Infectious Disease Modelling 10 (2025) 1002–1019

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

Evaluating the impact of the Modifiable Areal Unit Problem on ecological model inference: A case study of COVID-19 data in Queensland, Australia

Shovanur Haque [a], Aiden Price [b, c, d], Kerrie Mengersen [b, c], Wenbiao Hu [a, *]

[a] Ecosystem Change and Population Health Research Group, School of Public Health and Social Work, Queensland University of Technology, Brisbane, Australia
[b] School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia
[c] Centre for Data Science (CDS), Queensland University of Technology (QUT), Brisbane, Australia
[d] Australian Urban Research Infrastructure Network (AURIN), University of Melbourne, Melbourne, Australia

ARTICLE INFO

Article history:
Received 23 January 2025
Received in revised form 24 April 2025
Accepted 8 May 2025
Available online 10 May 2025
Handling Editor: Dr Yiming Shao

Keywords:
Modifiable Areal Unit Problem (MAUP)
Bayesian models
Spatial patterns
Model inference
COVID-19
Socio-Economic Indexes for Areas (SEIFA)

ABSTRACT

Accurate identification of spatial patterns and risk factors of disease occurrence is crucial for public health interventions. However, the Modifiable Areal Unit Problem (MAUP) poses challenges in disease modelling by impacting the reliability of statistical inferences drawn from spatially aggregated data. This study examines the effect of MAUP on ecological model inference using locally and overseas-acquired COVID-19 case data from 2020 to 2023 in Queensland, Australia. Bayesian spatial Besag-York-Mollié (BYM) models were applied across four Statistical Area (SA) levels, as defined by the Australian Statistical Geography Standard, with and without covariates: Socio-Economic Indexes for Areas (SEIFA) and overseas-acquired (OA) COVID-19 cases. OA COVID-19 cases were also considered a response variable in our study. Results indicated that finer spatial scales (SA1 and SA2) captured localized patterns and significant spatial autocorrelation, while coarser levels (SA3 and SA4) smoothed spatial variability, masking potential outbreak clusters. Incorporating SEIFA as a covariate in locally-acquired (LA) cases reduced spatial autocorrelation in residuals, effectively capturing socioeconomic disparities. Conversely, OA cases showed limited effectiveness in reducing autocorrelation at finer scales. For LA cases, higher socioeconomic disadvantage was associated with increased COVID-19 incidence at finer scales, but this association became non-significant at coarser scales. OA cases showed significant positive association with higher SEIFA scores at finer scales. Model parameters displayed narrower credible intervals at finer scales, indicating greater precision, while coarser levels had increased uncertainty. SA2 emerged as an arguably optimal scale, striking a balance between spatial resolution, model stability, and interpretability. To improve inference on COVID-19 incidence, it is recommended to use data from both SA1 and SA2 levels to leverage their respective strengths. The findings emphasize the importance of selecting appropriate spatial scales and covariates or evaluating the inferential impacts of multiple scales, to address MAUP to facilitate more reliable spatial analysis. The study advocates exploring intermediate aggregation levels and multi-scale approaches to better capture nuanced disease dynamics and extend these analyses across Australia and replicating in other countries with low population densities to enhance generalizability.

* Corresponding author.
E-mail address: w2.hu@qut.edu.au (W. Hu).
Peer review under the responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.idm.2025.05.003

1. Introduction

The spatial modelling of health data, particularly through disease mapping techniques, has become increasingly important as data accessibility improves (Wakefield & Lyons, 2010). Disease mapping provides spatially smoothed visualizations of disease outcomes across regions, supporting policymakers in formulating targeted interventions at the community level (Cramb et al., 2020; Duncan et al., 2019; Lee, 2011). Aggregating spatial data remains a common practice in disease mapping, as it enables the identification of disease clusters, helps understand the variability in disease burden, and ensures the confidentiality of health information (Best et al., 2005). However, the use of aggregated spatial data introduces biases, most notably the Modifiable Areal Unit Problem (MAUP).

The MAUP, first introduced by Openshaw in 1977 (Openshaw, 1977), highlighted how statistical inferences may vary depending on the spatial configuration and resolution used in data aggregation. The MAUP poses a significant challenge in spatial analyses because changes in spatial units or aggregation levels can substantially influence outcomes. This issue becomes particularly crucial when analyses depend on data aggregated at a single spatial level, as discussed by Tuson et al. (Tuson et al., 2020). MAUP manifests in two primary ways: the zoning effect, where boundaries are modified while the total number of areas remains constant, and the scaling effect, where spatial resolution or aggregation levels are adjusted, altering the total number of areas (Wong, 2004). These variations can result in inconsistent findings when analysing the same dataset at different spatial scales, as demonstrated in various studies of disease patterns (Fotheringham & Wong, 1991; Gregorio et al., 2005). Different zoning schemes and scales may lead to divergent interpretations of disease clusters, ultimately affecting conclusions about high-risk areas. In light of these potential discrepancies, carefully selecting the most appropriate aggregation levels is crucial to ensure robust and reliable inferences.

Several studies have explored the effects of the MAUP in environmental and health analyses, where different aggregation levels led to varying outcomes (Besag et al., 1991; Tuson et al., 2019). For instance, MAUP has been shown to impact district-level analyses of COVID-19 outbreaks and environmental factors, causing biases in model results (Wang & Di, 2020). To address these effects, researchers have developed methods such as empirical zoning distributions, Bayesian shared-effects modelling frameworks, and multi-scale aggregation strategies to mitigate the impact of MAUP (Briz-Redón, 2022; Burden & Steel, 2016; Tuson et al., 2018). These approaches provide zone-independent estimates and help identify spatial units contributing to discrepancies, thus enhancing the reliability of spatial analyses (Burden & Steel, 2016; Wang & Di, 2020).

The ecological and disease mapping literature has historically underemphasized the impact of the MAUP, with early reviews showing that only a small proportion of studies addressed it (Manley, 2021; Tuson et al., 2020). However, the growing interest in MAUP has led to a substantial body of work focusing on its implications, providing a strong foundation for integrating these considerations into spatial modelling (Roquette et al., 2017). Recent research has focused on quantifying MAUP effects at both global and local levels using Bayesian modelling, offering more refined insights into how spatial unit selection affects model results (Briz-Redón, 2022).

Australia's vast geographic area and diverse population distribution present unique challenges for health data analysis, particularly in regions with sparse populations such as rural and remote areas. While previous studies have applied Bayesian spatial models to explore MAUP in well populated areas, limited attention has been given to its impact in rural or sparsely populated regions (Kok et al., 2021; Tuson et al., 2020). The country's demographic and geographic diversity, along with limited access to high-resolution health outcome data, necessitates a thorough investigation into how the MAUP affects spatial modelling, particularly for infectious diseases like COVID-19 where data points may be low. Understanding these implications is essential for ensuring accurate and reliable public health interventions that reflect the realities of these regions.

This study aims to address this research gap by investigating the effects of the MAUP on modelling COVID-19 data in Queensland, Australia. We utilized locally-acquired (LA) and overseas-acquired (OA) COVID-19 data from 2020 to 2023, aggregated across various spatial resolutions: SA1, SA2, SA3, and SA4. Our aim is to evaluate how different levels of spatial aggregation influence model inferences and spatial patterns, and to assess the effect of incorporating covariates—namely Socio-Economic Indexes for Areas (SEIFA) and OA COVID-19 cases—on model outcomes. OA COVID-19 cases were also considered a response variable, with SEIFA IRSD quintile scores used as covariates across various aggregation levels. This analysis sought to explore the relationship between imported COVID-19 cases and areas that are more affluent or experience lower levels of disadvantage.

To achieve this, we applied Bayesian spatial Besag-York-Mollié (BYM) models at different spatial scales, both with and without covariates. Spatial patterns were assessed using Moran's I to detect spatial autocorrelation in observed counts and model residuals. We also generated choropleth maps and credible interval plots to visualize the spatial distribution of observed and fitted Standardized Incidence Ratios (SIR). This research provides critical insights into the effects of the MAUP

on disease mapping in sparsely populated regions and offers valuable guidance on selecting appropriate spatial units in similar contexts to support more accurate and equitable public health policies.

2. Materials and methods

2.1. Geographic framework

The Australian Statistical Geography Standard (ASGS) 2021, developed by the Australian Bureau of Statistics (ABS), provides a hierarchical framework for statistical areas in Australia. The hierarchy comprises of five levels of resolution. At the finest resolution are 'Mesh Blocks', the most granular level of analysis. The remaining levels, known as 'Statistical Areas' (SA), range from Statistical Area Level 1 (SA1) to Statistical Area Level 4 (SA4), with decreasing resolution. The structure is fully nested, with SA1s comprising groups of Mesh Blocks, SA2s comprising groups of SA1s, and so on, up to SA4. In Queensland, the ASGS 2021 defines 12,545 SA1s with a median population of 417, 507 SA2 with a median population of 9,588, 82 SA3 regions with a median population of 58,671, and 19 SA4 regions with a median population of 243,798. The hierarchical framework allows us effectively to incorporate geographic information into the model, improving its ability to capture spatial dynamics and predict disease risk across different regions. By aligning the analysis with geographic realities of the study area, this approach supports more informed decision-making for public policy and intervention planning.

2.2. Data

The data for this study were sourced from the Queensland Government's open data portal (Queensland COVID-19 Case Line List by Location and Source of Infection). Locally and overseas-acquired COVID-19 cases reported at the SA2 level were collected from 2020 to 2023. Population estimates collected from the Australian Bureau of Statistics (Australian Bureu of Statistics - Regional population) in 2023 were used to determine incidence rates across geographic scales.

The observed COVID-19 cases represent the total number of confirmed cases reported in each SA2. These counts were directly extracted from the source dataset. COVID-19 case data were then aggregated from SA2 level to SA3 and SA4 using publicly available geographical correspondence files, ensuring accurate alignment of cases to the larger spatial units (Australian Statistical Geography Standard (ASGS) Edition 3). Data were also disaggregated from SA2 to SA1 using population-weighted geographical correspondence files (Correspondences). This approach redistributed the observed cases proportionally across SA1s based on population share within each SA2. The counts and populations for each aggregation level were determined by summing the counts and populations of the corresponding areas. The detailed description of the database structure, key variables, and record counts used in this study is provided in Supplementary Table S1.

Raw disease counts alone do not accurately reflect disease risk, as they do not account for population differences across areas. To address this, the SIR, calculated as the ratio of observed to expected COVID-19 cases, was used to capture the underlying pattern of COVID-19 incidence in the data. The SIR was calculated at each level of aggregation (SA1, SA2, SA3, and SA4), providing a measure of disease risk that accounts for population size in each area.

The SIR for each area i (where $i = 1, \ldots, n$) is defined as:

$$SIR_i = Y_i / E_i$$

where Y_i is the observed number of COVID-19 cases in area i, and E_i is the expected number of cases that represent the total number of cases that would be expected if the population of area i had the same incidence rate as the standard population.

