Spatial Statistical Machine Learning Models to Assess Health Vulnerabilities in Children

Wala Draidi Areed (w.areed@qut.edu.au)
Queensland University of Technology
Aiden Price
Queensland University of Technology
Kathryn Arnett
Children's Health Queensland
Kerrie Mengersen
Queensland University of Technology

Research Article

Keywords: statistical machine learning methods, spatial random forest, health vulnerabilities

Posted Date: December 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1106160/v1

RESEARCH

Spatial Statistical Machine Learning Models to Assess Health Vulnerabilities in Children

Wala Draidi Areed[1*], Aiden Price[1], Kathryn Arnett[2] and Kerrie Mengersen[1]

[*]Correspondence:
w.areed@qut.edu.au
[1]School of Mathematical Science, Center for Data Science, Queensland University of Technology, Queensland, Australia
Full list of author information is available at the end of the article

Abstract

Background: The health and development of children during their first year of school is known to impact their social, emotional, and academic capabilities throughout and beyond early education. Physical health, motor development, social and emotional well-being, learning styles, language and communication, cognitive skills, and general knowledge are all considered to be important aspects of a child's health and development. It is important for many organisations and governmental agencies to continually improve their understanding of the factors which determine or influence health vulnerabilities among children. This article studies the relationships between health vulnerabilities and educational factors among children in Queensland, Australia. In Queensland, the percentage of children who are developmentally vulnerable in at least one domain in 2018 was around 26%, and the overall percentage of attendance at preschool was around 75.4% These are the lowest rates among all states and territories of Australia. There is also substantial geographic variation in rates across the state.

Methods: Spatial statistical machine learning models are reviewed and compared in the context of a study of geographic variation in the association between health vulnerabilities and attendance at preschool among children in Queensland, Australia. A new spatial random forest (SRF) model is suggested that can explain more of the spatial variation in data than other approaches.

Results: In the case study, spatial models were shown to provide a better fit compared to models that ignored the spatial variation in the data. The SRF model was shown to be the only model which can explain all of the spatial variation in each of the health vulnerabilities considered in the case study. The spatial analysis revealed that the attendance at preschool factor has a strong influence on the physical health domain vulnerability and emotional maturity vulnerability among children in their first year of school.

Conclusion: This study confirmed that it is important to take into account the spatial nature of data when fitting statistical machine learning models. A new spatial random forest model was introduced and was shown to explain more of the spatial variation and provide a better model fit in the case study of health vulnerabilities among children in Queensland. At small-area population level (statistical area level 2 (SA2)), increased attendance at preschool was strongly associated with reduced physical and emotional health vulnerabilities among children in their first year of school.

Keywords: statistical machine learning methods; spatial random forest; health vulnerabilities

Introduction

Hospitals have started engaging their local populations in recent years to improve outreach and preventive health activities. Many of these efforts are being carried out under the name of enhancing "population health". As Casalino and colleagues [1] stated, "Everyone in health care is working to improve population health these days. Or will be very soon. Or feel that they ought to be". Hospitals which have typically focused on primary health care have started to acknowledge population health as a core component of their community commitment and strategic programs. Mutual service, health improvement, physical and environmental change and economic growth are supported through population health services [2].

Research conducted in 2017 by the American Hospital Association found that children's hospitals invested a higher share of their overall community service costs than adult general hospitals [3]. Some children's hospitals see population health as an opportunity for new initiatives to be adopted, to resolve the social determinants of health and to understand the need to shift current cultural institutional society to meet their objectives [1].

Participating in preschool programs the year before entering school has been reported to help children acquire healthy habits and can help to lessen disparities in developmental outcomes for vulnerable groups [4]. Preschool attendance has emerged as a national policy issue in many countries, including Australia. A variety of variables might impact whether or not a child attends preschool; for example, cultural obstacles to preschool participation might exist for non-English speaking and Indigenous households. Furthermore, the quality and quantity of preschool services available to children in rural and remote places may be less than in major cities [5].

The Australian Early Development Census (AEDC) [6] is a population-based cross-sectional census of early childhood development, derived from the Canadian Early Development Instrument. The AEDC elicits information about children's demographics and early developmental outcomes (physical health and well-being, social competence, emotional maturity, language and cognitive skills (school-based), communication skills and general knowledge). Teachers complete the AEDC for all Australian children in their first year of compulsory school. Figure (1) shows the five domains of health vulnerabilities measured by AEDC for children in their first year of school.

