scientific reports

OPEN

Spatial non-parametric Bayesian clustered coefficients

Wala Draidi Areed[1]✉, Aiden Price[1], Helen Thompson[1], Reid Malseed[2] & Kerrie Mengersen[1]

In the field of population health research, understanding the similarities between geographical areas and quantifying their shared effects on health outcomes is crucial. In this paper, we synthesise a number of existing methods to create a new approach that specifically addresses this goal. The approach is called a Bayesian spatial Dirichlet process clustered heterogeneous regression model. This non-parametric framework allows for inference on the number of clusters and the clustering configurations, while simultaneously estimating the parameters for each cluster. We demonstrate the efficacy of the proposed algorithm using simulated data and further apply it to analyse influential factors affecting children's health development domains in Queensland. The study provides valuable insights into the contributions of regional similarities in education and demographics to health outcomes, aiding targeted interventions and policy design.

Spatial data analysis in public health often involves statistical models for areal data, which aggregate health outcomes over administrative units like states, counties, or zip codes[1] . Statistical models for areal data typically aim to provide estimates of geospatial outcomes of interest, find boundaries between abrupt changes in spatial patterns, describe smoothly varying spatial trends, identify and characterise spatial clusters, and so on. These models have been extensively employed in various fields, including geography, econometrics, ecology, epidemiology and public health.

When spatial dependence is present in the data, traditional statistical models that assume independent observations are inadequate and can produce biased estimates of outcomes of interest. This necessitates the use of alternative spatial models that incorporate this dependence[2]. There is a very wide range of spatial models tailored to the inferential aim, the nature of the spatial dependence, and the type of data.

Recent advancements in methods for spatial boundary detection have focused on model-based approaches which focus on probabilistic uncertainty quantification[3,4]. An exemplar paper is by Lee[5], who employs stochastic models for adjacency matrices in order to identify edges between regions with significant differences in health outcomes. Elaborations of this method have included control for multiple comparisons[6] and Bayesian hierarchical approaches[7]. Other approaches in boundary detection include integrated stochastic processes[8], stochastic edge mixed-effects models[9], and methods for estimating adjacencies in areal modelling contexts[10,11]. This modelling introduces spatial dependence using stochastic models on graphs, where nodes represent regions, and edges connect neighbouring regions[12] . Other methods include Markov random fields using undirected graphs[13,14] or directed acyclic graphical autoregression (DAGAR) models[15].

The most common regression methods for geographically referenced data are spatial linear regression[16] and spatial generalized linear regression[17]. However, these models assume that the coefficients of the explanatory variables are constant across space, which can be overly restrictive for large regions where the regression coefficients may vary spatially. Many methods include longitude and latitude coordinates as location variables, while others account for spatial variability in the model by including an additive spatial random effect for each location. This technique has been applied to linear models by Cressie[16], and generalized linear models by Diggle[17].

Numerous methods have been developed to identify and describe smoothly varying patterns of regression coefficients, such as Gelfand's[18] spatially varying coefficient processes (SVCM) and the spatial expansion methods proposed by Casetti[19]. In a follow-up paper, Casetti and Jones[20] treat the regression coefficients that vary spatially as a function of expansion variables.

An alternative popular method for capturing smoothly varying spatial patterns is through geographically weighted regression (GWR)[21]. The GWR fits a local weighted regression model at the location of each observation and captures spatial information by accounting for nearby observations, using a weight matrix defined by a kernel function[21]. This approach has been extended in a variety of ways, for example to a Cox survival model for spatially dependent survival data to explore how geographic factors impact time-to-event outcomes[22]. Unlike SVCM, GWR does not assume a specific functional form for the relationship between covariates and the response

[1]School of Mathematical Science, Centre for Data Science, Queensland University of Technology, Brisbane, QLD, Australia. [2]Children's Health Queensland, Brisbane, QLD, Australia. ✉email: wala.draidi@hotmail.com

variable. This flexibility is advantageous when dealing with complex spatial data where relationships may vary across space[23]. The GWR can also be computationally more efficient, especially for large datasets, and the results from GWR are often more interpretable as they provide localized parameter estimates for each spatial location[23]. This allows for a deeper understanding of how relationships between variables vary across space, which may be more intuitive for practitioners and policymakers.

The traditional GWR models are typically fit using a frequentist framework. A critical limitation of this approach is the violation of the usual assumption of non-constant variation between observations, and the resultant normality assumption for the errors[24]. Additionally, the usual frequentist approach struggles to address issues of model complexity, overfitting, variable selection and multicollinearity[24]. The stability and reliability of frequentist GWR might also yield unstable results or high variance when dealing with small sample sizes[25].

Bayesian GWR (BGWR) provides an appealing solution to these problems[26]. An exemplar paper is by Gelfand[27], who built a Bayesian model with spatially varying coefficients by applying a Gaussian process to the distribution of regression coefficients. Lesage[28] suggested an early version of BGWR, where the prior distribution of the parameters depends on expert knowledge. More recently, Ma[29] proposed BGWR based on the weighted log-likelihood and Liu[30] proposed BGWR based on a weighted least-squares approach. Spline approaches have been explored to estimate bivariate regression functions[31] and to accommodate irregular domains with complex boundaries or interior gaps[32]. Other studies, including those by Li et al.[30]and Wang et al.[33], also address this problem over irregular domains. However, all of these methods have a significant limitation, in that they cannot handle the possibility of a spatially clustered pattern in the regression coefficients. A recent development by Li et al.[34]is the spatially clustered coefficient (SCC) regression, which employs the fused LASSO to automatically detect spatially clustered patterns in the regression coefficients. Ma et al.[35] and Luo et al.[36] have proposed spatially clustered coefficient models using Bayesian approaches. Ma et al.[35] identified coefficient clusters based on the Dirichlet process, whereas Luo et al.[36] used a hierarchical modelling framework with a Bayesian partition prior model from spanning trees of a graph. Sugasawa et al. proposed spatially clustered regression (SCR)[37]. The selection of the appropriate number of clusters is a crucial aspect of clustering analysis. Most traditional methods require the number of clusters to be specified beforehand, which can limit their applicability in practice. This applies for K-means[38], hierarchical clustering[39] and Gaussian Mixture Models (GMM)[40]. These constraints pose challenges in scenarios where the optimal number of clusters isn't obvious from the start or varies across datasets, limiting the flexibility and adaptability of these clustering approaches in real-world applications. Dirichlet process mixture models (DPMM) have gained popularity in Bayesian statistics as they allow for an unknown number of clusters, increasing the flexibility of clustering analysis. However the DPMM does not account for the spatial information in the clusters.

In this paper, we synthesise two approaches, namely, a Bayesian GWR and a Bayesian spatial DPMM, to create a new method called the Bayesian spatial Dirichlet process clustered heterogeneous regression model. This method can detect spatially clustered patterns while considering the smoothly varying relationship between the response and the covariates within each group. We used a Bayesian geographically weighted regression algorithm to model the varying coefficients over the geographic regions and incorporated spatial neighbourhood information of regression coefficients. We then combined the regression coefficient and a spatial Dirichlet mixture process to perform the clustering. The approach is demonstrated using simulated data and then applied to a real-world case study on children's development in Queensland, Australia.

This approach meets the inferential aims of clustering and localised regression, for areal data. The clustering approach is preferred over the boundary detection approach in this context, since abrupt explainable changes in the spatial process are not anticipated and the prioritisation is to identify and profile broadly similar geospatial areas. The proposed use of GWR is also preferred over SCR, since GWR explicitly accounts for spatial variation in relationships between variables by estimating separate regression parameters for different locations, whereas SCR typically assumes spatial homogeneity within clusters and estimates a single set of parameters for each cluster. GWR is therefore capable of capturing fine-scale spatial variation, as it estimates parameters at the level of individual spatial areal units, whereas SCR aggregates data into clusters, potentially smoothing out fine-scale variation and overlooking localized patterns.

The strength of the proposed clustering method lies in several key features that set it apart from traditional clustering algorithms. Unlike K-Means and hierarchical clustering, which lack uncertainty measures, the proposed method provides clusters with associated uncertainty measures, enhancing interpretability and making them more valuable for decision-making and analysis. Additionally, the proposed method incorporates spatial neighbourhood information. This ensures that the resulting clusters not only reflect data similarity but also account for spatial heterogeneity. Furthermore, this Bayesian framework allows for better handling of outliers and uncertainties in the data by incorporating prior information. This adaptability is particularly beneficial in scenarios where data quality varies or is incomplete. The paper proceeds as follows: In Sect. "Results", we present the results obtained from applying the proposed algorithm to both simulated and real case studies. Following this, Sect. "Methods" provides a detailed explanation of the proposed algorithm. Section "Discussion" includes the discussion for the results and its limitation. In Sect. "Bayesian estimation and inference", we delve into the sampling procedure utilized in the proposed algorithms, along with an analysis of cluster accuracy. Finally, Sect. "Conclusion" concludes the paper by summarizing the findings.