The expected cases E_i is calculated using indirect standardization as:

$$E_i = r^{(s)} n^{(i)},$$

Where $r^{(s)}$ is the rate in the standard population, calculated as the total number of cases divided by total population across all areas. $n^{(i)}$ is the population of area i.

Table 1 highlighted the impact of the MAUP by demonstrating how spatial aggregation alters data patterns for LA and OA COVID-19 incidence across geographic levels (SA1 to SA4). At finer scales (SA1), LA incidence showed a lower mean (601.3) but higher variability (SD = 1714.1), capturing localized clustering. However, at SA2, the mean increased to 913.9 before dropping to 700.9 and 832.7 at SA3 and SA4, respectively. This trend shift reflected MAUP effects as spatial distribution changed with the unit boundaries. Similarly, while the maximum LA incidence at SA1 (39166.6) and SA2 (39166.7) remained nearly the same, the median increased by 20 %, further emphasized MAUP-specific distortions. OA incidence also displayed decreasing variability at coarser scales (SA3 and SA4) compared to SA2, masking fine-scale trends. The SEIFA scores demonstrated reduced variability as scales became coarser, indication the loss of socioeconomic detail with increasing aggregation. Supplementary Table S1 provides details of the database structure and key components used in this study.

2.2.1. Socio-Economic Indexes for Areas (SEIFA) calculation

Previous research have showed that socio-economic factors are significantly associated with COVID-19 cases and fatalities, indicating that greater socioeconomic disadvantage is linked to higher incidence and mortality (Faramarzi et al., 2022;

Table 1
Population and number of locally-acquired and overseas-acquired COVID-19 case identification by different geographic levels.

Geographic levels					
		SA1	**SA2**	**SA3**	**SA4**
Number of areas		12,545	507	82	19
Population	Mean	422.6	10,461	64,679	279,139
	SD[a]	210.4	5286.3	35164.4	131972.3
	Median	417	9588	58,671	243,798
	Min	15.8	652	12,735	82,158
	Max	3049.4	27,198	201,478	677,285
LA COVID-19 cases	Mean	2.7	61.6	380.8	1643.7
	SD	11.2	144.2	977.4	3189.7
	Median	0.4	11	46	280
	Min	0	1	3	42
	Max	525.6	1169	6336	13,361
OA COVID-19* cases	Mean	0.13	3.2	15.3	63.7
	SD	0.17	3.6	17.01	54.5
	Median	0.09	2.0	13.0	45.0
	Min	0.002	1.0	1.0	11.0
	Max	4.1	45.0	139.0	242.0
LA COVID-19** incidence (per 100,000)	Mean	601.3	913.9	700.9	832.7
	SD	1714.1	2906.5	1951.4	1702.2
	Median	99.9	118.4	91.2	134.1
	Min	5.03	5	9.2	21.8
	Max	39166.6	39166.7	14012.4	5975
OA COVID-19 incidence (per 100,000)	Mean	28.8	31.7	27.3	28.5
	SD	29.4	34.5	16.6	13.4
	Median	22.9	24.4	25.02	25.02
	Min	3.8	3.8	6.70	16.3
	Max	435.5	435.5	139.2	78.1
LA SEIFA irsdScore	Mean	993.3	988.1	993.6	990.3
	SD	93.2	84.9	52.5	46.3
	Median	1012.0	1006.0	991.0	979.0
	Min	317.0	501.0	894.0	917.0
	Max	1207.0	1119.0	1106.0	1076.0
OA SEIFA irsdScore	Mean	1001	1003	994.5	990.3
	SD	91.4	71.4	53.3	46.3
	Median	1021	1018	992	979
	Min	369	649	894	917
	Max	1207	1119	1106	1076

[a] SD: stands for standard deviation; * LA: locally-acquired; ** OA: overseas-acquired.

Hawkins et al., 2020; Politi et al., 2021). To assess the impact of the MAUP on model inference at different spatial scales, we included area-level socioeconomic disadvantage as a covariate. We utilized the Index of Relative Socioeconomic Disadvantage (IRSD) from the ABS, categorized into quintiles, with regions in the first quintile representing the most disadvantaged areas and those in the fifth quintile representing the least disadvantaged (ABS. Socio-Economic Indexes for Areas, 2021). The IRSD scores and quintiles are publicly available at the SA1 and SA2 levels. For this study, SA2-level IRSD scores were aggregated to higher regions, SA3 and SA4, following ABS-recommended methods (ABS. Socio-Economic Indexes for Areas (SEIFA), Australia methodology), and categorized into quintiles.

The observed SIR maps in Fig. 1 highlighted the MAUP-specific impacts on spatial analysis by demonstrating how different aggregation levels altered the visibility of disease patterns. The finer resolutions (SA1 and SA2) revealed much more detailed spatial variability, with clear demarcation of areas with both high and low SIR values. At SA1, higher heterogeneity in the patterns is observed, which better captured local variations in disease incidence. SA2 maintained meaningful spatial trends while providing a clearer overall picture of disease spread. At coarser levels like SA3 and SA4, the spatial variability became increasingly obscured, with SA4 displaying large homogeneous blocks that oversimplified disease distribution. This smoothing effect due to aggregation masked true hotspots, reduced spatial heterogeneity, and potentially led to misleading conclusions about disease spread and intervention strategies.

2.3. Statistical methods

To assess the effects of the MAUP on model inference across different aggregation levels, Bayesian spatial Besag-York-Mollié (BYM) models were fitted at SA1, SA2, SA3, and SA4 levels, both with and without covariates. This study evaluates how varying levels of aggregation influence model outcomes and inferences. Posterior summaries of model parameters were analysed and compared across these levels. Spatial autocorrelation in the observed data and model residuals was assessed using Moran's *I*, an inferential statistic that quantifies spatial autocorrelation (Getis, 2009; Kim, 2021).

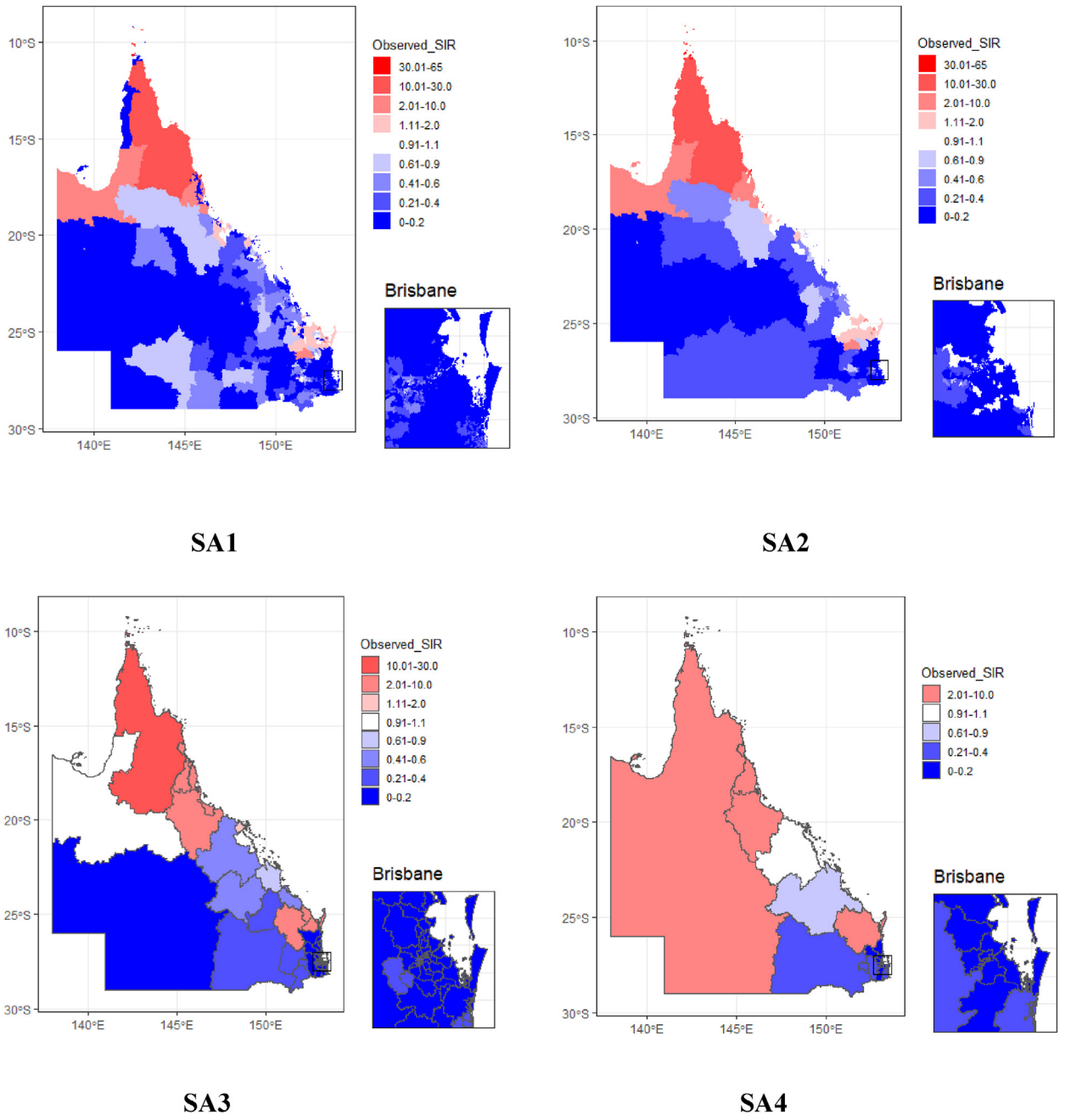

Fig. 1. Choropleth maps of the observed SIR from the LA COVID-19 data at multiple aggregate levels in Queensland, Australia.

2.3.1. Assessing spatial autocorrelation with Moran's I

Ignoring spatial autocorrelation can lead to biased or misleading results in spatial analyses; therefore, methods that account for spatial structure are recommended to enhance the accuracy and validity of spatial data analysis (Cliff & Ord, 1968). Among the various indices and tests available to detect spatial autocorrelation, we employed Moran's I statistic for this purpose. We first examined how spatial autocorrelation varies across different aggregation levels in the raw data, then checked for any remaining spatial autocorrelation in the model residuals.

Let x_i denote the number of observed COVID-19 cases and n_i represent the population at risk in geographic area, where $i = 1,2,\ldots,N$, and N is the total number of geographic areas. Let w_{ij} be the spatial weight that defines the relationship between the

pair of the geographic areas i and j (where $i \neq j$) (Getis, 2009; Oyana, 2020). Here, w_{ij} reflects the strength of the spatial relationship between pairs i and j. The spatial weight w_{ij} is defined as:

$$w_{ij} = \begin{cases} 1 & \text{if } i,j \text{ are adjacent neighbors} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Moran's I index (Moran, 1950) of spatial autocorrelation can be expressed as:

$$I = \left(\frac{1}{S_x^2}\right) \frac{\sum_i^N \sum_j^N w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i^N \sum_j^N w_{ij}} \tag{2}$$

where, $\bar{x} = \sum_i \frac{x_i}{N}$, $s_x^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$

If we define,

$$s_{xx,w} = \frac{\sum_i^N \sum_j^N w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i^N \sum_j^N w_{ij}}$$

then equation (2) can be simplified as:

$$I = \frac{s_{xx,w}}{s_x^2}$$

The value of Moran's I typically ranges between -1 and 1, though the range of I depends on the values of the weight function in equation (2). The expected value of I is:

$$E(I) = -\frac{1}{N-1}$$

Large positive values of I indicate strong geographic patterns of spatial clustering; large negative values suggest strong dissimilarity between neighbouring areas, and values close to zero represent spatial randomness.