For reasons of privacy or communication, population health data and associated socio-demographic data collected about patients, families and constituent communities are often released at the level of small area aggregates. These SA2 areas are typically determined on the basis of health management or statistical divisions. It is common to practice to map these statistical area data and assess demographic patterns in order to promote resource distribution and evidence-based policy making and planning. However, statistical analysis of aggregated spatial data presents specific challenges, particularly in assessing spatial patterns or identifying associations between health, potential socio-demographic factors and other potential explanatory variables. Using regression or classification approaches that ignore the spatial structure of data can be insufficient [7].

A range of statistical machine learning models are now available that take into

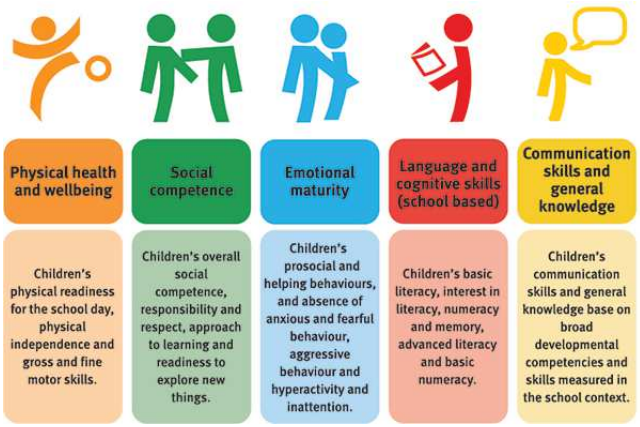

Figure 1 five Domains of Health Vulnerabilities from AEDC.

account the spatial nature of the data. Simple approaches include adding geographic coordinates or distance metrics to familiar models such as linear regression, random forests and neural networks [8]. More sophisticated geographic extensions of these approaches, as well as combinations of models, have also been proposed [9, 10, 11, 12, 13].

Interestingly, these spatial models may not capture all of the spatial autocorrelation in the data. The presence of spatial autocorrelation in the residuals after fitting a model suggests that the model estimates and predictions could be imprecise or biased [14]. In this article, we suggest a spatial random forest (SRF) model that can explain more of the spatial variation in the data than other common statistical machine learning approaches. We describe this approach in the context of a review of established popular aspatial and spatial statistical machine learning models, and compare the methods in a case study of health vulnerabilities among children in Queensland, Australia. The aims of the study are two fold: to evaluate spatial variation in these vulnerabilities, and to assess the relationship between the proportion of vulnerable children and the proportion of children attending preschool, based on aggregated small area data (SA2 level).

Materials and Methods

This section discusses the case study area and sources of data, then provides a short review of aspatial and spatial linear models, random forests and neural networks. A new spatial random forest method is also introduced in this section.

Study Area

Queensland is the second largest and third most populous Australian State or Territory, and is located in the northeast of the country. With strengths in mining, agriculture, tourism, international education, insurance, and banking. Queensland also has the third largest economy [15, 16]. The State is divided geographically into 528 non-overlapping statistical area level 2 (SA2) regions (according to the ASGS 2011 boundaries of the Australian Bureau of Statistics, ABS). SA2 regions

are medium-sized general purposed areas that are designed to represent a community that interacts together socially and economically (www.abs.gov.au). This is the smallest area for the release of ABS non-census and inter-censal statistics, including the estimated resident population and health data, and data from the 2016 Census of Population and Housing.
In this study, health and socio-demographic data are obtained at the SA2 level for 526 SA2s, excluding those with zero population and with offshore/migratory or undefined location.