Results

The proposed method, described in detail in the "Methods" section below, is evaluated through a simulation study and applied to a real-world case study. These applications demonstrate the effectiveness of the approach in simultaneously estimating and clustering spatially varying regression coefficients, with associated measures of uncertainty.

Simulation study

The simulation was structured based on the Georgia dataset with 159 regions introduced by Ma[35], where spatial sampling locations represented geographical positions for data collection. Specifically, we used centroids of geographical areas as the sampling locations.

For the simulation, six covariates (X_1 to X_6) were introduced as independent variables, each representing distinct features or characteristics at each sampling location. To incorporate spatial autocorrelation, we generated the covariates using multivariate normal distributions with spatial weight matrices derived from the distance matrix and parameter bandwidth.

The response variable (Y) in the simulation was generated using the GWR model[21]:

$$y(s) = \beta_0(s) + \sum_{k=1}^{K} \beta_k(s) \cdot X_k(s) + \varepsilon(s) \tag{1}$$

It is noteworthy that the true parameters (β_1 to β_6) of the GWR model varied spatially, implying that they differed across sampling locations based on the spatial weight matrices. This spatial variation allowed us to capture spatially dependent effects in the simulation[37]. We generate simulated spatial data with six covariates using the following steps. First, we generate 159 spatial locations, denoted as $s_1, \ldots, s_n$, based on the centroids of geographic areas. The locations are determined based on specific conditions related to the x and y coordinates of the centroids. Next, six covariates, $x_1(s_i), x_2(s_i), \ldots, x_6(s_i)$, are generated for each spatial location s_i. These covariates are derived from a spatial Gaussian process with mean zero and a covariance matrix defined by an exponential function $w_{ij} = \exp\left(-\frac{\|s_i - s_j\|}{\phi}\right)$, where ϕ is the bandwidth parameter with $\phi = 0.9$. This parameter influences the strength of spatial correlation in the covariates. Finally, the response at each location, $y(s_i)$, is generated according to a spatially varying linear model. This model includes the coefficients $\beta_1(s_i), \beta_2(s_i), \ldots, \beta_6(s_i)$ corresponding to the six covariates $x_1(s_i), x_2(s_i), \ldots, x_6(s_i)$, respectively. The error terms $\epsilon(s_i)$ are mutually independent.

To create distinct spatial patterns in the data, we visually partitioned the counties of Georgia into three large regions based on the spatial coordinates of centroids, defining true clustering settings. This approach enabled us to incorporate spatial autocorrelation, spatial variability, and true clustering effects in the simulated data. Figure 1 visualizes partition of the counties into three large regions with sizes, 51, 49 and 59 areas.

The code for the proposed algorithm can be found in the first author's GitHub https://github.com/waladraidi/Spatial-stick-breaking-BGWR.

The simulation was repeated 100 times. Figure 2 illustrates the spatial distribution of the posterior mean parameter coefficients for each location over the 100 replicates. This figure showcases the diverse spatial patterns and disparities in these parameter coefficients across the study area, providing insights into the geographical variation.

The performance of these posterior estimates was evaluated by mean absolute bias (MAB), mean standard deviation (MSD), mean of mean squared error (MMSE), as follows:

$$MAB = \frac{1}{159} \sum_{s=1}^{159} \frac{1}{100} \sum_{r=1}^{100} |\hat{\beta}_{s,k,r} - \beta_{s,k}|, \tag{2}$$

$$MSD = \frac{1}{159} \sum_{s=1}^{159} \sqrt{\frac{1}{99} \sum_{r=1}^{100} (\hat{\beta}_{s,k,r} - \bar{\hat{\beta}}_{s,k})^2}, \tag{3}$$

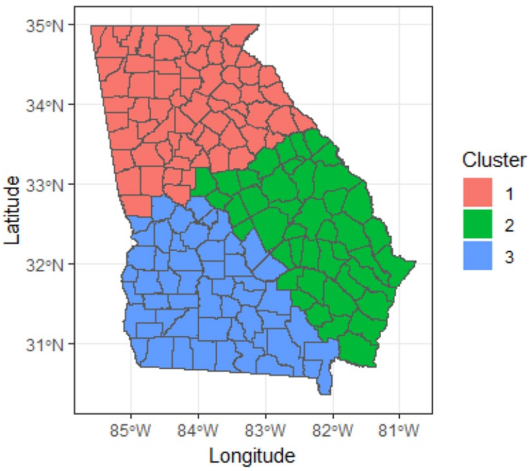

Figure 1. Regional cluster assignment for Georgia counties used for simulation study.

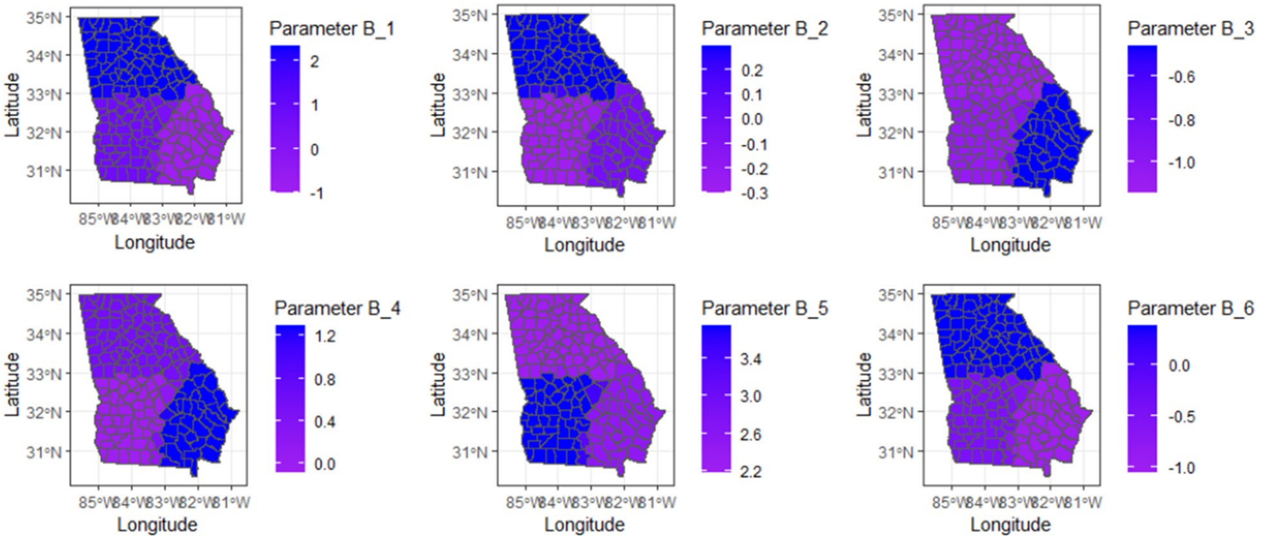

Figure 2. The spatial distribution of the posterior mean for the parameters obtained from the proposed model.

$$MMSE = \frac{1}{159}\sum_{s=1}^{159}\frac{1}{100}\sum_{r=1}^{100}(\hat{\beta}_{s,k,r} - \beta_{s,k})^2, \tag{4}$$

where $\bar{\hat{\beta}}_{s,k}$, is the average parameter estimate and has been calculated by the average of $\hat{\beta}_{s,k}$ ($s = 1, \ldots, 159; \quad k = 1, \ldots, 6$) in 100 simulations, and $\hat{\beta}_{s,k,r}$ denotes the posterior estimate for the k-th coefficient of county s in the r-th replicate. In each replicate, the MCMC chain length is set to be 10,000, and the first 2000 samples are discarded as burn-in. Therefore, we have 8000 samples for posterior inference. Table 1 reports the the three performance measures in Eqs. (2)–(4) for the simulated data. The parameter estimates are very close to the true underlying values and have a small MAB, MSD, and MMSE.

Three distinct clusters were found within the 159 regions. The spatial layout of these clusters is visualized in Fig. 3, where two cluster configurations are described in the later section on cluster configurations. Notably, when examining Fig. 3, it is clear that the cluster assignments derived from Dahl's and mode allocation approaches

Parameter	MAB	MSD	MMSE
$\hat{\beta}_1$	0.84	1.06	0.23
$\hat{\beta}_2$	0.32	0.42	1.35
$\hat{\beta}_3$	0.37	0.26	1.98
$\hat{\beta}_4$	0.37	0.42	1.19
$\hat{\beta}_5$	1.37	2.72	0.43
$\hat{\beta}_6$	0.94	1.21	0.55

Table 1. The performance of parameter estimates from the proposed model, where *MAB* mean absolute basis, *MSD* mean squared deviation, and *MMSE* mean of mean squared error.