We employed Moran's I, using the spdep package version 1.2–8 (Lam, 2002), to examine spatial autocorrelation in the observed data and model residuals. The moran.test function in the R statistical software (Team, 2013) was used to compute Moran's I statistic, testing under the null hypothesis of randomization. Additionally, a Markov Chain simulation was conducted to further test the hypothesis.

2.3.2. Bayesian spatial Besag-York-Mollié (BYM) model

To investigate the effects of the MAUP on model inference across different spatial aggregation levels and evaluate the role of covariates, we employed the Bayesian spatial Besag-York-Mollié (BYM) model (Besag et al., 1991). The model was chosen because it is specifically designed to handle spatially correlated data, where regions are of different size and where the response variable and covariates may not vary smoothly across the whole geographic space. The BYM model incorporates a spatial random effect to capture correlations between neighbouring areas based on a predefined neighbourhood structure and an unstructured random effect to model independent noise. This dual-component structure allows for locally smoothing of data, making the BYM model ideal for analysing spatial patterns and dependencies in infectious diseases like COVID-19. By accommodating spatial heterogeneity and noise, the BYM model provides robust inference and deeper insights into how covariates influence disease distribution across different spatial scales.

The BYM model operates as a three-level hierarchical Bayesian model. At the first level, the observed disease counts are modelled using a suitable likelihood function suitable for count data. The second level incorporates spatial association through both structured and unstructured random effects. The spatially structured component typically uses a Gaussian Markov Random Field (GMRF) model, particularly in the form of a Conditional Autoregressive (CAR) prior, which conditions each area i on its neighbouring areas (CARLIN & Xia, 1999; Escaramís et al., 2008). The third level assigns hyperprior distributions to the model parameters (Best et al., 2005). The fraction of spatial variation (FSV) is a metric used to quantify the relative contribution of spatially structured random effects versus unstructured random effects in the BYM model or similar spatial hierarchical frameworks. An FSV closer to 1 indicates most of the variation is spatially structured, while an FSV closer to 0 indicates most of the variation is unstructured (random noise).

Initially, the BYM model was applied to LA COVID-19 count data across various spatial scales, both with and without the two covariates, SEIFA and OA COVID-19 cases. We employed the BYM model to the LA COVID-19 case counts across various levels of geographic aggregation (SA1, SA2, SA3, and SA4) to investigate how the influence of socioeconomic disadvantage, represented by SEIFA IRSD scores, varies across spatial scales. The model was run with SEIFA IRSD quintile scores incorporated as covariates. Subsequently, the model was applied to OA COVID-19 cases, using SEIFA as a covariate.

Let Y_i, be the observed COVID-19 counts in the i th area, $i = 1, 2, \ldots, N$

Since a Poisson likelihood is commonly used for count data, Y_i can be expressed as follows:

$$Y_i \sim \text{Poisson}\,(E_i e^{\mu_i}), \text{for } i = 1, 2, \ldots, N \text{ areas}, \tag{3}$$

where E_i is the expected COVID-19 counts for area i, and N varies depending on the boundary levels (Table 1, Table S1). The log-relative risk, μ_i is expressed as a regression equation,

$$\mu_i = \alpha + \boldsymbol{x}_i^T \boldsymbol{\beta} + \psi_{\boldsymbol{i}} \tag{4}$$

where, the intercept α indicates an overall fixed effect, $\boldsymbol{\beta}$ represents the covariate effects associated with the vector of covariates $\boldsymbol{x}_i$. In our study $\boldsymbol{x}_i$ denotes the SEIFA IRSD quintiles for area i. Quintile 1 (most disadvantaged) is the reference category, making $\boldsymbol{\beta}$ a 4×1 vector for the effects of quintiles 2, 3, 4, and 5. Gaussian priors $N(0, \sigma_\alpha^2)$ and $N(0, \sigma_\beta^2)$ are specified for α and $\boldsymbol{\beta}$, with large values assigned to σ_α^2 and σ_β^2 (in this study, 1,000,000 as per the CARBayes package defaults (Lee, 2013)). The spatial random effects, $\psi_{\boldsymbol{i}}$ is decomposed into $u_i + v_i$, where u_i represents the spatially structured effect with a CAR prior, and v_i is an unstructured random effect. The conditional distribution of each u_i, given all other u_j ($j \neq i$), is:

$$u_i | \boldsymbol{u}_j j \neq i \sim N\left(\frac{\sum_j w_{ij}\, u_j}{\sum_j w_{ij}}, \frac{\sigma_u^2}{\sum_j w_{ij}}\right) \tag{5}$$

where w_{ij} denotes the spatial weight between areas i and j, defined as:

$$w_{ij} = \begin{cases} 1 & \text{if areas } i, j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}$$

The unstructured random effect v_i follows an independent normal distribution,

$$v_i \sim N\left(0, \sigma_v^2\right) \tag{6}$$

Hyperpriors for the variance parameters σ_u^2 and σ_v^2 are specified as Inverse-Gamma (a, b) distributions with shape parameter $a = 1$ and scale $b = 0.01$ (Lee, 2013).

The BYM model was employed in a similar manner to investigate the effect of the OA COVID-19 cases covariate on local COVID-19 case counts across different spatial scales. Here, the same sampling distribution (3) was employed, with the log-relative risk, μ_i, which is expressed as,

$$\mu_i = \alpha + \gamma (OA)_{\boldsymbol{i}} + \psi_{\boldsymbol{i}}$$

where, the coefficient γ associated with the OA COVID-19 cases allows us to assess the influence of imported cases on local transmission rates. As above, Gaussian priors $N(0, \sigma_\alpha^2)$ and $N(0, \sigma_\gamma^2)$ are specified for α and γ, with large values assigned to σ_α^2 and σ_γ^2.

u_i and v_i follow (5) and (6), respectively with equivalent hyperparameters.

We further fitted a BYM model to OA COVID-19 cases, incorporating SEIFA IRSD quintile scores as covariates. This is to investigate the association between more affluent or least disadvantaged areas and imported COVID-19 cases.

A total of twelve BYM models were fitted in the statistical software R (Team, 2000). Models were fitted at each aggregation level (SA1, SA2, SA3, and SA4) and for each model specification (with and without covariates). Fully Bayesian inference was conducted using Markov chain Monte Carlo (MCMC) methods through the CARBayes package, version 5.2.5 (Lee, 2013). For each model, a single chain of 1,500,000 iterations was run, with the first 500,000 discarded as burn-in. Posterior samples were thinned by retaining every 100th iteration, resulting in 10,000 draws. Model convergence was evaluated using Geweke diagnostics (Geweke, 1991; Lam, 2002).

Table 2 presented Moran's I statistic and p-values for observed counts and residuals from models with and without covariates, SEIFA and OA cases, at different spatial aggregation levels SA1, SA2, SA3, and SA4. Including covariates reduced spatial autocorrelation (Moran's I statistic) across all scales compared to the model without covariates. As the aggregation level became coarser, from SA2 to SA4, the p-values for model residuals, both with and without covariates, became non-significant. This indicated that the model captured some spatial structure at these levels. A well-fitted model significantly reduced the spatial autocorrelation present in the observed data, suggesting that the spatial structure had been adequately accounted for, leaving residuals that were spatially uncorrelated and more suitable for reliable inference.

Table 2
Moran's *I* of observed counts and modelled residuals of LA COVID-19 cases.

Geographic level	Observed counts		Residuals (without covariate)		Residuals (with covariate_SEIFA)		Residuals (with covariate_OA)	
	statistic	p-value	statistic	p-value	statistic	p-value	statistic	p-value
SA1	0.374	<0.0001	0.699	<0.0001	0.591	<0.0001	0.594	<0.0001
SA2	0.634	0.001	0.047	0.053	0.032	0.110	0.036	0.087
SA3	0.361	0.001	0.044	0.180	0.007	0.170	0.020	0.279
SA4	0.259	0.001	−0.065	0.494	−0.163	0.765	−0.118	0.657

Table 3
Moran's *I* of observed counts and modelled residuals for OA COVID-19 cases.

Geographic level	Observed counts		Residuals (with SEIFA)	
	statistic	p-value	statistic	p-value
SA1	0.667	<0.0001	0.158	<0.001
SA2	0.270	<0.0001	0.039	0.131
SA3	0.263	<0.001	−0.194	0.996
SA4	0.219	0.010	−0.041	0.429

Table 4
Posterior summary of BYM model results without covariates for LA COVID-19 cases.

Geographic level	Parameter estimates			
	α Median (95 % CI)	σ_u^2 Median (95 % CI)	σ_v^2 Median (95 % CI)	FSV $= \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2}$
SA1	−2.04 (−2.10, −1.99)	0.50 (0.46, 0.54)	0.003 (0.001, 0.006)	0.0045
SA2	−1.37 (−1.41, −1.34)	1.00 (0.75, 1.31)	0.32 (0.25, 0.41)	0.23
SA3	−1.55 (−1.60, −1.51)	1.55 (0.93, 2.58)	0.26 (0.09, 0.63)	0.12
SA4	−1.23 (−1.26, −1.19)	3.29 (0.96,11.04)	0.60 (0.03, 2.36)	0.07

Table 5
Posterior summary of BYM model results with covariate (SEIFA) for LA COVID-19 cases.

