M. Pacheco, E. Sanjuán. A bayesian spatio-temporalanalysis of mortality rates in spain: application to the covid-19 2020 outbreak. Population Health Metrics 19 (1) (2021) 27. https://doi.org/10.1186/s12963-021-00259-y

[43] A. K. Sahu V. T. Amritha Nand R. Mathew P. Aggarwal J. Nayer, S. Bhoi. Covid-19 in health care workers–a systematic review and meta- analysis. The American Journal of EmergencyMedicine 38 (9) (2020) pp. 1727–1731. https://doi.org/10.1016/j.ajem.2020.05.113

[44] S. K. Sahu, D. Böhning. Bayesian spatio-temporal joint disease mapping of COVID-19 cases and deaths in local authorities of england. Spatial Statistics 49 (2022) 100519. https://doi.org/10.1007/s11356-021-12925-2

[45] P. Satorra, C. Tebé. Bayesian spatio-temporal analysis of the covid-19 pandemic in Catalonia Scientific Reports 14 (1) (2024) 4220. https://doi.org/10.1007/s11749-021-00769-4

[46] O. Schabenberger, C. A. Gotway. Statistical methods for spatial data analysis (2017). Chapmanand Hall/CRC.

[47] K. F. Sellers S. Borle, G. Shmueli. The com-poison model for count data: a survey of methods and applications. Applied Stochastic Models in Business and Industry 28 (2) (2012) pp. 104–116.

[48] S. Sisman, A. C. Aydinoglu. A modelling approach with geographically weighted regressionmethods for determining geographic variation andinfluencing factors in housing price: A case inistanbul. Land use policy 119 (2022) 106183. https://doi.org/10.1016/j.landusepol.2022.106183

[49] R. J. Thomas. Female consumption and evaluation of traditionally male orientated products: aself-monitoring perspective (2010). University of South Wales (United Kingdom).

[50] P. H. Verburg K. Kok R. G. Pontius Jr, A. Veldkamp. Modelling land-use and land-cover change. In: Land-use and land-cover change: local processes and global impacts, pp. 117–135 Springer (2006).

[51] L. A. Waller L. Zhu C. A. Gotway D. M. Gor-man, P. J. Gruenewald. Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models. Stochastic Environmental Re-search and Risk Assessment 21 (2007) pp. 573–588. https://doi.org/10.1007/s00477-007-0139-9

[52] S. Wang X. Yang L. Li P. Nadler R. Arcucci Y. Huang Z. Teng, Y. Guo. A Bayesian updating scheme for pandemics: estimating the infection dynamics of COVID-19. IEEE Computational Intelligence Magazine 15 (4) (2020) pp. 23–33. https://doi.org/10.1016/j.jbi.2020.103347

[53] Y. Wang X. Chen, F. Xue. A review of Bayesian spatiotemporal models in spatial epidemiology. ISPRS International Journal of Geo-Information 13 (3) (2024) 97. https://doi.org/10.3390/ijgi13030097

[54] A. Woo ditch N. J. Johnson R. Solymosi J. Medina Ariza, S. Langton. Getting to know your data. In: A Beginner's Guide to Statistics forCriminology and Criminal Justice Using R, pp. 21–38. Springer (2021).