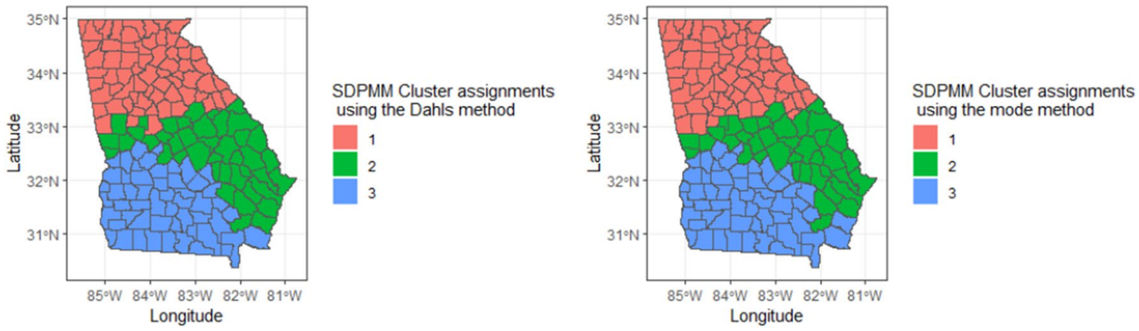

Figure 3. (LHS) Cluster assignment for Georgia counties using Dahl's method from the proposed algorithm. (RHS) The cluster assignment obtained from the proposed algorithm using the mode method.

(see "Methods" section below) exhibit a high degree of similarity. The corresponding parameter estimates are shown in Table 2.

Since the kernel type, bandwidth prior and the number of knots play a crucial role in the spatial stick-breaking construction of the the proposed mode for the DPMM, the priors and kernel functions from Table 8 ("Methods" section) were utilised to test the accuracy of the proposed model. We explored different options to determine the optimal fit for the data using the Watanabe-Akaike onformation criterion (WAIC)[39], and we performed a sensitivity analysis on the proposed model with respect to number of knots. The results are summarized in Table 3.

According to the Table 3, the optimal number of knots for the simulated data as 9, as evidenced by the lowest WAIC value. In Table 4, we present a sensitivity analysis for our proposed algorithm, discussing its performance under various bandwidth priors and kernel functions. This table categorizes the results under two primary kernel types: uniform and squared exponential.

For each kernel type, two bandwidth priors were evaluated. Based on the WAIC values, the squared exponential kernel with a bandwidth prior of $\varepsilon_{1i}, \varepsilon_{2i} \equiv \frac{\lambda^2}{2}$ emerged as the most effective. Importantly, our algorithm not only demonstrated the stability of clusters when compared to two established methods but also managed to accurately assign clusters with an accuracy of 0.87. In the simulated dataset, our method effectively identified three distinct clusters.

Cluster	1	2	3
$\hat{\beta}_1$	– 1.0	0.62	2.31
	(– 1.01, – 0.92)	(– 0.03, 0.68)	(2.06, 2.32)
$\hat{\beta}_2$	– 0.04	– 0.29	0.29
	(– 0.06, – 0.04)	(– 0.30, – 0.19)	(0.19, 0.30)
$\hat{\beta}_3$	– 0.46	– 1.06	– 1.14
	(– 0.49, – 0.46)	(– 1.07, – 0.82)	(– 1.14, – 1.13)
$\hat{\beta}_4$	1.29	– 0.06	0.53
	(1.22, 1.29)	(– 0.09, 0.47)	(0.44, 0.54)
$\hat{\beta}_5$	2.46	3.71	2.19
	(2.46, 2.53)	(3.24, 3.74)	(2.18, 2.42)
$\hat{\beta}_6$	– 1.05	– 0.51	0.37
	(– 1.05, – 1.02)	(– 0.73, – 0.49)	(0.24, 0.37)

Table 2. Parameter estimates and their 95% highest posterior density (HPD) intervals for the three clusters identified.

Number of knots	WAIC
9	117044.3
14	117066.4
19	117051.6
29	117055.4

Table 3. Sensitivity analysis for the number of knots in the spatial stick-breaking with the squared exponential kernel.

	Bandwidth prior	RI	Summary	WAIC

Uniform kernel	$\varepsilon_{1i}, \varepsilon_{2i} \equiv \lambda$	0.86	C1=56,
			C2=70,
			C3=33
				117067.4

	$\varepsilon_{1i}, \varepsilon_{2i} \equiv Exp(\lambda)$	0.87	C1=51,
			C2=42,
			C3=10
				117058.4

Exp kernel	$\varepsilon_{1i}, \varepsilon_{2i} \equiv \frac{\lambda^2}{2}$	0.88	C1=54,
			C2=66,
			C3=39
				117044.3

	$\varepsilon_{1i}, \varepsilon_{2i} \sim$ Inverse Gamma $(1.5, \lambda^2/2)$	0.84	C1=21,
			C2=55,
			C3=39,
			C4=44
				126134.4

Table 4. Sensitivity analysis for the proposed algorithm with different bandwidth priors and kernel functions with the number of knots is 9.

Real data analysis

Case study

Children's Health Queensland (CHQ) has developed the CHQ Population Health Dashboard, a remarkable resource providing data on health outcomes and socio-demographic factors for a one-year period (2018-2019) across 528 small areas (Statistical area level 2 (SA2) in Queensland, Australia. The dashboard presents over 40 variables in a user-friendly format, with a focus on health outcomes, particularly vulnerability indicators measuring children's developmental vulnerability across five Australian Early Development Census (AEDC) domains. These domains include physical health, social competence, emotional maturity, language and cognitive skills, and communication skills with general knowledge.

The AEDC also includes two additional domain indicators: vulnerable on one or more domains (Vuln 1) and vulnerable on two or more domains (Vuln 2). Socio-demographic factors, including Socio-Economic Indexes for Areas (SEIFA) score, preschool attendance, and remoteness factors are also incorporated, offering insights into potential links to health outcomes. The SEIFA score summarizes socio-economic conditions in an area, while remoteness factors categorize regions into cities, regional, and remote areas.

In the field of population health, publicly available data are often grouped according to geographical regions, such as the statistical areas (SA) defined in the Australian Statistical Geography Standard (ASGS). These areas, called SA1, SA2, SA3, and SA4, range respectively from the smallest to the largest defined geographical regions. Due to privacy and confidentiality concerns, personal-level information is typically not released. Therefore, in this paper, we focus on group-level data, and in the case study, we use data that have been aggregated at the SA2 level[41].

Data for the analysis are sourced from the 2018 AEDC, and focus on the proportion of vulnerable children in each SA2. Some missing data is handled through imputation using neighboring SA2s, with two islands having no contiguous neighbors excluded from the analysis. The study utilises the remaining data from 526 SA2 areas to conduct the analysis.

Our study uses the proposed methodology to analyse the influential factors affecting the development of children who are vulnerable in one or more domains (Vuln 1) in the Queensland SA2 regions. Data were found on the Australian Bureau of Statistics (ABS) official website and the AEDC. For each SA2 region, we considered several dependent variables, including the proportion of attendance at preschool, the remoteness factor which is converted using the one hot coding to three variables including zero and one and the index of relative socio-economic disadvantage (IRSD) factor which is considered continuous in this case study. Before fitting the model, we scaled the variables using the logarithm. As a result, all the models are fitted without an intercept term. In our Bayesian Geographically Weighted Regression (GWR) analysis, we illustrate substantively important regions by plotting the 95% credible intervals for each coefficient. With (BGWR), a separate parameter estimate is indeed generated for each region. This means that the credible intervals obtained are specific to each region and its corresponding parameter. Therefore, the 95% credible interval associated with each region reflects the uncertainty in the parameter estimate for that particular geographical area. This is illustrated in the Fig. 4, where the blue line represents the mean.

Spatial cluster inferences

Figure 5 offers a geographical representation of five posterior mean parameters plotted on a map of Queensland. These values have been obtained through the proposed method, revealing that the relationship between the response variable (Vuln 1) and the covariates varies across different locations.

To find the suitable structure of spatial weighs kernel for SDPMM for the case study, we used the WAIC. The WAIC values of the uniform and exponentially weighted kernels associated with different bandwidth priors can be found in Table 5. Comparison of the WAIC value leads to the conclusion that the exponential kernel is the most suitable for the real dataset. Additionally, Table 6 shows that as the complexity of the model (in terms of the number of knots) increases, the fit of the model to the data (WAIC) also improves. However, since the total number of SA2 is just 526, in this case study we assumed the number of knots to be 9.

Figure 6 showed the cluster distribution on the map obtained from the proposed algorithm with 6 clusters using Dahl's method. The cluster sizes are 124, 103, 90, 101 , 101 and 7. The strength of the proposed algorithm lies in its capability to create smaller cluster sizes compared to other cluster algorithms. This is beneficial for policy interventions targeting specific regions in Queensland, especially for identifying regions with high developmental vulnerabilities. Further, we provide a summary (Table 7) for each of these clusters according to the parameter estimation and 95% highest posterior density (HPD) interval.