Geographic level	Parameter estimates							
	α Median (95 % CI)	β_1 Median (95 % CI)	β_2 Median (95 % CI)	β_3 Median (95 % CI)	β_4 Median (95 % CI)	σ_u^2 Median (95 % CI)	σ_v^2 Median (95 % CI)	FSV $= \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2}$
SA1	−1.92 (−1.99, −1.85)	−0.11 (−0.16, −0.06)	−0.15 (−0.21, −0.09)	−0.15 (−0.22,-0.09)	−0.18 (−0.26, −0.10)	0.49 (0.45, 0.53)	0.002 (0.001, 0.007)	0.0056
SA2	−1.02 (−1.24, −0.82)	−0.39 (−0.70, −0.07)	−0.44 (−0.77, −0.12)	−0.45 (−0.79, −0.12)	−0.46 (−0.84, −0.07)	0.98 (0.74, 1.28)	0.31 (0.24, 0.39)	0.24
SA3	−1.38 (−1.99, −0.80)	−0.11 (−0.91, 0.91)	−0.19 (−1.12, 0.66)	−0.43 (−1.70, 0.65)	−0.12 (−1.38, 1.09)	1.68 (0.95, 3.12)	0.29 (0.08, 0.72)	0.12
SA4	0.09 (−1.79, 1.48)	−2.17 (−3.94, 1.31)	−3.10 (−6.47, 1.34)	−0.76 (−5.47, 2.05)	−0.66 (−2.67, 3.40)	6.28 (1.21, 23.25)	1.02 (0.09, 3.82)	0.09

Table 3 presented Moran's *I* statistic and p-values for observed counts and residuals from models incorporating SEIFA as a covariate at different spatial aggregation levels (SA1, SA2, SA3, and SA4) for OA COVID-19 cases. Including SEIFA reduced spatial autocorrelation (Moran's *I* statistic value) across all scales to a certain degree. At SA1, observed counts displayed spatial autocorrelation (Moran's $I = 0.667$, $p < 0.0001$), which was reduced in the residuals (Moran's $I = 0.158$, $p < 0.001$). As aggregation increased, spatial autocorrelation in the observed counts decreased, with non-significant p-values for model residuals at SA2, SA3, and SA4. This indicated that at these levels, the model adequately accounted for spatial structure.

In Table 4, the α parameter (intercept) became less negative at coarser aggregation levels, reflecting changes in overall incidence rates due to spatial scaling. The spatial random effects (σ_u^2) increased from SA1 to SA4, highlighting that spatial heterogeneity was more prominent at coarser levels. The unstructured random effects (σ_v^2) also increased, but to a lesser degree than the structured spatial effects, indicating more spatial noise was present at coarser resolutions. At SA1, the FSV was low (0.0045), suggesting that spatially structured variation accounted for only a small proportion of the total variance, with unstructured random effects dominating. As aggregation increased, the FSV decreased, from 0.23 at SA2 to 0.07 at SA4, indicating that coarser scales masked spatial dependencies and introduced greater uncertainty in spatial structure.

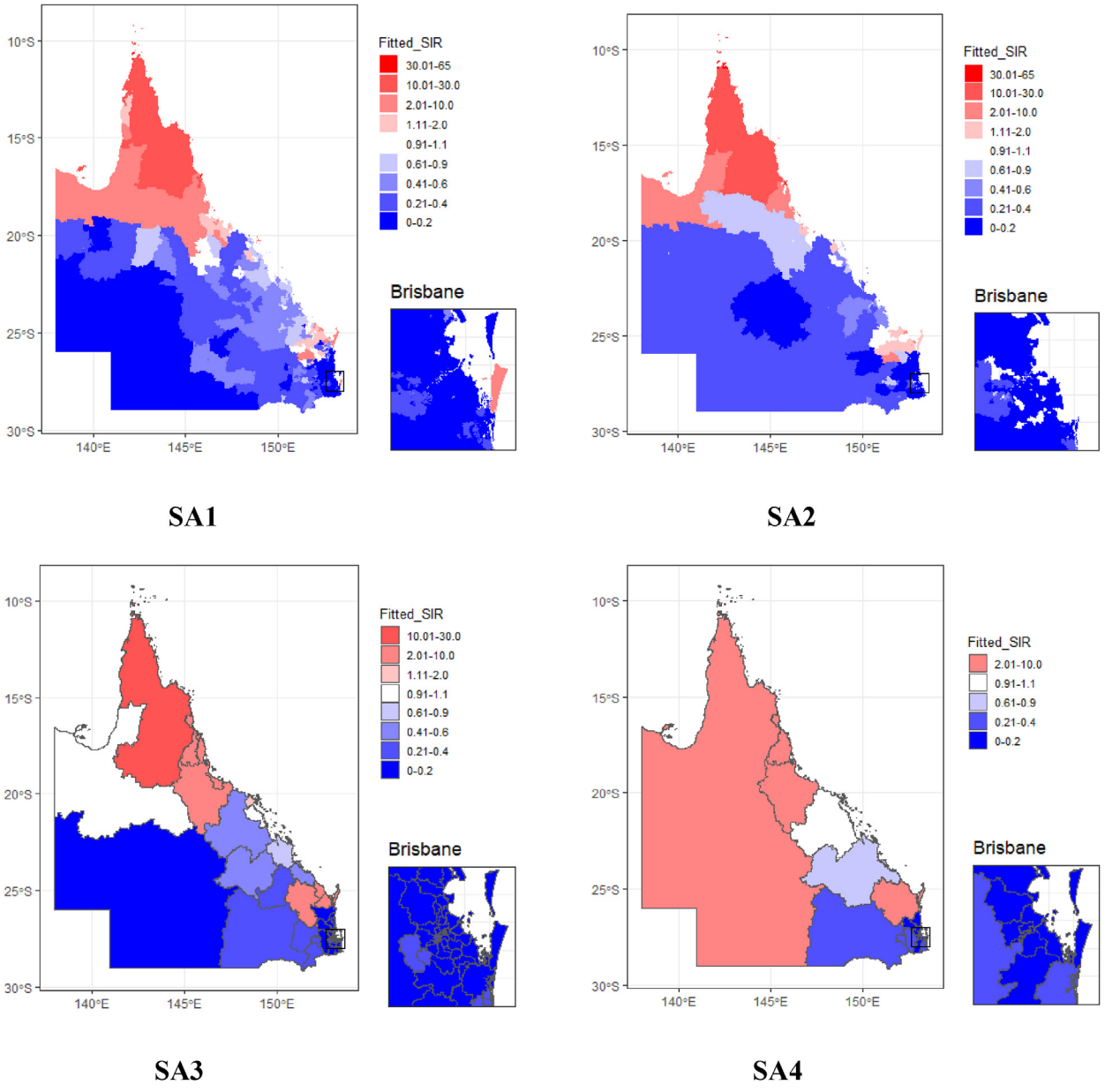

Fig. 2. Choropleth maps of the fitted SIR without covariates from the LA COVID-19 counts at multiple aggregate levels in Queensland, Australia.

The fitted SIR maps in Fig. 2 displayed a smoothing effect, reducing some of the extreme variations seen in the observed data. Differences between observed (Fig. 1) and fitted SIRs were particularly notable at finer scales, where observed risk tended to show more extreme values, suggesting that the model captures broad spatial trend but may reduce local variability.

The discrepancy between observed and fitted SIRs at different scales demonstrated the MAUP, highlighting how spatial aggregation influences disease risk interpretation across geographic levels.

The inclusion of SEIFA as a covariate improved model fit by reducing both spatial and unstructured random effects at finer scales (SA1 and SA2), which demonstrated the importance of socioeconomic factors in explaining COVID-19 variation. However, at SA4, the spatial random effects remained high, and the precision of the estimates (credible intervals) decreased. This suggested that SEIFA was less effective at explaining variability in highly aggregated areas. Negative β coefficients across different scales indicated that higher SEIFA scores (representing better socioeconomic conditions) were associated with lower COVID-19 cases, highlighting the covariate's significance in the model.

Table 6
Posterior summary of BYM model results with covariate (SEIFA) for OA COVID-19 cases.

Geographic level	Parameter estimates							
	α Median (95 % CI)	β_1 Median (95 % CI)	β_2 Median (95 % CI)	β_3 Median (95 % CI)	β_4 Median (95 % CI)	σ_u^2 Median (95 % CI)	σ_v^2 Median (95 % CI)	FSV $= \frac{\sigma_u^2}{\sigma_u^2+\sigma_v^2}$
SA1	−4.57 (−6.13, −3.53)	1.44 (0.46, 2.76)	1.43 (0.46, 2.76)	1.42 (0.47, 2.78)	1.23 (0.27, 2.65)	1.62 (1.09, 2.36)	0.0052 (0.002,0.019)	0.0027
SA2	−0.54 (−0.77, −0.30)	0.18 (−0.12, 0.47)	0.47 (0.19, 0.75)	0.47 (0.19, 0.75)	0.72 (0.45, 1.01)	0.002 (0.001, 0.004)	0.24 (0.17, 0.32)	0.99
SA3	−0.48 (−0.73, −0.22)	0.05 (−0.29, 0.39)	0.24 (−0.11, 0.59)	0.58 (0.24, 0.91)	0.49 (0.11, 0.86)	0.18 (0.01, 0.39)	0.012 (0.002, 0.121)	0.17
SA4	−0.49 (−0.77, −0.22)	0.23 (−0.19, 0.69)	0.28 (−0.21, 0.77)	0.40 (−0.01, 0.81)	0.80 (0.37, 1.23)	0.02 (0.002, 0.31)	0.03 (0.003, 0.11)	0.60

Table 7
Posterior summary of BYM model results with covariate OA for LA COVID-19 cases.

Geographic level	Parameter estimates				
	α Median (95 % CI)	γ Median (95 % CI)	σ_u^2 Median (95 % CI)	σ_v^2 Median (95 % CI)	FSV $= \frac{\sigma_u^2}{\sigma_u^2+\sigma_v^2}$
SA1	−1.99 (−2.04, −1.94)	0.31 (0.15, 0.46)	0.62 (0.58, 0.67)	0.002 (0.001, 0.007)	0.006
SA2	−1.37 (−1.46, −1.28)	−0.002 (−0.037, 0.034)	1.02 (0.78, 1.33)	0.311 (0.24, 0.39)	0.23
SA3	−1.55 (−1.74, −1.36)	−0.0005 (−0.028, 0.026)	1.619(0.951, 2.966)	0.273 (0.087, 0.684)	0.12
SA4	−1.12 (−2.29, 0.38)	−0.001 (−0.02, 0.01)	3.86 (0.82, 14.04)	1.02 (0.17, 4.24)	0.20

In Table 6, positive associations (β coefficients) were observed between OA COVID-19 incidence and socioeconomic status across all levels, with more precise estimates at finer scales (SA1 and SA2). These positive associations indicated that higher SEIFA scores (i.e., better socioeconomic conditions) were associated with higher OA COVID-19 incidence, likely due to increased international travel and exposure. Spatially structured variance (σ_u^2) was highest at SA1 (1.62), capturing detailed spatial heterogeneity, but decreased significantly at coarser scales (SA2: 0.002). Conversely, unstructured variance (σ_v^2) increased at coarser scales, indicating reduced resolution in capturing localized variability. The fraction of spatial variation was highest at SA2 (0.993), which suggests that this level balances spatial resolution and model stability. At SA1, the fraction was very low (0.0027), indicating limited practical utility due to data sparsity.

In Fig. 3, higher SIR values were concentrated in remote and northern regions, which indicated a significant socioeconomic influence on COVID-19 incidence. Areas with lower SEIFA scores, representing higher disadvantage, showed elevated SIRs, particularly at finer spatial scales. As aggregation increased, the patterns smoothed out, but regional disparities remained visible, emphasizing SEIFA's impact.