The Data Repository

The outcome variables considered in this study were health vulnerabilities, provided by the Australian Early Development Census (AEDC). The AEDC takes place every three years and is the world's most extensive data gathering for children. Teachers complete the census for their students in their first year of school, and their answers are used to construct domain scores. Each child is given a score between zero and ten for each of the AEDC domains, using the cut-offs established as a baseline in 2009, children falling below the 10^{th} percentile in a domain, taking into account the age differences, are categorised as "developmentally vulnerable".
In this study, the outcome variable of interest is the SA2 level health vulnerability score for each domain, which is the age matched proportion of developmentally vulnerable children in the SA2. Five health vulnerabilities were considered in this study. These include: physical health and well-being domain vulnerability (PHD), social competence domain vulnerability (SCD), emotional maturity domain vulnerability (EMD), language and cognitive skills domain vulnerability (LCS), communication skills and general knowledge domain vulnerability (CS), and two health domain indicators which are vulnerable on one or more domain(VOD), and vulnerable on two or more domains (VTD).
The covariate information was extracted from the ABS and AEDC for each SA2. The covariates of interest obtained from the ABS included a geographic remoteness category, a Socio-Economic Index for Area (SEIFA) score, specifically an Index of Relative Socio-Economic Disadvantage (IRSD), mother's language, country of birth, Indigenous status, and attendance at preschool. These covariates are also gathered as part of the survey AEDC and aggregated for research purposes.
The ABS classification of geographical remoteness is major city, inner-regional, outer-regional, remote and very remote. In Queensland there are 294 SA2 areas categorised as major cites, 113 SA2 areas as inner regional, 96 SA2 areas as outer regional, 11 SA2 areas as remote and 14 SA2 areas as very remote area [17].
The SEIFA score is a broad socioeconomic index that summarises a variety of data on individual and family economic and social condition in a given area. This factor is coded from 1 to 10. A low score suggests that the area in general is at a disadvantage. For example, low-income households, or people without qualifications or in low skill occupations.
Binary classifications were used for mother's language (English, other), Indigenous status (Indigenous, not),Country of birth (Australia/not Australia) and attendance at preschool (yes, no).
The data custodians listed the above data over different time periods. In this study,

we collect annual data only from 2018-2019. All count covariates acquired in this study have been transformed into proportions of children in an SA2 region with the feature of interest. There were a small number of missing observations in this dataset. Missing continuous data has been imputed using spatial neighbourhood averages. For categorical data, imputation was instead taken as the highest frequency neighbourhood category. In two instances, missing values for two islands could not be filled, as the regions have no contiguous neighbours. As a result, the analysis carried out in this study was reduced to the remaining 526 SA2 regions.

Overall Measures of Spatial Variation

Moran's I [18] and Geary's C [11] are popular measures to determine whether the data are geographically clustered, randomly distributed, or uniformly distributed in space. The semi-variogram, which depicts the range and rate at which spatial autocorrelation decreases, is another tool for measuring spatial dependency [19]. The semi-variance of a dataset with spatial autocorrelation typically grows to a maximum value before levelling off. Moran's I [18] can be calculated as

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{S \sum_{i=1}^{n} z_i^2}, \tag{1}$$

where $z_i = x_i - \overline{x}$ and $S = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}$. Here, x_i is the independent variable, $\overline{x}$ is the associated sample mean and w_{ij} is an element of the spatial S matrix, which shows the degree of spatial connection between regions i and j [20]. The range of Moran's I is between -1 and 1, where -1 is perfect dissimilarity clustering, 0 means that there is no spatial autocorrelation, and 1 indicates perfect similarity clustering.

Tangos' maximized excess events test (MEET) [21] is another way to detect the spatial variation inside the data. This measure assumes a range of spatial scale parameters and depends on a weight function. Tango's (MEET) has been shown to have very good statistical power in detecting global disease clustering [21]. Tango [22] proposed a distance based exponential weight function for MEET, but other choices of weights are also possible. one feature of this test is that it considers a range of spatial scale parameters, adjusting for the multiple testing Tango's (MEET) has been shown to have very good statistical power in detecting global disease clustering. Let c_i be the observation in a region i, n_i the population size of the region i, C the total number of observations, N the total population, d_{ij} the distance between region i and j, and $u_{j(i)}$ the population size in region i and its j nearest neighbors. then Tango's EET can be defined as:

$$\text{Tango's EET} = \sum_{i} \sum_{j} w_{ij} (c_i - n_i \frac{C}{N})(c_j - n_j \frac{C}{N}) \tag{2}$$

where w_{ij} is typically defined as $w_{ij} = e^{-4(\frac{d_{ij}}{\lambda})^2}$ [22] or $w_{ij} = e^{(-\frac{d_{ij}}{\lambda})}$ [23] where, λ is a measure of spatial scale clustering.