Cluster 1 (124 out of 526) stands out due to its negative effect on the regression parameters for "Attendance at Preschool" with a narrow credible interval. The positive effects for the three levels of "Remoteness" are reliable, with the broadest uncertainty observed for the "Cities" parameters in comparison with the rest of the clusters. There's also some uncertainty in the "IRSD" parameters, which exhibit a negative effect, although they remain influential. Cluster 2 (103 out of 526) is characterized by a significant negative effect for "Attendance at Preschool" with a narrowest credible interval across the six clusters, indicating a strong impact and high confidence. Additionally, There are more positive effects for the "Cities", "Regional" and "Remote" parameters compared to Cluster 1, there is a more negative relationship for "IRSD" parameters compared to Cluster 1. Cluster 3 (90 out of 526) also exhibits a significant negative effect for "Attendance at Preschool" parameters but with a broader credible interval, indicating a strong impact with more uncertainty. The positive effects for "Remoteness" parameters are still significant and confident, with the broadest uncertainty for the "Regional" parameters across the six clusters, "IRSD" exhibits the most negative parameters in this cluster in comparison with the rest. Cluster 4 (101 out of 526) maintains a significant negative effect for "Attendance at Preschool" with a narrow credible interval, with a positive effects for "Remoteness" parameters. In this cluster "IRSD" has a positive effect with a narrow credible

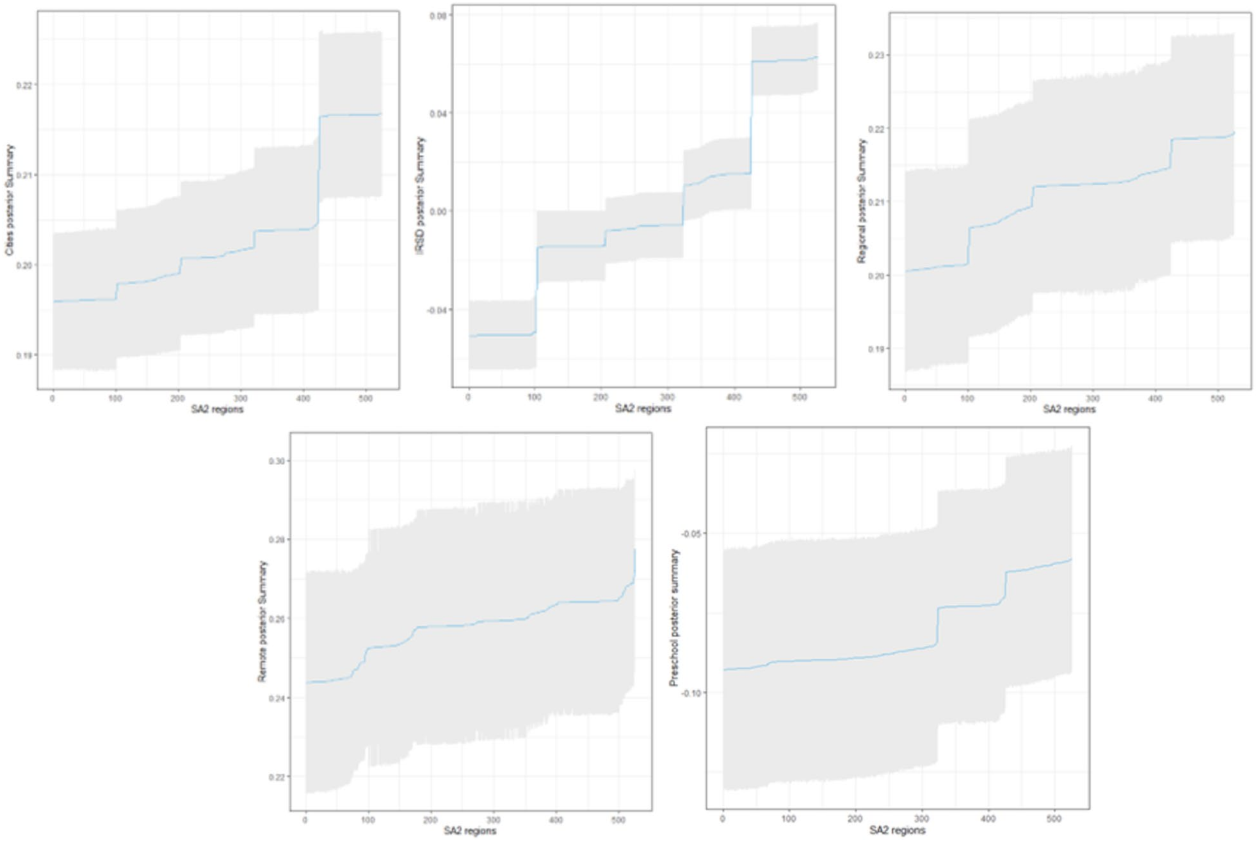

Figure 4. 95% posterior credible interval form the proposed algorithm.

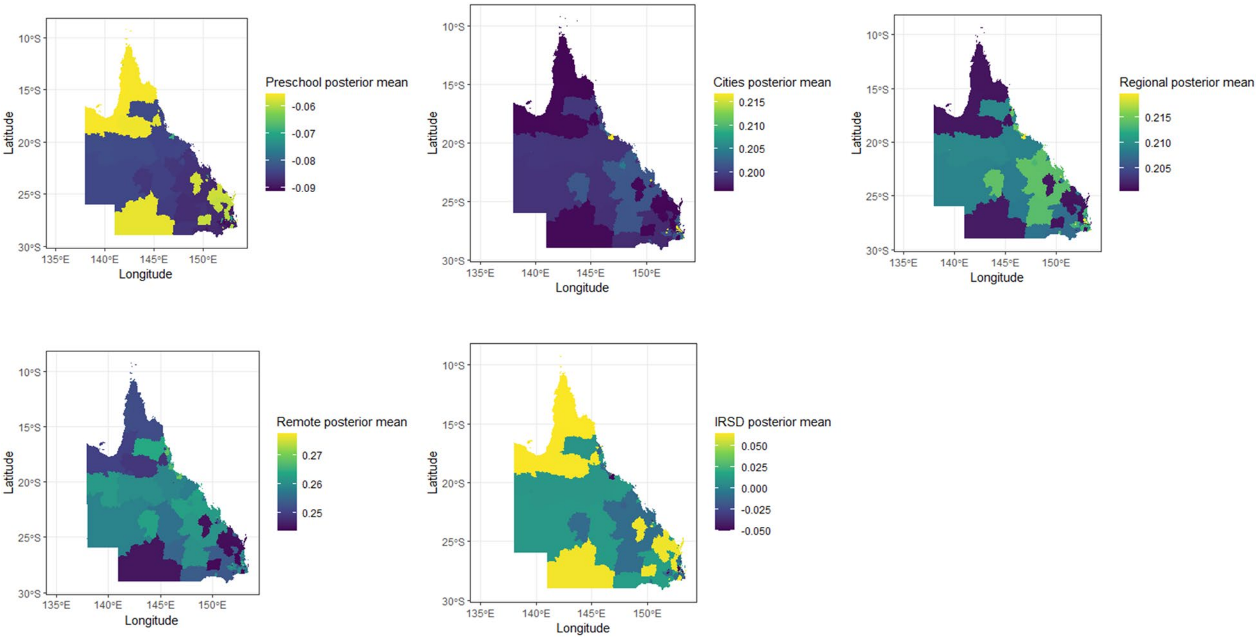

Figure 5. The spatial distribution of the posterior mean parameters derived from the proposed model.

interval in comparison with clusters 1, 2 and 3. Cluster 5 (101 out of 526) has a negative effect for "Attendance at Preschool" and a narrow credible interval. Similar to Cluster 4, the positive effects for "Remoteness". But in this cluster "IRSD" has more positive effect in comparison with cluster 4. Cluster 6 (7 out of 526) stands out with it negative effect for "Attendance at Preschool" even though it has a wider credible interval. The positive effects for "Remoteness" are similar to the previous clusters, with the narrowest credible intervals for the "Cities, Regional, and Remote" parameters, also the "IRSD" has a negative effect with a narrow credible interval.

Kernel type	Bandwidth prior	WAIC
Uniform	$\varepsilon_{1i}, \varepsilon_{2i} \equiv \lambda$	3564.28
	$\varepsilon_{1i}, \varepsilon_{2i} \sim$ Inverse Gamma$(1.5, \lambda^2/2)$	3568.94
Exponential	$\varepsilon_{1i}, \varepsilon_{2i} \equiv \frac{\lambda^2}{2}$	3550.91
	$\varepsilon_{1i}, \varepsilon_{2i} \sim$ Inverse Gamma$(1.5, \lambda^2/2)$	3565.42

Table 5. Sensitivity analysis for the proposed algorithm with different bandwidth priors and kernel functions with 9 knots.

Number of knots	WAIC
9	3550.91
19	2705.90
24	1679.01
32	1279.32

Table 6. Sensitivity analysis for the number of knots in the spatial stick-breaking.