The inclusion of the OA covariate had minimal impact on improving model fit, particularly at coarser spatial scales. At SA1, the γ coefficient was positive (0.31) with a relatively narrow credible interval, indicating a weak association between OA cases and locally acquired COVID-19 cases. However, at SA2, SA3, and SA4, the γ coefficient was close to zero with credible intervals crossing zero, suggesting little to no effect of OA as a predictor. The spatial random effects (σ_u^2) remained high at coarser scales, which indicated persistent spatial heterogeneity even with the inclusion of OA. The unstructured random effects (σ_v^2) also increased with aggregation, further suggesting that OA was not an effective explanatory variable for local COVID-19 transmission.

In the OA-based maps in Fig. 4, similar trends of higher SIR values, as seen in Fig. 3, were observed; however, the influence of OA cases was less pronounced compared to SEIFA covariate. While the inclusion of OA reduced spatial autocorrelation, it did not capture socioeconomic disparities as effectively. At finer spatial scales, OA-based maps displayed less distinct clustering than SEIFA-based maps. As aggregation increased, the patterns became smoother, but regional variations became less defined, which demonstrated the limited explanatory power of OA cases in capturing localized COVID-19 transmission dynamics.

The plots in Fig. 5 collectively highlighted how aggregation levels influenced the estimation of these parameters, with finer levels (SA1 and SA2) offering more precise estimates (narrower credible intervals) and coarser levels (SA3 and SA4) introducing increased uncertainty (larger credible intervals).

The six plots in Fig. 6 displayed the posterior summaries of model parameters with SEIFA covariate on LA COVID-19 cases, including baseline log-risk (α), socioeconomic covariate coefficients (β_1, β_2, β_3, β_4), spatial structured variance (σ_u^2), and spatial unstructured variance (σ_v^2) across geographic levels (SA1, SA2, SA3, and SA4). These plots illustrated the effect of

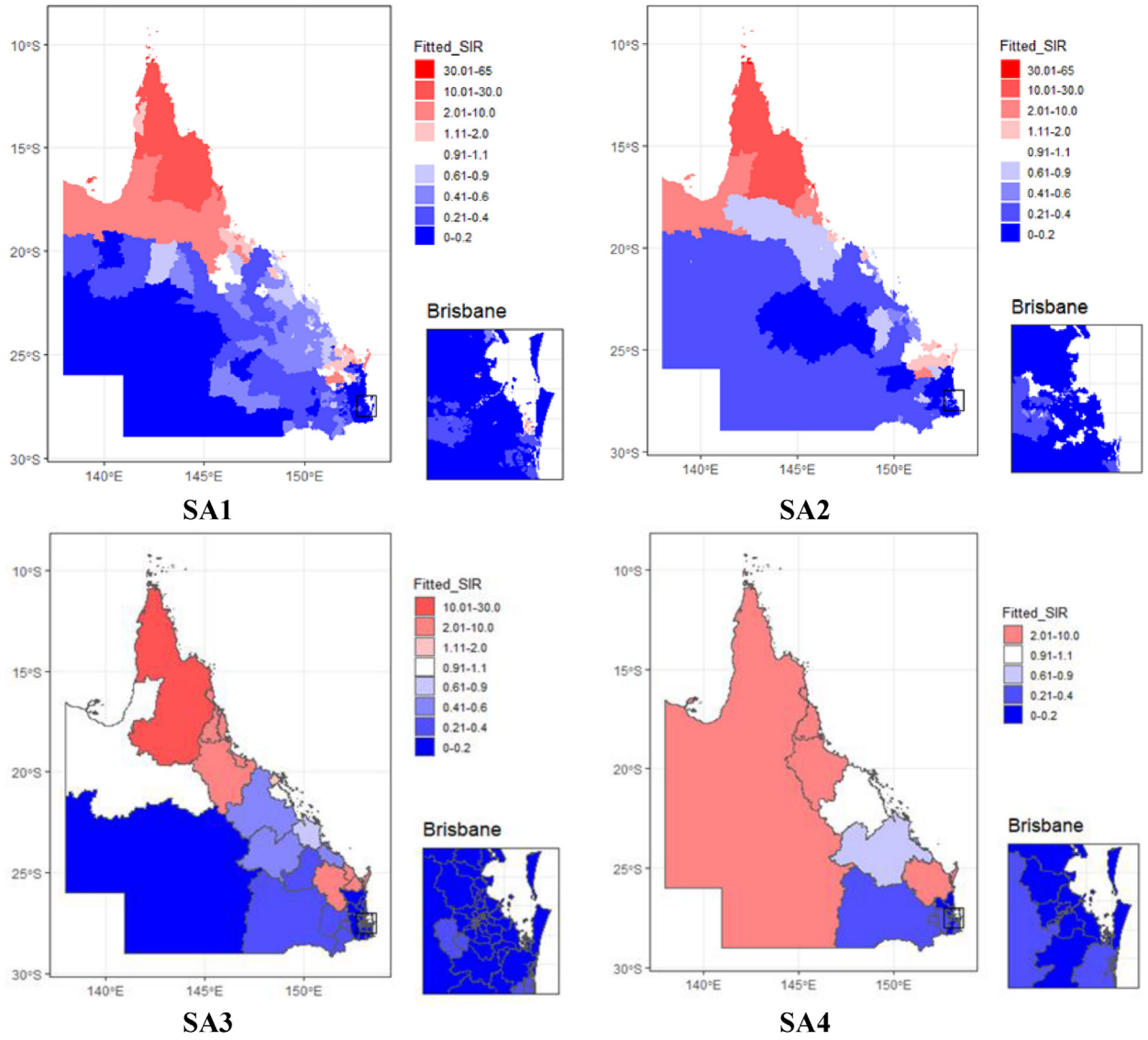

Fig. 3. Choropleth maps of the fitted SIR with covariate SEIFA for the LA COVID-19 counts at multiple levels in Queensland, Australia.

aggregation levels on parameter estimation, showing narrower credible intervals at finer scales and wider intervals with increased uncertainty at coarser scales. The variations in β coefficients highlighted the influence of socioeconomic covariates on the model outcomes.

In Fig. 7, parameters include baseline log-risk (α), socioeconomic coefficients (β_1-β_4), and spatial variances (σ_u^2, σ_v^2). SEIFA might not fully account for all variability in OA cases at the SA1 level, which reduced the model's precision at this level. This highlights the importance of considering data characteristics and covariate relevance when interpreting model results at different geographic levels.

In Fig. 8, the wider credible interval for γ at SA1 reflected the challenges of modeling data at highly granular levels, especially when the covariate does not fully account for localized variability or when data sparsity is a concern.

2.4. Results and discussion

The study investigated the impact of the MAUP on spatial analysis of sparse or uneven disease data using a Bayesian spatial BYM model in a geographically complex setting. The BYM model effectively captures both structured spatial dependencies and unstructured spatial random effects, making it particularly suitable for addressing spatial autocorrelation, a critical challenge in infectious disease epidemiology. Its advanced smoothing capabilities improve estimation in areas with limited or uneven data, which is often encountered in public health research. Compared to simpler spatial models or fixed-effects

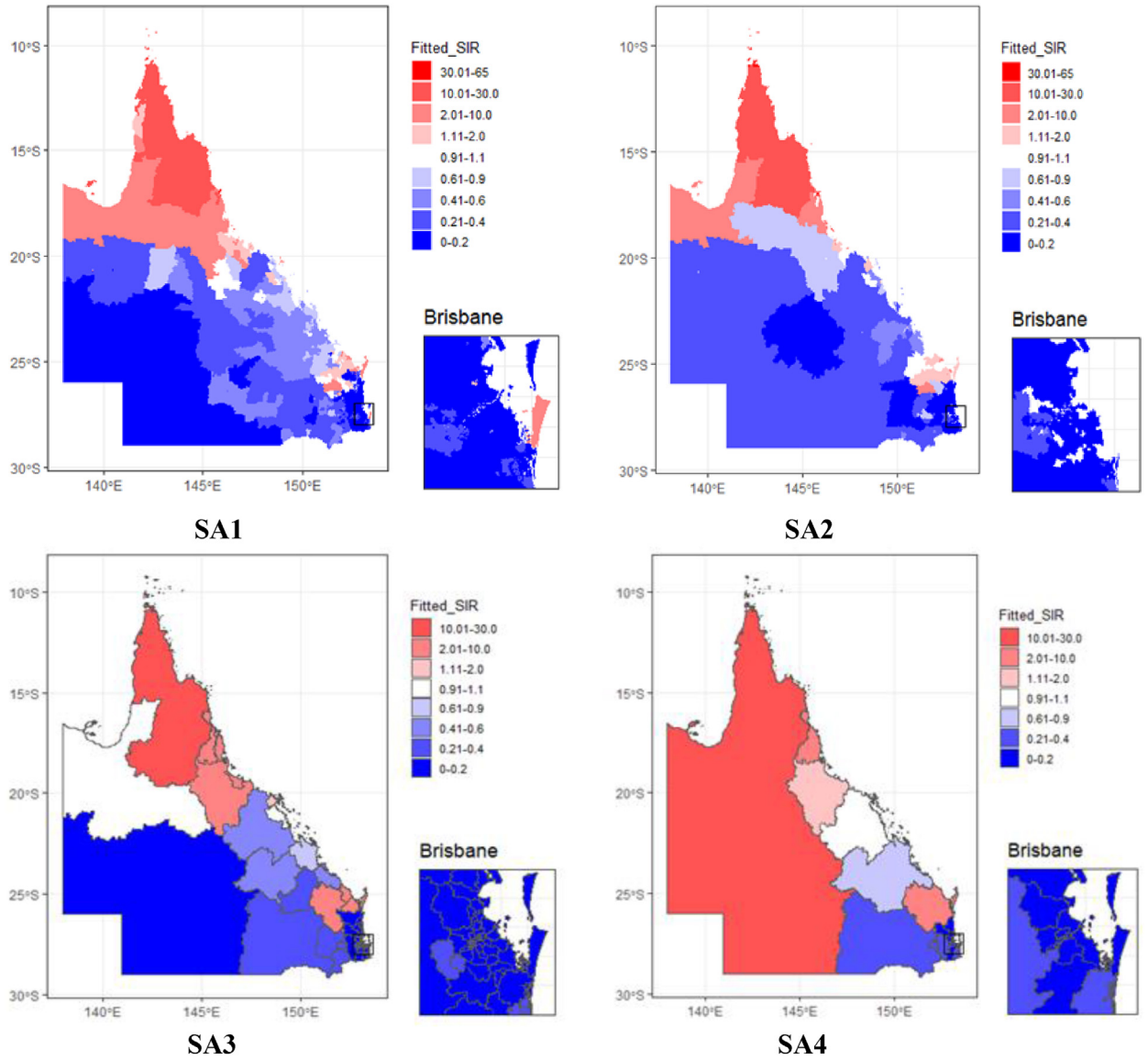

Fig. 4. Choropleth maps of the fitted SIR with covariate OA from the LA COVID-19 counts at multiple levels in Queensland, Australia.

frameworks, the BYM model offers greater flexibility and precision by accounting for both intrinsic spatial correlations and random noise, leading to more robust and reliable inference, particularly in complex spatial data structures. The analysis identified substantial differences across all evaluated aspects, including spatial patterns, estimates, and coefficient inferences.