Random Forest

A popular type of machine learning method is the decision tree, which has been shown to be fast, flexible, and can deal with large amounts of data [24]. A decision tree is built by grouping data using a recursive binary partitioning algorithms into more homogeneous groups. Each binary split is selected based on specified splitting criteria.

The random forest (RF) approach suggested by Breiman [25] creates and combines a large number of individual decision trees generated from the data set of interest. Random forests address some of the drawbacks of having a single classification and regression tree (CART), such as over-fitting, correlation between variables and trees sets of splits to describe non linear relationships. It is a combination of random sub-space methods [26] and bagging [27]. Every decision tree is randomly generated by sampling a proportion(usually approximately two-thirds) of the training data and leaving the reminder out of training. Furthermore, at each decision node, only a subset of features is picked at random during the tree construction process. The final result is generated using the majority vote (classification) or the average prediction of all trees (regression). Simultaneously, the data left out of training for each tree is used to compute an output goodness of fit assessment called the out-of-bag (OOB) error estimate [25]. The importance of the covariates can be calculated using the OOB error, where the most effective approach is to use the increase in the mean squared error (iMSE) measure to determine variable importance. The values of each feature are permuted randomly, and the OOB error iMSE (increase in mean square error) is calculated,

$$iMSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2, \tag{3}$$

where n is the total number of points, and y_i, $\hat{y_i}$ are the observed values and the estimated values, respectively [28]. This method can be used for both regression and classification problems in many different fields [29, 30, 31, 32, 33].

Random Forest for Spatial Data

A number of approaches have been proposed for applying a random forest to spatial data. Longitude and latitude were introduced as covariates in several efforts to integrate a spatial context into machine learning [34, 12, 35]. For example, Behrens [12] used x- and y-coordinates and distances to the corners and center of a bounding box around the sampling locations as covariates. Random Forest for Spatial Prediction (RFsp) was developed by Hengl [8], and uses buffer distance maps from observation points as covariates. In the next section we discuss another popular approach, the geographical random forest (GRF).

Geographical Random Forest

The GRF is a disaggregation consisting of several local sub-models [13]. It uses a similar idea to geographical weighted regression (GWR) [36]. Here, a local RF is computed for each location i based only on nearby observations. Thus for each training data point, a RF is developed, each with its own efficiency, predictive

ability, and feature importance. As a result, the stability of the RF is measured locally rather than globally. To understand more about this algorithm, we use a regression equation of the form,

$$Y_i = ax_i + e, \quad i = 1 : n \tag{4}$$

where Y_i is the ith observation's value of the response variable, ax_i is the RF's nonlinear prediction based on a set of x covariates, and e is the error term. For GRF, equation (4) is extended as follows.

$$Y_i = a(u_i, v_i)x_i + e, \quad i = 1 : n; \tag{5}$$

where $a(u_i, v_i)x_i$ is the RF model prediction for location i. and (u_i, v_i) are the spatial coordinates. The neighbourhood (or kernel) is the field in which the sub-model runs. The bandwidth is the maximum distance between a data point and its kernel [37]. There are two kinds of kernels that are commonly used, "adaptive" and "fixed" [38]. When sampling density varies across space, using an adaptive kernel is beneficial [13].
A GRF can be used to achieve two goals: firstly to enhance predictions over a standard RF, and secondly to extract spatially differentiated model parameter inferences. The degree of spatial variation in the data and the required bandwidth selection determine the increase in efficiency. Moreover, a GRF model can be used as a simple guide to investigate the data's local structure and improve our understanding of how spatial processes affect this structure.

Spatial Random Forest
In this section we introduce an alternative to the GRF, based on an extension of the global random forest algorithm. Here, a second stage is added to the RF to absorb residual spatial autocorrelation in the data.
This algorithm is described as a set of three steps.
Step 1: Determine a neighbourhood for each spatial region. (In our case study we adopt a contiguous neighbour definition that accepts any region that shares at least one boundary). See figure (2).
Step 2: Find the global random forest (RF):

$$RF_1 \sim (y, x_i), \tag{6}$$

Step 3: Find the residual using the neighbourhoods

$$r_i = \frac{\sum_{j \sim i}(y_j - \hat{y_j})}{n_j} \tag{7}$$

Here, y_i is the observed values, $\hat{y_i}$ is the estimated values using RF_1, and $j \sim i$ denotes all regions j in the neighbourhood of the ith region. Note that, in contrast to common measures such as mean absolute error (MAE) and mean square error (MSE), the neighbouring residuals are simply summed in the above equation. This is consistent with the concept of spatial correlation, in the set of residuals with

different signs indicate a weaker spatial sign nature compared to a set with consistently positive region or negative signs.
Step 4: Apply

$$RF_2 \sim (y, \{x_i, r\}) \tag{8}$$

Note that this method borrows conceptually from the conditional autoregression (CAR) approach.