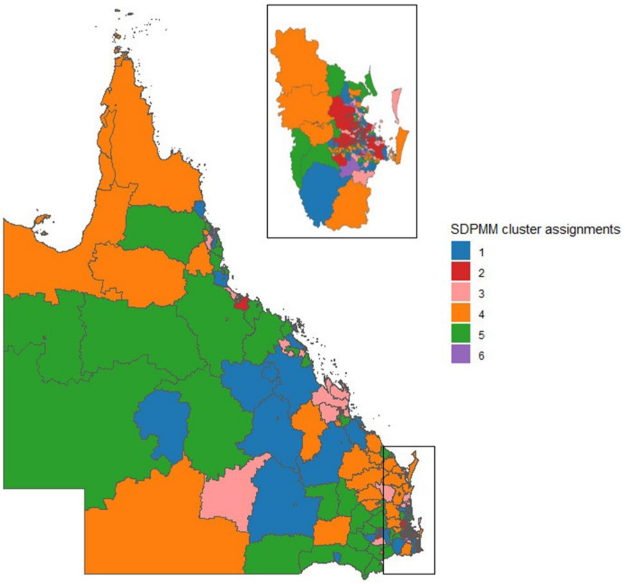

Figure 6. Cluster distribution from the proposed algorithm for the case study.

Cluster	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
1	– 0.083 (– 0.126, – 0.003)	0.206 (0.199, 0.216)	0.214 (0.197, 0.235)	0.259 (0.221, 0.293)	– 0.014 (– 0.029, – 0.001)
2	– 0.075 (– 0.112, – 0.035)	0.216 (0.205, 0.230)	0.217 (0.200, 0.232)	0.262 (0.231, 0.294)	– 0.049 (– 0.070, – 0.031)
3	– 0.095 (– 0.138, – 0.046)	0.198 (0.189, 0.207)	0.213 (0.199, 0.231)	0.259 (0.221, 0.299)	– 0.005 (– 0.018, 0.009)
4	– 0.056 (– 0.112, – 0.004)	0.196 (0.186, 0.205)	0.202 (0.178, 0.221)	0.244 (0.221, 0.278)	0.063 (0.044, 0.086)
5	– 0.087 (– 0.134, – 0.040)	0.199 (0.189, 0.207)	0.207 (0.187, 0.224)	0.255 (0.224, 0.296)	0.013 (– 0.003, 0.036)
6	– 0.125 (– 0.173, – 0.102)	0.200 (0.194, 0.208)	0.192 (0.177, 0.199)	0.256 (0.235, 0.278)	– 0.017 (– 0.028, – 0.010)

Table 7. Parameter estimates and their 95% highest posterior density (HPD) intervals for the six clusters identified. $\hat{\beta}_1$: attendance at preschool parameters, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$: remoteness parameters, with three levels (cities, regional, and remote), and $\hat{\beta}_5$: IRSD parameters.

These clusters are distinguished primarily by the magnitude and certainty of the effect of "Attendance at Preschool" and the reliability of the "Remoteness" and "IRSD". Cluster 1, 4,5, and 6 share a strong negative impact of "Attendance at Preschool" with narrow credible intervals. Cluster 2 , with a broader credible interval, indicates more uncertainty in the impact of "Attendance at Preschool". Cluster 3, despite having a wider credible interval for "Attendance at Preschool" still shows a significant negative effect. Additionally, the "IRSD" exhibits variations across clusters, adding another layer of distinction.

Discussion

In this paper, we introduce a new statistical framework aimed at addressing the challenges in clustering posed by spatially varying relationships within regression analysis. Specifically, we present a Bayesian model that integrates geographically weighted regression with a spatial Dirichlet process to cluster relevant model parameters. This solution therefore not only identifies clusters of the model parameters but effectively captures the inherent heterogeneity present in spatial data. Our exploration encompasses various weighting schemes designed to effectively model the complex spatial interaction between neighborhood characteristics and the positioning of key points (or "knots"). This modelling is supported by a discussion of Bayesian model selection criteria, a crucial step in the analysis process that ensures selection of an appropriate and well-fitted model. Spatial variation in the effects of covariates empowers our model to provide a better fit to spatial data compared to conventional models, offering insights into the complex patterns of heterogeneity across diverse geographical locations. Additionally, making smaller group sizes helps decision-makers identify which regions need more help. To demonstrate the efficacy of our methodology, we have presented a simulation study. Moreover, we have extended our investigation to a real-world application: a thorough analysis of the factors influencing children's development indicators in Queensland. Through this practical example, we showcase the benefits of our proposed approach, emphasizing its ability to find hidden dynamics that might otherwise remain obscured.

In our case study, we aimed to explore the influential factors affecting child development vulnerability in Queensland's statistical area level 2 (SA2) regions. Our analysis utilised a dataset consisting of 526 observations, each corresponding to one of Queensland's SA2 regions. The dataset included various explanatory variables, including preschool attendance, remoteness factors, and socio-economic factors. The primary objective was to identify spatial clusters of children's vulnerability and gain insights into the regional disparities in children's development domains. To select the appropriate spatial weights kernel for our model, we employed WAIC and found the uniform kernel, suggesting that it provides a better fit for our real dataset. Furthermore, we performed a sensitivity analysis to determine the optimal number of knots in the spatial stick-breaking process and found increasing the complexity of the model by adding more knots improved its fit to the data. We selected 9 knots as the optimal number for our analysis. Using the selected model with an exponential kernel and 10 knots, we applied the proposed algorithm to identify spatial clusters of child vulnerability. Our analysis revealed a total of 6 clusters across Queensland's SA2 regions. These clusters vary in size, with the largest containing 124 regions and the smallest comprising only 7 regions. The ability of the proposed algorithm to create smaller cluster sizes is noteworthy, as it allows for more targeted policy interventions in regions with specific developmental needs. Moreover, we conducted a detailed analysis of the clusters to understand their characteristics and implications. For instance, the presence of smaller clusters may indicate isolated areas with unique developmental challenges that require tailored interventions. In contrast, larger clusters could represent regions with similar vulnerabilities, suggesting the need for broader policy strategies. These findings offer valuable insights for policymakers and stakeholders interested in addressing child development disparities in Queensland.

Both the Bayesian Geographically Weighted Regression (GWR) and the Bayesian Spatially Varying Coefficient Model (Bayesian SVCM) offer powerful tools for understanding spatially varying relationships within data. Comparing and contrasting these two approaches can help in justifying the consideration of Bayesian GWR.

Firstly, both Bayesian GWR and Bayesian SVCM operate within a Bayesian framework, allowing for the incorporation of prior knowledge and uncertainty into the modelling process. However, they differ in their approaches to capturing spatial variation. Bayesian GWR explicitly models spatial heterogeneity by allowing regression coefficients to vary across space, making it well-suited for exploring localized relationships between variables. On the other hand, Bayesian SVCM focuses on estimating spatially varying coefficients for a global regression model, which may overlook finer-scale variations present in the data.

Furthermore, it is important to note that coefficients obtained from these methods may differ. Bayesian GWR produces coefficients that are specific to each geographic location, reflecting the spatially varying nature of relationships within the data. In contrast, Bayesian SVCM estimates coefficients that represent spatially varying effects within the context of a global model. These differences in coefficient estimation highlight the distinct strengths and interpretation nuances of each approach, the Bayesian GWR approach can complement existing non-Bayesian techniques such as the Spatial Clustered Regression (SCR) proposed by Sugasawa and Murakami[37] While SCR provides an alternative for capturing spatial clustering effects, it may lack the flexibility to adequately model spatially varying relationships. Bayesian GWR, with its emphasis on local estimation, can offer additional insights into how relationships between variables change across different geographic areas .

Bayesian SVCM and our method have their computational challenges. Our proposed algorithm is not computationally intensive, in comparison with other clustering Bayesian methods. The time to run the simulated data was around 25 minutes, while for the real data set it took around 2:23 hours using the high performance computer (HPC).

While our paper represents a step forward in the field of spatial regression, it is essential to acknowledge the avenues for further exploration that our research did not study. For instance, while we thoroughly examined the full model that incorporates all relevant covariates, we did not delve into methodologies for variable selection within the context of clustered regression. This presents a clear direction for future research, where approaches for selecting the most influential variables among a clustered regression could enhance the performance of our model even further. Additionally, our work touched upon the utilisation of the spatial Dirichlet process mixture model (SDPMM) to derive cluster information for regression coefficients. However, we acknowledge that the posterior distribution of the cluster count might not always be accurately estimated through the SDPMM, as demonstrated by Miller and Harrison[42]. Our simulation studies confirm this observation. This area emerges as a critical focus for future studies. Another area for future work involves expanding our methodology to accommodate non-Gaussian data distributions, a direction that holds promise for a wider range of applications. Moreover, the pursuit of adapting our model to handle multivariate response scenarios represents an essential avenue for future exploration, offering the potential to unlock insights and applications across various domains. Lastly, extend the proposed algorithm to a semi-parametric GWR scenario where certain exploratory variables remain fixed while others vary spatially[43]. Further, GWR allows regression coefficients to vary by location; it typically assumes a linear relationship between all predictor variables and the response within each location. However, in some real-world scenarios, not all predictor variables may exhibit linear relationships with the response variable. Some variables might have non-linear patterns or lack a certain discernible pattern altogether. These features could be included in the linear model through polynomials, splines, interactions, and so on, and alternative non-parametric regression models could be developed. It would be interesting to extend the proposed algorithm to allow for more flexibility in modeling complex relationships between predictor variables and the response[44].