A total of 31,276 LA COVID-19 cases were analysed, with the median number per area increasing from 0.4 at SA1 level to 280 at SA4 level. The mean incidence values also varied across aggregation levels, from 601.3 at SA1 to 913.9 at SA2, before dropping at SA3 and SA4 (Table 1). This fluctuation reflects a core feature of the MAUP, where aggregation alters incidence trends and statistical summaries. For example, while the maximum incidence values at SA1 and SA2 were nearly identical, the median jumped by 20 %, highlighting discrepancies in summary statistics caused by aggregation (Table 1). These changes are a clear indication of the MAUP, where spatial boundaries and aggregation levels influence data interpretation and model outcomes.

Significant spatial autocorrelation was observed at finer levels, with Moran's *I* values decreasing and becoming non-significant at broader scales (Table 2). The reduction in Moran's *I* reflected how aggregation smooths out localized clusters, masking fine-scale variations crucial for identifying disease hotspots. Missing fine-resolution clusters may lead to inaccurate public health responses, especially in sparsely populated areas where outbreaks are more likely to be localized. Our findings showed that spatial autocorrelation, while a key indicator of clustering, decreased as data were aggregated, demonstrating the influence of the MAUP on model inference.

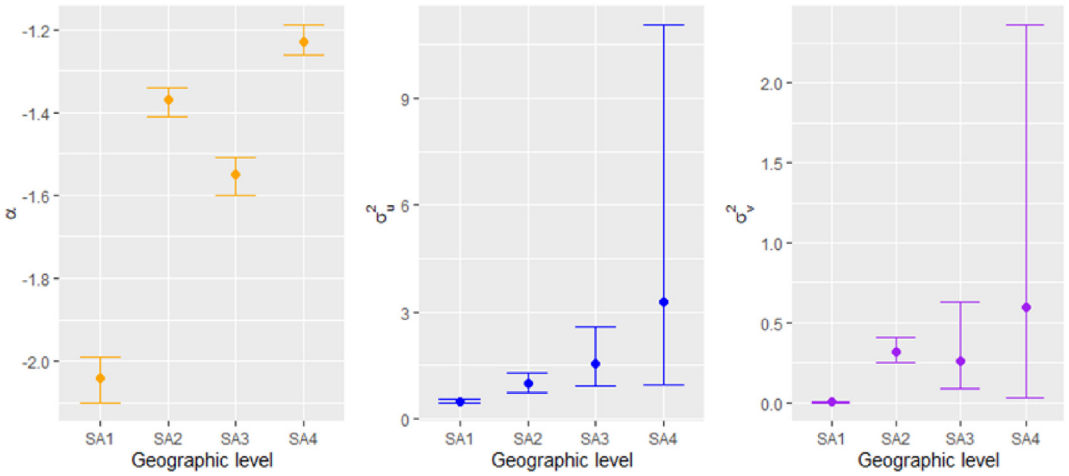

Fig. 5. The posterior summaries of model (without covariate) parameters, the baseline log-risk (α), spatial structured variance (σ_u^2), and spatial unstructured variance (σ_v^2) across different geographic levels (SA1, SA2, SA3, and SA4) for LA COVID-19 cases.

Incorporating the SEIFA covariate reduced spatial autocorrelation across scales, particularly at SA1 and SA2, by capturing socioeconomic-related spatial dependencies (Table 2). This reduction emphasized SEIFA's relevance in addressing the MAUP by retaining finer spatial details and reducing spatial clustering at smaller scales. In contrast, OA COVID-19 cases, used as an exploratory covariate, had limited relevance for predicting LA COVID-19 cases, particularly at finer scales where its effect was minimal ($\gamma = 0.31$) and diminished further at coarser levels (Table 2, Table 7). This suggests that imported cases are influenced by distinct factors, such as travel policies, quarantine protocols, and international exposure risks, which do not directly impact community-level transmission (Chinazzi et al., 2020; Wells et al., 2020). The decoupling between international and local transmission dynamics, combined with strict border controls, made OA cases a poor proxy for the local spread of COVID-19 (Zachreson et al., 2022).

LA COVID-19 cases were more influenced by local socioeconomic conditions captured by SEIFA, making it a more effective predictor in spatial models (Fontanet et al., 2023; Kok et al., 2021). The ability of SEIFA to capture socioeconomic disparities was particularly important at finer spatial scales (SA1, SA2); however, this effect lessened at coarser aggregation levels (SA3, SA4), where lower-resolution results yielded non-significant Moran's *I* p-values, indicating that the loss of information quality was likely due to spatial aggregation rather than true reductions in spatial clustering. These findings highlighted the importance of selecting appropriate covariates and spatial scales to mitigate MAUP effects in spatial models.

The observed and fitted SIR maps (Figs. 1 and 2) revealed substantial differences in spatial patterns across Queensland. Remote regions generally showed higher SIR values, with greater spatial detail and localized patterns at finer scales (SA1 and SA2). Coarser levels (SA3 and SA4) smoothed out these variations, potentially obscuring localized outbreaks. This illustrated the MAUP's role in influencing disease risk interpretation. SEIFA-based maps (Fig. 3) highlighted elevated COVID-19 incidence in socioeconomically disadvantaged regions, particularly in remote and northern areas, reinforcing the impact of socioeconomic gradients (β coefficients). These results emphasize the importance of selecting covariates that are contextually relevant to the disease being modelled.

Model parameters (α, β, γ, σ_u^2 and σ_v^2) displayed narrower credible intervals (Figs. 5, 6 and 8) at finer scales (SA1 and SA2), reflecting greater precision in estimates. Higher levels (SA3 and SA4) showed wider intervals, indicating increased uncertainty. The fraction of spatial variation, which quantifies the contribution of spatially structured random effects, was highest at SA2 across models, suggesting that SA2 strikes a balance between spatial resolution and model stability (Table 5). At SA1, the fraction of spatial variation was very low due to data sparsity, making large-scale modelling less reliable at this level despite the detailed spatial patterns it offers (Fontanet et al., 2023; Kok et al., 2021). This trade-off highlighted how the MAUP affects model stability, with finer scales capturing more localized variations but being more susceptible to issues like data sparsity. In contrast, broader levels (SA3 and SA4) masked localized patterns and introduced greater uncertainty in spatial effects, reducing their reliability for detailed infectious disease modelling.

Further analysis of 1289 OA COVID-19 cases using the BYM model across geographic aggregation levels (SA1 to SA4) reinforced the impact of MAUP (Table 1). While the mean case counts naturally increased with aggregation, the variability across geographic levels also rose, as indicated by the increasing SD values (Table 1). Positive spatial autocorrelation was observed at all levels, with Moran's *I* values decreasing from SA1 to SA4 (Table 3), reflecting the masking of localized patterns at broader scales. Incorporating SEIFA IRSD as a covariate significantly reduced spatial autocorrelation, particularly at finer levels, and resulted in non-significant spatial autocorrelation at broader levels (Table 3), indicating SEIFA's effectiveness in capturing socioeconomic-related spatial dependencies. However, unlike LA COVID-19 incidence, positive β coefficients (Table 6) across different scales indicated that higher SEIFA scores were associated with higher OA COVID-19 incidence, likely due to increased international

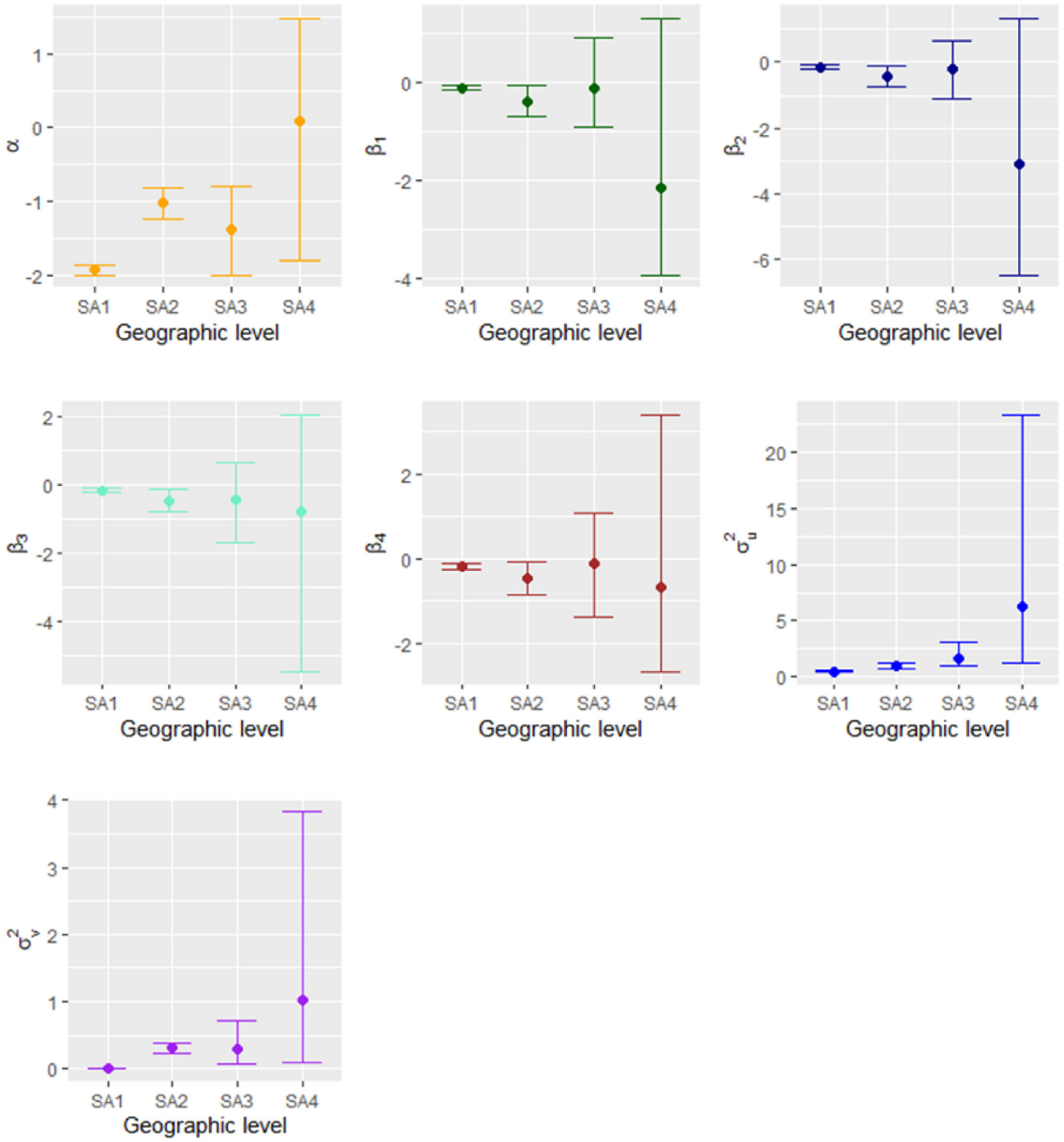

Fig. 6. Credible intervals of model parameters with SEIFA covariate on LA COVID-19 cases.

travel, exposure risks, and access to resources. This aligns with the findings in (Zhang et al., 2022) where the initial waves of COVID-19 were predominantly characterized by imported cases among individuals with higher socioeconomic status (SES). The spatially structured variance (σ_u^2) was highest at SA1 (1.62), capturing detailed spatial heterogeneity, but decreased at higher scale, SA2 (0.002) (Table 6). Conversely, unstructured variance (σ_v^2) increased at coarser scales, indicating reduced resolution in capturing localized variability. The fraction of spatial variation was highest at SA2 (0.99), suggesting that this level balances spatial resolution and model stability. At SA1, the fraction was very low (0.0027), indicating limited practical utility due to data sparsity (Table 6).