Neural Network

Like the RF, Neural Networks (NN) provide more representational flexibility and freedom from the constraints of a linear model [39]. A NN is made up of many or perhaps millions of tightly linked basic processing nodes. The majority of today's neural networks are structured into layers of nodes and are "feed-forward," meaning that data flows in just one direction through them. A single node may be linked to multiple nodes in the layer below it from which it receives data, as well as several nodes in the layer above it from which it transmits data. A node will assign a numerical "weight" to each of its incoming connections. Each node receives a new data item a distinct number, and multiplies it by the associated weight. The resulting items are then added together to produce a single number. If that number falls below a certain threshold, the node does not send any data to the next layer. An input, hidden, and output layer make up a basic NN. The connection weights of a basic NN from the hidden to the output layer can be interpreted as the coefficients of a linear model of non-linearly transformed variables, namely the outputs of the hidden neurons [40].

Neural Network for Spatial Data

One way of using neural networks for spatial data is to use the longitude and latitude as a covariate. We call this method a spatial neural network (SNN). Another recent extension of NN for spatial data is the geographically weighted artificial neural network (GWANN) [40]. Each output neuron of GWANN has as a geographic location associated to it. This allows the spatial distances between the observations and the output neuron's location to be calculated. As a result, the connection weights between the hidden and output layers can be understood as a geographical weighted regression GWR model when estimated using a geographically weighted error function.
The key distinction between a GWANN and a basic ANN is that a GWANN calculates an error signal using a geographical weighted error function rather than the standard quadratic error function [40] [10]. The geographically weighted error function is given by:

$$E = \frac{1}{2} \sum_{i=1}^{n} v_i (t_i - o_i)^2, \tag{9}$$

where t_i is the target value, o_i the output of output neuron i, v_i the geographically weighted distance between the observation and the location of output neuron i,

and n the number of targets. From equation (9), the difference between the output neuron's value and the target value is weighted by the spatial distance between output neuron's location and the observation; when the output neuron's location and observation are close, the difference is given more weight than when they are farther apart.

A 10-fold cross validation (CV) is typically used to calculate the number of GWANN training iterations. The models are trained within each fold until their performance on the current fold's test data does not increase after many iterations. The additional iterations are designed to offer networks a chance to break free from local minima. This method, known as "early stopping with patience" [41] minimises the chance of overfitting the training data significantly. The iteration with the best mean performance across all folds, as well as the performance value obtained, are then presented.

Garson [42] devised a method for calculating the relative importance of each of the input variables based on the connection weights. In this algorithm each variable's input is stored as a weight in the network model, and the contribution of each of these variables to the output is largely determined by the magnitude and direction of these link weights. A positive connection weight enhances the magnitude of the network output, whereas a negative weight suppresses the value of the response variable [43]. Furthermore, when compared to the other factors, a variable with a considerably larger connection weight is regarded to have a bigger influence on the network output. Thus,

$$RI_x = \sum_{y=1}^{m} w_{xy} w_{yz} \tag{10}$$

where RI_x denotes the relative importance of input neuron x, w_{xy} the final weights of the connection from input neuron to hidden neurons, w_{yz} the final weights of the connection from hidden neuron to output neuron. y represents the total number of hidden neurons, and z is the output neuron.

Linear Model

Given a dependent or response vector y, and a matrix of input variables X, a general linear model can be expressed as

$$Y = X\beta + \epsilon \tag{11}$$

where β is a vector of regression coefficients and ϵ is a vector of residuals. In normal linear regression, ϵ is assumed to have a zero mean Gaussian distribution. The matrix X contains independent variables or covariates $\{X_1, X_2, ..., X_p\}$, as well as any interactions or functions of these variables (e.g, quadratic or cubic terms to allow for non linear relationships between the independent and response variables).