Methods

This section outlines the proposed model called the spatial Dirichlet process clustered heterogeneous regression model. The model utilises a non-parametric spatial Dirichlet mixture model applied to the regression coefficients of the geographically weighted regression model. The model is cast in a Bayesian framework.

Bayesian geographical weighted regression

The Bayesian geographically weighted regression (BGWR) model can be described as follows. Given diagonal weight matrix $W(s)$ for a location s, the likelihood for each $y(s)$ is:

$$y(s)|\beta(s), x(s), W_i(s), \sigma^2(s) \sim \mathcal{N}(x^T(s)\beta(s), \sigma^2(s)W_i^{-1}(s)) \quad (5)$$

where $y(s)$ is the ith observation of the dependent variable, $x(s)$ is the ith row (or observation) from the design matrix X and $W_i(s)$ is the ith diagonal element from the spatial weight matrix $W(s)$. The weighted matrix $W(s)$ is constructed to identify the relative influence of neighbouring regions on the parameter estimates at locations.

When working with areal data, the graph distance is an alternative distance metric that can be used. It is based on the concept of a graph, where $V = \{v_1, ..., v_m\}$ represents the set of nodes (locations) and $E(G) = \{e_1, ..., e_n\}$ represents the set of edges connecting these nodes. The graph distance is defined as the distance between any two nodes in the graph[35].

$$W(s) = \begin{cases} |V(e)| & \text{if } e \text{ is the shortest distance connecting a pair of nodes,} \\ \infty & \text{if the two nodes are not connected} \end{cases}$$

where $|V(e)|$ is the number of edges in e[45]. The graph distance-based weighted function is given as:

$$W(s) = \begin{cases} 1 & \text{if } d_i(s) \leq b, \\ f(d_i(s)^b) & \text{otherwise} \end{cases}$$

where $d_i(s)^b$ is the graph distance between locations i and s, f is a weighting function, and b represents the bandwidth. In this study, we suppose that $f()$ is a negative exponential function[29], so,

$$W(s) = \begin{cases} 1 & \text{if } d_i(s) \leq 1, \\ e^{(-d_i(s)/b)} & \text{otherwise} \end{cases}$$

where b represents the bandwidth that controls the decay with respect to distance[46]. Here, $d_i(s)$ indicates that an observation far away from the location of interest contributes little to the estimate of parameters at that location. In this paper, we used the graph distance and the greater circle distances[47] and both of these methods show consistent parameters. The proposed model is constructed in a Bayesian framework with conjugate priors on the regression parameters and other model terms. The full model is given in the (Full Bayesian spatial Dirichlet process mixture prior cluster heterogeneous regression) section.

Heterogeneous regression with spatial Dirichlet process mixture prior

In a Bayesian framework, coefficient clustering can be achieved by using a Dirichlet process mixture model (DPMM). This approach links the response variable to the covariates through cluster membership. The DPMM is defined by a probability measure G that follows a Dirichlet process, denoted as $G \sim (\alpha, G_0)$, where α is the concentration parameter and G_0 is the base distribution[35]. Hence,

$$(G(A_1), ..., G(A_k)) \sim Dirichlet(\alpha G_0(A_1), ..., \alpha G_0(A_k)) \quad (6)$$

where $(A_1, ..., A_k)$ is a finite measurable partition of the space Ω, and the variable k represents the number of components or clusters in a (DPMM). Several formulas have been proposed in the literature for specifying the DPMM's parameters and incorporating spatial dependencies[48,49]. A popular approach is the spatial stick-breaking algorithm[50,51], which in a BGWR setup is applied at each location as follows:

Name	$\boldsymbol{l_i(s)}$	**Model for $\boldsymbol{\varepsilon_{1i}}$ and $\boldsymbol{\varepsilon_{2i}}$**
Uniform	$\prod_{j=1}^{2} I\left(\lvert s_j - \psi_{ji}\rvert < \frac{\varepsilon_{ji}}{2}\right)$	$\varepsilon_{1i}, \varepsilon_{2i} \equiv \lambda$
Uniform	$\prod_{j=1}^{2} I\left(\lvert s_j - \psi_{ji}\rvert < \frac{\varepsilon_{ji}}{2}\right)$	$\varepsilon_{1i}, \varepsilon_{2i} \sim \mathrm{Exp}(\lambda)$
Squared exp.	$\prod_{j=1}^{2} \exp\left(-\frac{(s_j-\psi_{ji})^2}{2\varepsilon_{ji}^2}\right)$	$\varepsilon_{1i}, \varepsilon_{2i} \equiv \frac{\lambda^2}{2}$
Squared exp.	$\prod_{j=1}^{2} \exp\left(-\frac{(s_j-\psi_{ji})^2}{2\varepsilon_{ji}^2}\right)$	$\varepsilon_{1i}, \varepsilon_{2i} \sim \mathrm{IG}(1.5, \lambda^2/2)$

Table 8. Examples of kernel functions, where *IG* presents inverse Gamma function.

$$
\begin{aligned}
G(s) &= \sum_{i=1}^{K} p_i(s)\delta(\theta_i) \\
p_1(s) &= V_1(s) \\
p_i(s) &= V_i(s)\prod_{j=1}^{i-1}(1 - V_j(s)) \quad \text{for all } i > 1 \\
V_i(s) &= l_i(s)V_i \\
V_i &\sim \mathrm{Beta}(a, b)
\end{aligned}
\tag{7}
$$

where $\delta(\theta_i)$ is the Dirac distribution with a point mass at (θ_i), and $p_i(s)$ is a random probability weight between 0 and 1. The distribution of $G(s)$ depends on V_i and θ_i; the distribution varies according to the kernel function $l_i(s)$. However, the spatial distributions of the kernel function $l_i(s)$ vary, constrained within the interval [0, 1]. These functions are centered at knots $\psi_i = (\psi_{1i}, \psi_{2i})$, and the degree of spread is determined by the bandwidth parameter $\epsilon_i = (\epsilon_{1i}, \epsilon_{2i})$. Both the knots and bandwidths are treated as unknown parameters with independent prior distributions, unrelated to V_i and θ_i. The knots ψ_i are assigned independent uniform priors, covering the bounded spatial domain. The bandwidths can be modelled to be uniform for all kernel functions or can vary across kernel functions, following specified prior distributions[50,52]. The most common kernels are the uniform and the square exponential functions. This kernel can take different formats. Table 8 provides examples of the most popular kernels used for the spatial stick-breaking configuration.

A vector of latent allocation variables Z is generated to characterize the clustering explicitly. Let $Z_{n,k} = \{z_1., ..., z_n\}$, where $z_i \in \{1, ..., k\}$ and $1 \leq i \leq n$ represents all possible clustering of n observations into K clusters.

Full Bayesian spatial dirichlet process mixture prior cluster heterogeneous regression

Adapting the spatial Dirichlet process to the heterogeneous regression model, we focus on clustering of spatial coefficients $\beta(s_1), ..., \beta(s_n)$ and $\beta(s_i) = \beta_{z_i} \in \{\beta_1, ..., \beta_k\}$. The full model is described as follows with the most commonly adopted priors:

$$y(s)|\beta(s), x(s), W_i(s), \sigma^2(s) \sim \mathcal{N}(x^T(s)\beta(s_{z_i}), \sigma^2(s)W_i^{-1}(s)) \tag{8}$$

$$W_i(s) = f(d_i|b) \tag{9}$$

$$b \sim Uniform(0, D) \tag{10}$$

$$\beta_{z_i} \sim \mathcal{N}_p(\mu_{z_i}, \Sigma_{z_i}) \tag{11}$$

$$z_i \sim categorical(p_1(s), p_2(s), ..., p_k(s)) \tag{12}$$

$$\mu_k|\Sigma_k \sim \mathcal{N}_p(m_k, \Sigma_k) \tag{13}$$

$$\Sigma_k \sim IW(D_k, c_k) \tag{14}$$

$$\sigma^2(s) \sim IGamma(\alpha_1, \alpha_2) \tag{15}$$

$$P(z_i = k|p) = p_k(s) \tag{16}$$

$$p_1(s) = V_1(s), p_k(s) = V_k(s) \prod_{j=1}^{K-1} (1 - V_j(s)), V_k(s) = l_k(s) V_k \tag{17}$$

$$V_k \sim Beta(a_v, b_v) \tag{18}$$

Here, the response variable Y is assumed to follow a Gaussian distribution; the design matrix representing the predictors is denoted by X, and the spatial weight matrix $W(s)$ depends on two key aspects: the distance between observations, represented as d_i, and a parameter b, which controls the bandwidth. This bandwidth is assumed to follow a uniform distribution between 0 and a certain value D , which represents the bandwidth parameter. A common prior for the bandwidth is given by: b-Uniform(0,D), where D > 0. Without any prior knowledge, D can be selected to be large enough that we begin to approximate with a non-informative prior; i.e. we begin with an approximate global model in which all observations are weighted equally. We used a bandwidth parameter D set to 100. The maximum great circle distance in the spatial structure of the 159 regions is 10, so using a bandwidth of 100 induces a weighting scheme that ensures relative weights are assigned appropriately. If the distance between two regions is considerable, the relative weight is approximately exp(-10/100) = 0.904. This approximation thus allows the model to behave similarly to a global model where every observation is equally weighted, ensuring a sufficiently non-informative prior bandwidth b. please see the method section. f is the graph weighting function. The regression coefficients β_{z_i} are associated with a specific group, or cluster, z_i, for a particular observation. The mean and spread of cluster z_i are denoted as μ and Σ_{z_i} , respectively, and the maximum number of possible clusters is K.