These findings reinforce the need for appropriate spatial scales and covariates in spatial modelling. In our study, SA2 emerged as the optimal level for infectious disease modelling, providing detailed spatial patterns while mitigating the risks of

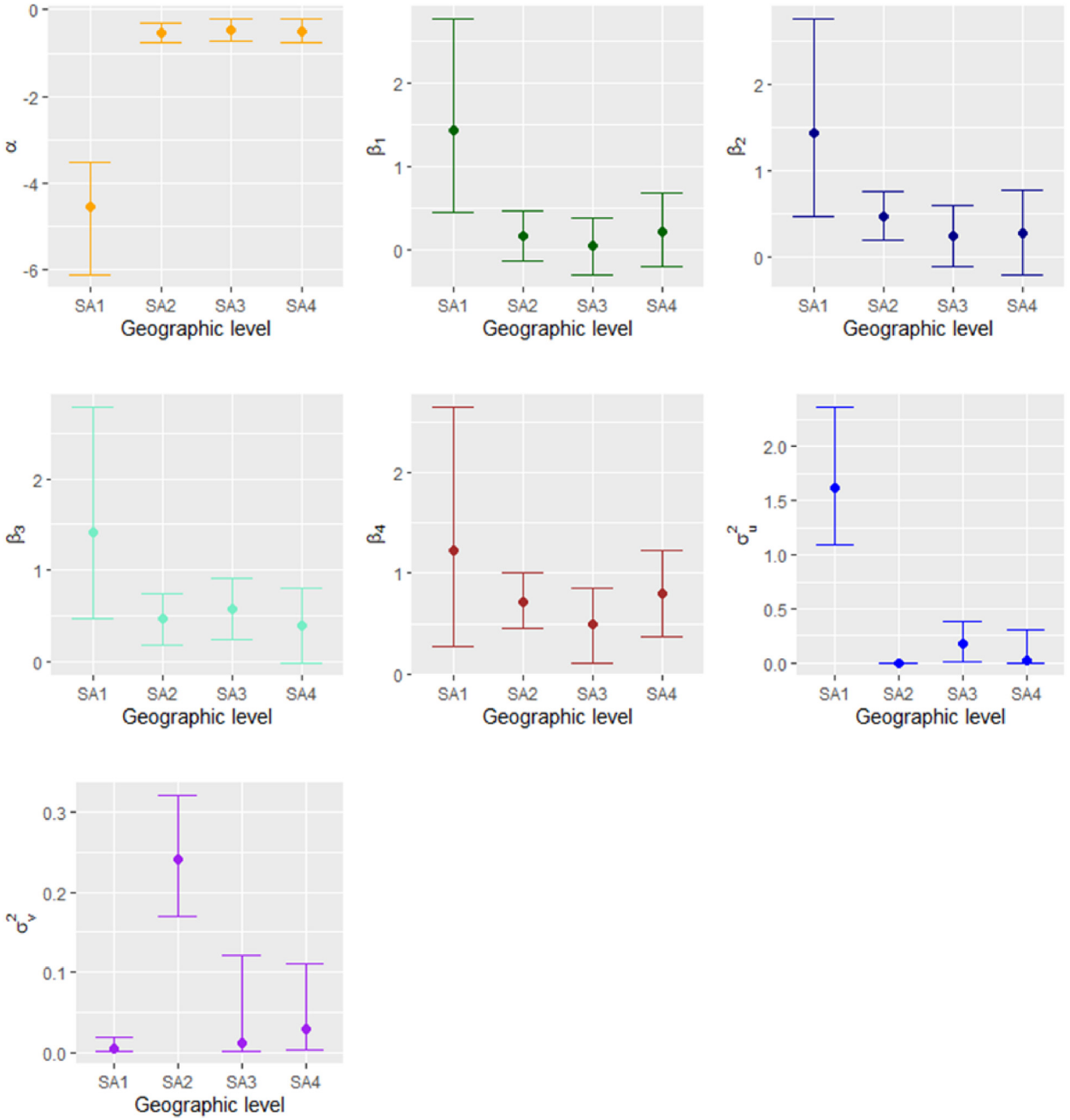

Fig. 7. Credible intervals of model parameters for the model fitted to OA COVID-19 cases with SEIFA as a covariate across geographic levels (SA1-SA4).

over-smoothing or excessive complexity. SEIFA effectively captured spatial dependencies and highlighted socioeconomic disparities in disease spread, underlines the importance of relevant covariates in addressing MAUP effects.

Our study highlights significant public health implications regarding the choice of spatial scale for disease analysis. Finer-scale resolutions (SA1 and SA2), were effective in detecting localized clusters of COVID-19, allowing for precise, community-level interventions, including targeted lockdowns, testing efforts, and resource allocation. These detailed scales are particularly valuable in urban settings or areas with heightened socioeconomic vulnerability, where rapid and localized responses are essential. In contrast, coarser levels (SA3 and SA4) provided a useful overview of regional trends, supporting strategic planning and public communication at higher levels. However, the increased aggregation at these levels often masked local variability, potentially obscuring emerging hotspots and reducing the effectiveness of localized public health measures. Recognizing these trade-offs is critical for designing spatially appropriate public health responses to COVID-19 and similar infectious disease threats.

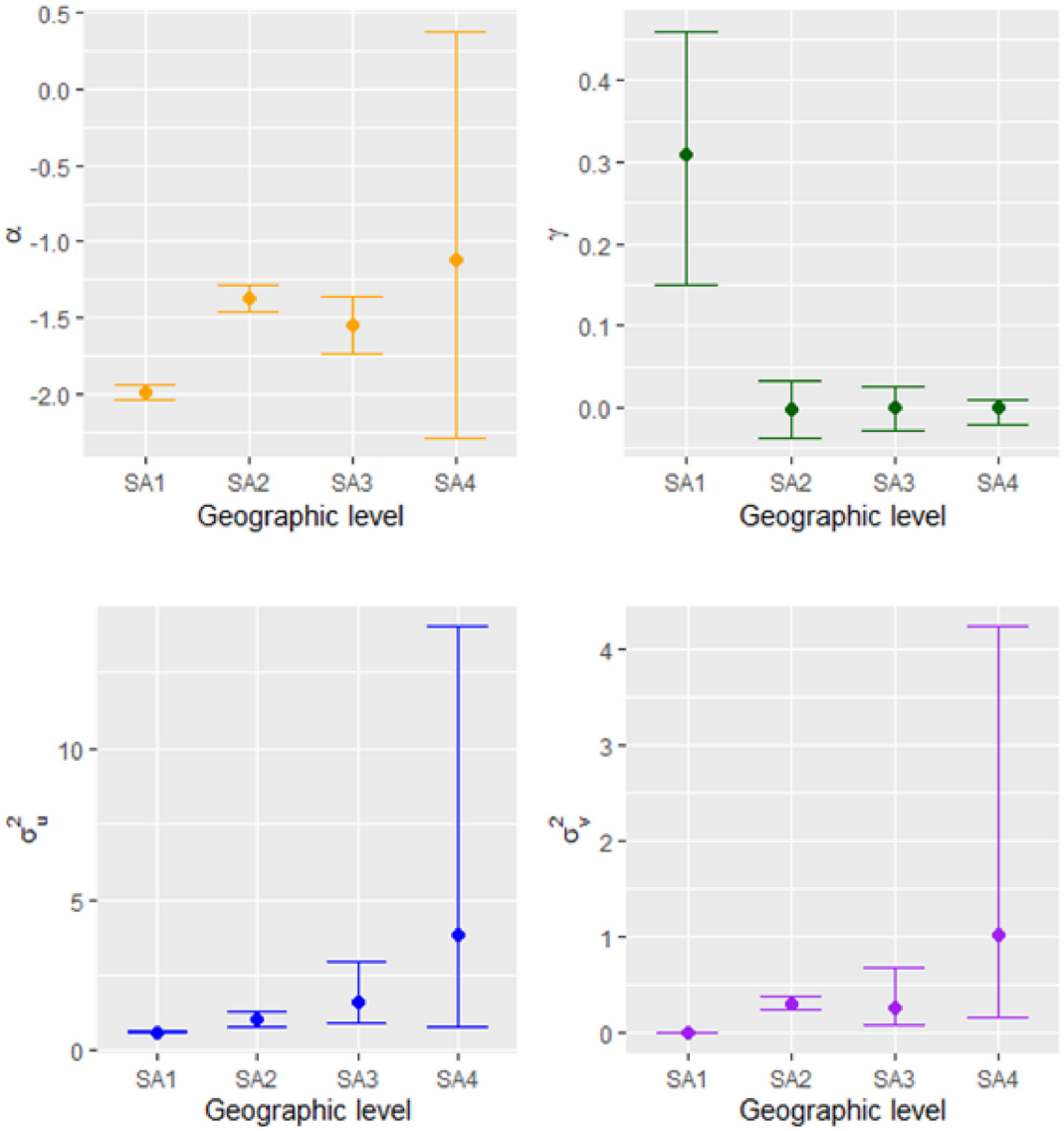

Fig. 8. Credible intervals of model parameters for the model fitted to LA COVID-19 cases with OA cases as a covariate across geographic levels (SA1-SA4).

This study also has several limitations. Due to data constraints, we derived SA1-level data from SA2-level data, which may have impacted the results despite following the recommended methodologies. This limitation highlights a key aspect of the MAUP, where both outcome measures and covariates are influenced by the level of spatial aggregation. Moreover, this study addressed only the scaling aspect of the MAUP. The zoning aspect, which refers to how different boundary definitions impact results, was not explored in this study. Zoning choices can affect the aggregation of data, potentially introducing biases or changes in the interpretation of spatial patterns, especially in the context of infectious disease transmission (Tuson et al., 2020). These limitations highlight the importance of carefully considering spatial aggregation when interpreting findings. Furthermore, the absence of temporal data limits the study's ability to examine how the MAUP influences spatial patterns across different time periods. Incorporating temporal dimensions, particularly for evolving diseases like COVID-19, could provide valuable insights into the dynamic implications of the MAUP.