Linear Models for Spatial Data

The generalized liner model (GLM) can be extended to include non-normal responses via a generalized linear model, or additive terms via generalised additive

model GAM. A spatial GLM or a spatial GAM is another way to model the spatial data. Non-Gaussian error distributions and non-linear correlations between response and predictor variables are supported by these regression techniques. GAMs are non-parametric extensions of linear model regressions in which non-parametric smoothers f are applied to each predictor and the component response is calculated additively, i.e,

$$E(Y) = \beta_0 + f_1(x_1) + f_2(x_2) + f_3(x_3, x_4) + ... + f_k(x_k). \tag{12}$$

In the most simple form, latitude and longitude can be used as model inputs [44]. The spatial autoregressive (SAR) model proposed by Whittle [45] is a spatial approach for describing the connection between dependent and independent variables by taking the spatial effect into account. It features an autoregressive structure that represents the spatial dependency of the attributes using a precision matrix that is generally a function of the proximity between regions [46]. Moran's I [18] can be used to confirm the presence of spatial variation before the SAR model is used. Weights are used to indicate the impact of location effects on the data [47]. The general formula of the spatial regression called SARMA (Spatial Autoregressive Moving Average [48]) is given as

$$Y = \rho WY + X\beta + \lambda Wu + \epsilon \tag{13}$$

where Y is the dependent variable, X is the matrix of independent variables, ρ is the spatial autoregressive parameter, W is a weights matrix, β is a regression coefficient vector, λ is a spatial error coefficient, and ϵ is a residual vector.

Conditional Autoregressive Model (CAR)
Bayesian models are especially well adapted to spatial modelling because the information particular to each region may be represented as priors, and both correlated and uncorrelated spatial effects can be investigated [49]. A popular spatial prior proposed by Leroux [50] includes a spatial random effect ψ such that,

$$\psi \sim MVN(0,\, D) \tag{14}$$

with covariance matrix D, where D is usually described by its generalized inverse [50]

$$\sigma^2 D^- = (1 - \rho) + \rho R. \tag{15}$$

Here, R is the intrinsic autoregression matrix which represents the neighbourhood structure of the regions with typical element R_{ij}, which equals n_i when $i = j$, where n_i is the numbers of neighbours of region i, and $I(i \sim j)$ otherwise, where $I(i \sim j)$ is an indicator function taking the value 1 when i and j are neighbours.
The term ρ is introduced as a spatial dependence parameter, $\rho \in [0, 1]$, whose two extreme cases give rise to the independence model (i.e., $\psi_i = v_i$ and $D = \sigma^2 I$). The spatial residual is typically considered to have an independent normal

distribution $v_i \sim \mathcal{N}(0, \sigma_v)$, and intrinsic auto regression (i.e., $\psi_i = u_i$ and $D = \sigma^2 R^-$), respectively [50]. For ρ close to 1, the conditional variance becomes close to σ^2/n_i and for ρ close to 0, the variance becomes close to σ^2, that is independent of the number of neighbours n_i [51].
The univariate full conditional distribution for $\psi_i|\psi_{-i}$ can be written as

$$\psi|\psi_{-i} \sim \mathcal{N}(\frac{\rho}{n_i\rho + 1 - \rho}\sum_{j\sim i}\psi_j, \frac{\sigma^2}{n_i\rho + 1 - \rho}) \tag{16}$$

where ψ_{-i} denotes the random effect vector with the ith component deleted.

Model Evaluations

We use three well-established and reliable measures to assess model fit and accuracy: coefficient of determination R^2, root mean square error (RMSE) and Moran's I. Here,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{17}$$

and

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}, \tag{18}$$

where n is the total number of points, y_i, $\hat{y}_i$ and $\overline{y}_i$ are the actual values, estimated values, and the averaged values, respectively. Moran's I [18] was discussed earlier and is another way to judge the consistency of a model applied to geographical and spatial data.
The importance of variables for the RF, SRF, GRF can be determined by the mean square error (iMSE) and impurity reduction. The impurity reduction introduced by a split is maximised using RF splitting criteria. A split with a significant decrease in impurity is considered important for the impurity. In addition, the impurity importance for a variable x_i is calculated by the sum of all impurity decrease measures of all nodes in the forest. Consider splitting a regression tree T at a node t. Let s be a proposed split for a variable X that splits t. Regression node impurity is determined by within node sample variance

$$\delta(t) = \frac{1}{N}\sum_{x_i \in t}(Y_i - \overline{Y_t})^2, \tag{19}$$

where $\overline{Y_t}$ is the sample mean for t and N is the sample size of t [52].