The hyper-parameter m_k is a prior mean value for the μ_{z_i} and Σ_k is a way to express how different a cluster can be. Similarly, D_k is the scale matrix, and $c_k > p - 1$ is the degrees of freedom. Another important aspect is the variation in the data, which is $\sigma^2(s)$. This variation changes across locations and follows a specific prior pattern, which is assumed to be an inverse Gamma distribution with parameters α_1 and α_2.

we focus on the probability $P(z_i = k)$ that observation i belongs to cluster k. This assignment probability at a specific location s is denoted as $p_k(s)$. For the clusters, we also consider "stick-breaking weights", denoted by $V_k(s)$, which change across locations. The values a_v and b_v are related to how these weights are determined using a beta distribution. Here $\sum_{j=1}^{k} p_k(s) = 1$ almost surely under the constraint that $V_k(s) = 1$ for all locations s[53].

Bayesian estimation and inference

This section covers using MCMC to obtain samples from posterior distributions of model parameters. It explains the sampling scheme, covers the use of posterior inference for cluster assignments, and methods for evaluating accuracy.

The MCMC sampling schemes

The main R function for the model is implemented using the nimble package[54]. This function encapsulates the model and provides an interface for executing the MCMC sampling scheme, performing posterior inference, and evaluating estimation performance and clustering accuracy. The model itself is wrapped within a nimbleCode function, which allows the nimble package to generate and compile C++ code to execute the MCMC sampling scheme efficiently. This can result in substantial speed improvements over pure R implementations, especially for models with large datasets or complex parameter space. In the context of the proposed algorithm, the nimble package provides several MCMC sampling methods, including the popular Gibbs and Metropolis-Hasting algorithms for inferring the posterior distribution of the regression and other model parameters. Nimble also allows for the specification of priors and likelihood functions for the parameters to customise the MCMC sampling process. In our study, the Gibbs sampling algorithm was used to obtain the clusters of the parameters.

Block Gibbs sampling is a MCMC technique used for sampling from the joint distribution of multiple random variables. The primary idea behind block sampling is to group related variables together into "blocks" and sample them jointly, which can improve the efficiency and convergence of the sampling process[55]. An explanation of this sampling algorithm for the proposed algorithm can be found in the Appendix.

Cluster configurations

Two methods are used to determine cluster configurations. In the first, the estimated parameters, together with the cluster assignments $Z_{n,k}$ are determined for each replicate from the best post-burn-in iteration selected using Dahl's method[56], which involves estimating the clustering of observations through a least-squares model-based approach that draws from the posterior distribution. In this method, membership matrices for each iteration, denoted as $B^{(1)}, \ldots, B^{(M)}$, with M being the number of post-burn-in MCMC iterations, are computed. The membership matrix for the c-th iteration, $B^{(c)}$ is defined as:

$$B^{(c)} = (B^{(c)}(i,j))_{i,j \in \{1:n\}} = \mathbf{1}(z_i^{(c)} = z_j^{(c)})_{n \times n} \tag{19}$$

where $\mathbf{1}(\cdot)$ represents the indicator function. The entries $B^{(c)}(i,j)$ take values in $\{0, 1\}$ for all $i, j = 1, \ldots, n$ and $c = 1, \ldots, M$. When $B^{(c)}(i,j) = 1$, it indicates that observations i, and j belong to the same cluster in the cth iteration.

To obtain an empirical estimate of the probability for locations i and j being in the same cluster, the average of $B^{(1)}, \ldots, B^{(M)}$ can be calculated as:

$$\overline{B} = \frac{1}{M}\sum_{c=1}^{M} B^{(c)} \tag{20}$$

where $\sum$ denotes the element-wise summation of matrices. The (i, j)th entry of $\overline{B}$ provides this empirical estimate.

Subsequently, the iteration that exhibits the least squared distance to $\overline{B}$ is determined as:

$$C_{LS} = \arg\min_{c\in\{1:M\}} \left[\sum_{i=1}^{n}\sum_{j=1}^{n}(B^{(c)}(i,j) - \overline{B}(i,j))^2\right] \tag{21}$$

where $B^{(c)}(i,j)$ represents the (i, j)th entry of $B(c)$, and $\overline{B}(i,j)$ denotes the (i, j)th entry of $\overline{B}$. The least-squares clustering offers an advantage in that it leverages information from all clusterings through the empirical pairwise probability matrix $\overline{B}$.

The second method utilised here is the posterior mode method. This method leverages posterior samples from iterations associated with z_i, where z denotes the cluster assignments specific to each region. Each iteration generates a new set of cluster assignments z, which are dependent on the parameters. Consequently, following multiple iterations, each region will have an empirical posterior distribution of cluster assignments z. The mode indicates the cluster with the highest probability of assignment for a given region.

Cluster accuracy

In order to assess the accuracy of the proposed algorithm, we compared the cluster configurations with the true labels provided for the simulated data. It is important to note that while the true labels are available for the simulated data, such information is not readily available for real-world datasets. In practice, true labels are often unknown, which poses a challenge for the evaluation of clustering accuracy. In this study, we utilised the Rand index (RI)[57]. This index measures the level of similarities between two sets of cluster assignments, labelled as C and C', with respect to a given dataset $X = \{x_1, x_2, \ldots, x_n\}$. Each data point $x(s)$ is assigned a cluster label c_i in C and c'_i in C'. The RI is computed using the following formula:

$$\mathrm{RI} = \frac{a+b}{a+b+c+d}. \tag{22}$$

Hhere a, represents the number of pairs of data points that are in the same cluster in both C and C' (true positives); b indicates the number of pairs of data points that are in different clusters in both C and C' (true negatives); c represents the number of pairs of data points that are in the same cluster in C but in different clusters in C' (false positives); and d stands for the number of pairs of data points that are in different clusters C but in the same cluster in C' (false negatives).

The Rand index ranges from 0 to 1, with a value of 1 denoting a complete concordance between the two clusterings (both C and C' perfectly agree on all pairs of data points). Conversely, a value close to 0 indicates a weak level of agreement between the two clusterings.

Conclusion

This paper introduces a method called the spatial Dirichlet process clustered heterogeneous regression model. The method employs a non-parametric Bayesian clustering approach to group the spatially varying regression parameters of a Bayesian geographically weighted regression, and also determines the best number and arrangement of clusters. The model uses advanced Bayesian techniques to cluster the parameters and determine the best number and arrangement of clusters. The model's abilities were demonstrated using simulated data and then applied to actual data related to children's development vulnerabilities in their first year of school. In this application, the model successfully identified key factors. This approach enhances our understanding of how children develop in various regions, revealing the factors that impact their health and well-being. With these insights, policymakers can create targeted policies that are suited to each area's unique characteristics. As a result, this innovative method not only improves the suite of analytical tools but also contributes to the broader goal of enhancing the health and development prospects of children in different places.

Data availibility

All the datasets used in this article are publicly accessible and free to download. Anyone interested can access them without special privileges. Likewise, the authors did not have any special privileges when accessing the data for analysis in this article. The datasets can be obtained from the following sources: Children's Health Data is sourced from the Australian Early Development Census, available upon request at https://www.aedc.gov.au/data-explorer/. The Explanatory Data is obtained from the Australian Bureau of Statistics and is publicly available at https://www.abs.gov.au/census/find-census-data/quickstats/2021/3.