2.5. Conclusions

This study highlights the significant impact of the MAUP on spatial epidemiological modelling of COVID-19 in Queensland, Australia. Our analysis revealed that finer spatial scales, such as SA1 and SA2, retained localized patterns and displayed significant spatial autocorrelation, while coarser scales (SA3 and SA4) smoothed out variations, potentially masking clusters of outbreaks. The inclusion of the SEIFA covariate effectively reduced spatial autocorrelation across all scales, particularly at

finer levels, reinforcing its importance in capturing socioeconomic disparities. OA COVID-19 cases as covariate also reduced spatial autocorrelation, but was less effective than SEIFA, especially at finer scales. LA COVID-19 cases showed negative association with higher SEIFA scores while modelling OA COVID-19 cases showed a positive association with higher SEIFA scores, indicating differing spatial influence.

Despite SA1 showing a lower fraction of spatial variation in our study, its ability to retain the highest spatial detail makes it a valuable scale for modelling when sufficient data exists. However, challenges like data sparsity and computational burden must be carefully managed. SA2 emerged as a more practical level, balancing spatial resolution, stability, and interpretability while mitigating the risks associated with both excessively fine (SA1) and overly coarse (SA3 and SA4) aggregations. Therefore, we recommend fitting models at both SA1 and SA2 levels where possible to leverage their respective strengths, allowing for a comprehensive analysis of infectious disease patterns. Similar conclusions may be drawn in countries with geographic areas and diverse population distributions comparable to Australia when analysing infectious disease data.

Future studies should focus on exploring intermediate aggregation levels or custom-defined zones based on factors like population density, access to health service, or socio-demographic characteristics. Implementing multi-scale modelling approaches that account for spatial dependencies at different levels could improve our understanding of complex disease dynamics and help mitigate the impact of the MAUP. Additionally, extending this analysis across Australia and replicating the study in other countries with diverse geographic and socio-economic contexts would enhance the validity and generalizability of finding.

In conclusion, this study emphasizes the critical role of selecting appropriate spatial scales and covariates in spatial modelling to mitigate the impact of the MAUP. While both finer and coarser aggregation levels offer valuable insights for public health data analysis, it is essential to understand how spatial aggregation influences the interpretation of disease trends. A multi-scale approach, particularly at finer resolutions (e.g., SA1 and SA2), combined with context-specific covariates, can help reduce MAUP-related biases and enhance the precision and equity of public health interventions.

CRediT authorship contribution statement

Shovanur Haque: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Aiden Price:** Writing – review & editing. **Kerrie Mengersen:** Writing – review & editing. **Wenbiao Hu:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Ethics approval and consent to participate

Ethics committee approval for the data was not sought as we analysed routinely collected, publicly available de-identified data.

Availability of data and materials

The data extracted, prepared, and analysed to support the findings of this study are publicly available and can be accessed at https://www.data.qld.gov.au/dataset/queensland-covid-19-case-line-list-location-source-of-infection/resource/1dbae506-d73c-4c19-b727-e8654b8be95a.

Funding

The National Health and Medical Research Council (NHMRC) Special Initiative in Human Health and Environmental Change (Grant No. 2008937).

Declaration of competing interest

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.

Acknowledgements

We acknowledge the HEAL (Healthy Environments And Lives) National Research Network, which receives funding from the National Health and Medical Research Council (NHMRC) Special Initiative in Human Health and Environmental Change (Grant No. 2008937). We also acknowledge the National Foundation for Australia-China Relations (Grant No. 220011), the Australian Department of Foreign Affairs and Trade.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.idm.2025.05.003.

References

ABS. Socio-Economic Indexes for Areas (SEIFA), Australia methodology. https://www.abs.gov.au/methodologies/socio-economic-indexes-areas-seifa-australia-methodology/2021.

ABS. (2021). Socio-economic Indexes for areas. https://www.abs.gov.au/statistics/people/people-and-communities/socio-economic-indexes-areas-seifa-australia/latest-release.

Australian Bureu of Statistics - Regional population. https://www.abs.gov.au/statistics/people/population/regional-population/latest-release#data-downloads.

Australian Statistical Geography Standard (ASGS) Edition 3. https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/latest-release.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics, 43*, 1–20.

Best, N., Richardson, S., & Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research, 14*(1), 35–59.

Briz-Redón, Á. (2022). A Bayesian shared-effects modeling framework to quantify the modifiable areal unit problem. *Spatial Statistics, 51*, Article 100689.

Burden, S., & Steel, D. (2016). Constraint choice for spatial microsimulation. *Population, Space and Place, 22*(6), 568–583.

Carlin, B. P., & Xia, H. (1999). Assessing environmental justice using bayesian hierarchical models: Two case studies. *Journal of Exposure Analysis and Environmental Epidemiology, 9*(1).

Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., Pastore y Piontti, A., Mu, K., Rossi, L., & Sun, K. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science, 368*(6489), 395–400.

Cliff, A. D., & Ord, J. K. (1968). *The problem of spatial autocorrelation*. University.

Correspondences. https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/access-and-downloads/correspondences.

Cramb, S., Duncan, E., Baade, P., & Mengersen, K. L. (2020). A comparison of bayesian spatial models for cancer incidence at a small area level: Theory and performance. *Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018*, 245–274.

Duncan, E. W., Cramb, S. M., Aitken, J. F., Mengersen, K. L., & Baade, P. D. (2019). Development of the Australian cancer atlas: Spatial modelling, visualisation, and reporting of estimates. *International Journal of Health Geographics, 18*, 1–12.

Escaramís, G., Carrasco, J. L., & Ascaso, C. (2008). Detection of significant disease risks using a spatial conditional autoregressive model. *Biometrics, 64*(4), 1043–1053.

Faramarzi, A., Javan-Noughabi, J., Mousavi, S. A., Bahrami Asl, F., & Shabanikiya, H. (2022). Socioeconomic status and COVID-19-related cases and fatalities in the world: A cross-sectional ecological study. *Health Science Reports, 5*(3), Article e628.

Fontanet, C. P., Carlos, H., Weiss, J. E., Diaz, M. C. G., Shi, X., Onega, T., & Loehrer, A. P. (2023). Evaluating geographic health disparities in cancer care: Example of the modifiable areal unit problem. *Annals of Surgical Oncology, 30*(12), 6987–6989.

Fotheringham, A. S., & Wong, D. W. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A, 23*(7), 1025–1044.

Getis, A. (2009). Spatial autocorrelation. In *Handbook of applied spatial analysis: Software tools, methods and applications* (pp. 255–278). Springer.

Geweke, J. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*.

Gregorio, D. I., DeChello, L. M., Samociuk, H., & Kulldorff, M. (2005). Lumping or splitting: Seeking the preferred areal unit for health geography studies. *International Journal of Health Geographics, 4*, 1–10.

Hawkins, R. B., Charles, E. J., & Mehaffey, J. H. (2020). Socio-economic status and COVID-19–related cases and fatalities. *Public Health, 189*, 129–134.

Kim, D. (2021). Predicting the magnitude of residual spatial autocorrelation in geographical ecology. *Ecography, 44*(7), 1121–1130.

Kok, M. R., Tuson, M., Yap, M., Turlach, B., Boruff, B., Vickery, A., & Whyatt, D. (2021). Impact of the modifiable areal unit problem in assessing determinants of emergency department demand. *Emergency Medicine Australasia, 33*(5), 794–802.

Lam, P. (2002). Convergence diagnostics. *Lecture presented at government*.

Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and spatio-temporal epidemiology, 2*(2), 79–89.

Lee, D. (2013). CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software, 55*(13), 1–24.

Manley, D. (2021). Scale, aggregation, and the modifiable areal unit problem. In *Handbook of regional science* (pp. 1711–1725). Springer.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*(1/2), 17–23.

Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the Institute of British Geographers*, 459–472.

Oyana, T. J. (2020). *Spatial analysis with r: Statistics, visualization, and computational methods*. CRC press.

Politi, J., Martín-Sánchez, M., Mercuriali, L., Borras-Bermejo, B., Lopez-Contreras, J., Vilella, A., Villar, J., Orcau, A., de Olalla, P. G., & Rius, C. (2021). Epidemiological characteristics and outcomes of COVID-19 cases: Mortality inequalities by socio-economic status, barcelona, Spain, 24 february to 4 may 2020. *Euro Surveillance, 26*(20), Article 2001138.

Queensland COVID-19 Case Line List by Location and Source of Infection. https://www.data.qld.gov.au/dataset/queensland-covid-19-case-line-list-location-source-of-infection/resource/1dbae506-d73c-4c19-b727-e8654b8be95a.

Roquette, R., Painho, M., & Nunes, B. (2017). Spatial epidemiology of cancer: A review of data sources, methods and risk factors. *Geospatial health, 12*(1).

Team, R. C. (2000). R language definition. Vienna, Austria. *R foundation for statistical computing, 3*(1), 116.

Team, R. C. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: Foundation for Statistical Computing.

Tuson, M., Kok, M. R., Yap, M., Vickery, A., Boruff, B., Murray, K., Turlach, B., & Whyatt, D. (2018). Reducing bias in multivariate analyses due to the modifiable areal unit problem. *International Journal of Population Data Science, 3*(4).

Tuson, M., Yap, M., Kok, M. R., Boruff, B., Murray, K., Vickery, A., Turlach, B. A., & Whyatt, D. (2020). Overcoming inefficiencies arising due to the impact of the modifiable areal unit problem on single-aggregation disease maps. *International Journal of Health Geographics, 19*, 1–18.

Tuson, M., Yap, M., Kok, M. R., Murray, K., Turlach, B., & Whyatt, D. (2019). Incorporating geography into a new generalized theoretical and statistical framework addressing the modifiable areal unit problem. *International Journal of Health Geographics, 18*, 1–15.

Wakefield, J., & Lyons, H. (2010). Spatial aggregation and the ecological fallacy. *Handbook of spatial statistics, 20103158*, 541–558.

Wang, Y., & Di, Q. (2020). Modifiable areal unit problem and environmental factors of COVID-19 outbreak. *Science of the Total Environment, 740*, Article 139984.

Wells, C. R., Sah, P., Moghadas, S. M., Pandey, A., Shoukat, A., Wang, Y., Wang, Z., Meyers, L. A., Singer, B. H., & Galvani, A. P. (2020). Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proceedings of the National Academy of Sciences, 117*(13), 7504–7509.

Wong, D. (2004). The modifiable areal unit problem (MAUP). *WorldMinds: Geographical perspectives on 100 problems*. Dordrecht: Springer.

Zachreson, C., Shearer, F. M., Price, D. J., Lydeamore, M. J., McVernon, J., McCaw, J., & Geard, N. (2022). COVID-19 in low-tolerance border quarantine systems: Impact of the delta variant of SARS-CoV-2. *Science Advances, 8*(14), Article eabm3624.

Zhang, A., Shi, W., Tong, C., Zhu, X., Liu, Y., Liu, Z., Yao, Y., & Shi, Z. (2022). The fine-scale associations between socioeconomic status, density, functionality, and spread of COVID-19 within a high-density city. *BMC Infectious Diseases, 22*(1), 274.