Implementation of case study

The case study was carried out in different steps as explained below.

- **Prepare the Data**

In this step we prepared the data set for analysis. The covariates data were provided on the form of counts per SA2 region

1 Converting the continuous covariates to proportions. This was done by dividing each SA2 data by the SA2 population of children in their first year in school. For the responses the data were already given as proportions from AEDC.

2 For spatial analysis The centroids, longitude, latitude and contiguous boundaries were determined for each SA2 region, and added to the data set. Figure (2) shows the contiguous centroids for each SA2 region in Queensland.

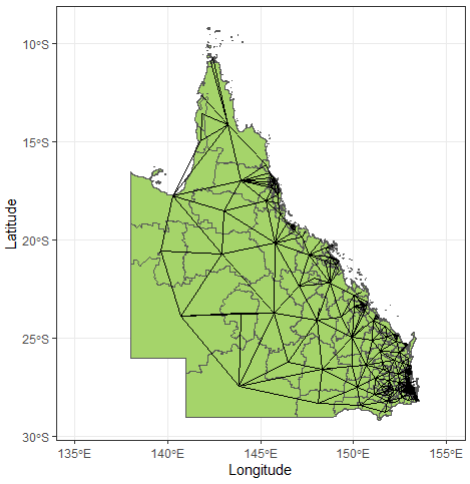

Figure 2 Queensland Centroid for SA2.

3 For neural network,we prepared the data by using dummy coding for categorical variables, to convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns.

4 Normalize the data. We took the logarithm of each variable to increase the applicability of the different statistical machine learning methods.

5 Specify model hyper-parameters. The numbers of hidden layers were specified for the neural networks models, and the optimal bandwidth, number of trees and number of iterations were specified for the random forest models using cross validation.

6 Split the data into testing data (20%) and training data (80%) for the random forest models. The testing data were used to find the importance plots.

7 In the Bayesian spatial linear regression models, a neighborhood matrix was calculated using contiguous regions with shared boundaries.

- **Analyse the Data**

1 The statistical analysis was conducted using the R programming environment [53, 54, 55] and utilised a number of packages, including `Random Forest` [56] for random forest calculations, `ggplot2` [57] for visualizing the data, `caret` [58] for data preparation and separation, `spatialML` [59] for geographical random forest (GRF) model, `neuralnet` [60] for neural network and spatial neural network, `GWANN` [40] for geographical weighted artificial neural network and `CARBayes` [61] for Bayesian spatial linear regression modelling.

2 In the random forest models analyses, the impurity reduction and the iMSE values were calculated for each parameter to determine variable importance.

3 The longitude and latitude were included as a covariates for spatial neural network and the relative importance was calculated.

4 For the GAM model, cubic spline smoothing functions were used between the cut points. Cross validation was used to determine the optimal number of knots, and interactions between the covariates were also included in the model.

- **Evaluate Model Fit**

After implementing the statistical machine learning methods, the values of R^2, $RMSE$ and Moran's I were calculated for each model.

Results

Figure (3) shows the correlation plot between the seven types of vulnerabilities in the case study. The strongest correlation is between vulnerability on one or more domains (VOD) and vulnerability on two or more domains (VTD), where the Pearson correlation coefficient is 0.9. while the weakest correlation is between physical health domain vulnerability (PHD) and emotional maturity domain vulnerability (EMD) which is around 0.51.

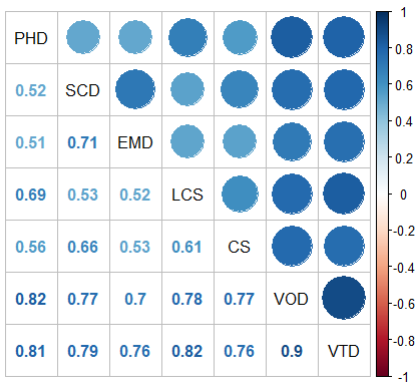

Figure 3 Correlation plot between different types of vulnerabilities

Tables (1) and (2) show the values of the coefficient of determination R^2 and the $RMSE$ for the models considered in this study.