Received: 21 November 2023; Accepted: 17 April 2024
Published online: 27 April 2024

References

1. Lawson, A. B., Banerjee, S., Haining, R. P. & Ugarte, M. D. *Handbook of Spatial Epidemiology* (CRC Press, 2016).
2. Anselin, L. Spatial dependence and spatial structural instability in applied regression analysis. *J. Reg. Sci.* **30**, 185–207 (1990).
3. Hanson, T., Banerjee, S., Li, P. & McBean, A. Spatial boundary detection for areal counts. *Nonparametric Bayesian Inference Biostat.*https://doi.org/10.1007/978-3-319-19518-6_19 *(2015).*

4. Ma, H., Carlin, B. P. & Banerjee, S. Hierarchical and joint site-edge methods for medicare hospice service region boundary analysis. *Biometrics* **66**, 355–364 (2010).
5. Lee, D. & Mitchell, R. Boundary detection in disease mapping studies. *Biostatistics* **13**, 415–426 (2012).
6. Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**, 479–498 (2002).
7. Aiello, L. & Banerjee, S. Detecting spatial health disparities using disease maps. Preprint at http://arxiv.org/abs/2309.02086 (2023).
8. Riley, D. D., Koutsoukos, X. & Riley, K. Simulation of stochastic hybrid systems using probabilistic boundary detection and adaptive time stepping. *Simul. Model. Pract. Theory* **18**, 1397–1411 (2010).
9. Gao, H. & Bradley, J. R. Bayesian analysis of areal data with unknown adjacencies using the stochastic edge mixed effects model. *Spat. Stat.* **31**, 100357 (2019).
10. Lu, H., Reilly, C. S., Banerjee, S. & Carlin, B. P. Bayesian areal wombling via adjacency modeling. *Environ. Ecol. Stat.* **14**, 433–452 (2007).
11. Lu, H. & Carlin, B. P. Bayesian areal wombling for geographical boundary analysis. *Geogr. Anal.* **37**, 265–285 (2005).
12. Dale, M. & Fortin, M.-J. From graphs to spatial graphs. *Annu. Rev. Ecol. Evolut. Syst.* **41**, 21–38 (2010).
13. Besag, J. Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B (Methodol.)* **36**, 192–225 (1974).
14. Rue, H. & Held, L. *Gaussian Markov Random Fields: Theory and Applications* (CRC Press, 2005).
15. Datta, A., Banerjee, S., Hodges, J. S. & Gao, L. Spatial disease mapping using directed acyclic graph auto-regressive (Dagar) models. *Bayesian Anal.* **14**, 1221 (2019).
16. Cressie, N. *Statistics for Spatial Data* Vol. 4 (Wiley, Terra Nova, 1992).
17. Diggle, P. J., Tawn, J. A. & Moyeed, R. A. Model-based geostatistics. *J. R. Stat. Soc. Ser. C Appl. Stat.* **47**, 299–350 (1998).
18. Gelfand, A. E., Kim, H.-M.J., Sirmans, C. F. & Banerjee, S. Spatial modeling with spatially varying coefficient processes. *J. Am. Stat. Assoc.* **98**, 387–396 (2003).
19. Casetti, E. Generating models by the expansion method: Applications to geographical research. *Geogr. Anal.* **4**, 81–91 (1972).
20. Casetti, E. & Jones, J. P. Spatial aspects of the productivity slowdown: An analysis of us manufacturing data. *Ann. Assoc. Am. Geogr.* **77**, 76–88 (1987).
21. Fotheringham, A. S., Charlton, M. E. & Brunsdon, C. Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis. *Environ. Plan. A* **30**, 1905–1927 (1998).
22. Xue, Y., Schifano, E. D. & Hu, G. Geographically weighted Cox regression for prostate cancer survival data in Louisiana. *Geogr. Anal.* **52**, 570–587 (2020).
23. Finley, A. O. Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods Ecol. Evolut.* **2**, 143–154 (2011).
24. Chan, H. S. R. *Incorporating the Concept of Community into a Spatially-weighted Local Regression Analysis* (University of New Brunswick, 2008).
25. Dormann, F. C. *et al.* Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography* **30**, 609–628 (2007).
26. Sodikin, I., Pramoedyo, H. & Astutik, S. Geographically weighted regression and Bayesian geograpically weighted regression modelling with adaptive Gaussian kernel weight function on the poverty level in West Java province. *Int. J. Humanit. Relig. Soc. Sci.* **2**, 21–30 (2017).
27. Gelfand, A. E. & Schliep, E. M. Spatial statistics and Gaussian processes: A beautiful marriage. *Spat. Stat.* **18**, 86–104 (2016).
28. LeSage, J. P. A family of geographically weighted regression models. In *Advances in Spatial Econometrics* (ed. LeSage, J. P.) 241–264 (Springer, 2004).
29. Ma, Z., Xue, Y. & Hu, G. Geographically weighted regression analysis for spatial economics data: A Bayesian recourse. *Int. Reg. Sci. Rev.* **44**, 582–604 (2021).
30. Liu, Y. & Goudie, R. J. Generalized geographically weighted regression model within a modularized bayesian framework. Preprint at http://arxiv.org/abs/2106.00996 (2021).
31. Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G. & Breidt, F. J. Non-parametric small area estimation using penalized spline regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70**, 265–286 (2008).
32. Wang, H. & Ranalli, M. G. Low-rank smoothing splines on complicated domains. *Biometrics* **63**, 209–217 (2007).
33. Wang, L., Wang, G., Lai, M.-J. & Gao, L. Efficient estimation of partially linear models for data on complicated domains by bivariate penalized splines over triangulations. *Stat. Sin.* **30**, 347–369 (2020).
34. Li, X., Wang, L., Wang, H. J. & Initiative, A. D. N. Sparse learning and structure identification for ultrahigh-dimensional image-on-scalar regression. *J. Am. Stat. Assoc.* **116**, 1994–2008 (2021).
35. Ma, Z., Xue, Y. & Hu, G. Heterogeneous regression models for clusters of spatial dependent data. *Spat. Econ. Anal.* **15**, 459–475 (2020).
36. Luo, Z. T., Sang, H. & Mallick, B. A Bayesian contiguous partitioning method for learning clustered latent variables. *J. Mach. Learn. Res.* **22**, 1748–1799 (2021).
37. Sugasawa, S. & Murakami, D. Adaptively robust geographically weighted regression. *Spat. Stat.* **48**, 100623 (2022).
38. Liu, F. & Deng, Y. Determine the number of unknown targets in open world based on elbow method. *IEEE Trans. Fuzzy Syst.* **29**, 986–995 (2020).
39. Watanabe, S. A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* **14**, 867–897 (2013).
40. Bouguila, N. & Fan, W. *Mixture Models and Applications* (Springer, 2020).
41. Buchin, K. *et al.* Clusters in aggregated health data. In *Headway in Spatial Data Handling* (eds Buchin, K. *et al.*) 77–90 (Springer, 2008).
42. Miller, J. W. & Harrison, M. T. A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems* Vol. 26 (eds Miller, J. W. & Harrison, M. T.) (Neural Information Processing Systems Foundation, Inc, 2013).
43. Laome, L., Budiantara, I. N. & Ratnasari, V. Estimation curve of mixed spline truncated and Fourier series estimator for geographically weighted nonparametric regression. *Mathematics* **11**, 152 (2022).
44. Laome, L., Budiantara, I. N. & Ratnasari, V. Construction of a geographically weighted nonparametric regression model fit test. *MethodsX* **12**, 102536 (2024).
45. Gao, X., Xiao, B., Tao, D. & Li, X. A survey of graph edit distance. *Pattern Anal. Appl.* **13**, 113–129 (2010).
46. Cho, S.-H., Lambert, D. M. & Chen, Z. Geographically weighted regression bandwidth selection and spatial autocorrelation: An empirical example using Chinese agriculture data. *Appl. Econ. Lett.* **17**, 767–772 (2010).
47. Bullock, R. Great circle distances and bearings between two locations. *MDT* **5**, 1–3 (2007).
48. Quintana, F. A., Müller, P., Jara, A. & MacEachern, S. N. The dependent Dirichlet process and related models. *Stat. Sci.* **37**, 24–41 (2022).
49. Yamato, H. Dirichlet process, Ewens sampling formula, and Chinese restaurant process. In *Statistics Based on Dirichlet Processes and Related Topics* (ed. Yamato, H.) 7–28 (Springer, 2020).
50. Reich, B. J. & Fuentes, M. A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Ann. Appl. Stat.* **1**, 249–264 (2007).
51. Sethuraman, J. A constructive definition of Dirichlet priors. *Stat. Sin.* **4**, 639–650 (1994).
52. Hosseinpouri, M. & Khaledi, M. J. An area-specific stick breaking process for spatial data. *Stat. Pap.* **60**, 199–221 (2019).

53. Ishwaran, H. & James, L. F. Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.* **96**, 161–173 (2001).
54. de Valpine, P. *et al.* Programming with models: Writing statistical algorithms for general model structures with nimble. *J. Comput. Graph. Stat.* **26**, 403–413 (2017).
55. Yu, G., Huang, R. & Wang, Z. Document clustering via Dirichlet process mixture model with feature selection. In *Proc. of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 763–772 (2010).
56. Dahl, D. B. Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian Inference Gene Expr. Proteomics* **4**, 201–218 (2006).
57. Rand, W. M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).

Acknowledgements

We would like to express our gratitude to the team at Children's Health Queensland and the Centre for Data Science for their invaluable assistance and support in this project.

Author contributions

Conceptualization: Wala Draidi Areed, Aiden Price, Helen Thompson, Kerrie Mengersen. Data curation: Wala Draidi Areed, Aiden Price, Reid Malseed, Kerrie Mengersen. Formal analysis: Wala Draidi Areed. Investigation: Wala Draidi Areed. Methodology: Wala Draidi Areed, Kerrie Mengersen. Software: Wala Draidi Areed. Supervision: Aiden Price, Helen Thompson, Reid Malseed, Kerrie Mengersen. Validation: Wala Draidi Areed. Visualization: Wala Draidi Areed. Writing – original draft: Wala Draidi Areed. Writing - review and editing: Wala Draidi Areed, Aiden Price, Helen Thompson, Reid Malseed, Kerrie Mengersen.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-59973-w.

Correspondence and requests for materials should be addressed to W.D.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.