Table 1 Values of the coefficient of determination R^2 for different statistical machine learning models.

Methods	PHD	SCD	EMD	LCS	CS	VOD	VTD
GLM	0.299	0.212	0.166	0.385	0.372	0.417	0.411
SAR	0.336	0.224	0.214	0.555	0.382	0.429	0.426
RF	0.702	0.730	0.587	0.729	0.734	0.752	0.717
GRF	0.759	0.722	0.669	0.782	0.788	0.811	0.778
SRF	0.771	0.704	0.616	0.737	0.755	0.691	0.707
GAM	0.307	0.239	0.208	0.525	0.377	0.404	0.462
GAM with interaction	0.506	0.348	0.287	0.625	0.379	0.469	0.467
SGAM	0.559	0.472	0.323	0.623	0.42	0.614	0.623
NN	0.611	0.604	0.568	0.684	0.684	0.590	0.669
SNN	0.719	0.737	0.679	0.713	0.726	0.684	0.689
GWANN	0.694	0.705	0.662	0.707	0.737	0.671	0.691
CAR	0.879	0.729	0.787	0.875	0.771	0.802	0.826

Table 2 Values of the $RMSE$ for different statistical machine learning models.

Methods	PHD	SCD	EMD	LCS	CS	VOD	VTD
GLM	0.091	0.082	0.096	0.082	0.063	0.093	0.061
SAR	0.042	0.040	0.0348	0.033	0.036	0.049	0.041
RF	0.036	0.035	0.028	0.028	0.031	0.048	0.038
GRF	0.031	0.021	0.026	0.022	0.02	0.041	0.032
SRF	0.034	0.032	0.027	0.026	0.031	0.045	0.034
GAM	0.054	0.049	0.039	0.052	0.052	0.107	0.051
GAM with interaction	0.047	0.044	0.037	0.033	0.052	0.068	0.048
SGAM	0.042	0.052	0.037	0.037	0.038	0.066	0.044
NN	0.054	0.094	0.114	0.057	0.081	0.082	0.075
SNN	0.050	0.081	0.111	0.051	0.079	0.081	0.071
GWANN	0.054	0.050	0.038	0.054	0.047	0.067	0.052
CAR	0.031	0.034	0.026	0.031	0.029	0.056	0.034

From these models we can see that the GAM with interaction performs better than the GAM without interactions, which indicates non linear and complex relationships between the socio-demographic and education covariates and the health vulnerabilities. This is reinforced by the improved fit of the RF and NN compared to the GAM and GLM models. The table also reveals that the value of including spatial information. The values of $RMSE$ are reduced and the values of R^2 are increased considerably for SAR, GRF, RF, SGAM, GWANN and CAR models compared to their non-spatial counterparts.

Among the spatial models, the Bayesian CAR model provided the largest R^2 value, and this model and GRF gave the smallest $RMSE$ values for most of health outcomes vulnerabilities.

The importance of attendance at preschool on the health outcomes vulnerabilities was assessed in the models that were considered to be reliable in term of goodness of fit R^2 and accuracy $RMSE$.

Table (3) shows the relative importance of attendance at preschool for the RF, GRF, and NN models. It can be seen that the attendance at preschool variable plays a

Table 3 The importance percentages for RF,GRF, SRF respectively, and the relative importance values for NN, for proportion of attendance at preschool (educational factor).

Responses	RF	GRF	SRF	SNN
PHD	27.80%	45.93%	33.12%	-0.06
SCD	9.47%	13.41%	8.25%	-0.02
EMD	39.30%	44.80%	23.77%	-0.03
LCS	6.89%	1.00%	4.056%	-0.07
CS	13.36%	5.76%	8.69%	-0.01
VOD	3.20%	3.63%	2.86%	-0.02
VTD	2.86%	3.30%	2.69%	-0.03

major role in the analyses of the physical health and well being domain, and the emotional maturity domain in the RF, GRF and SRF models. In contrast, attendance at preschool does not appear to play a major role for vulnerability on one or more domain or two or more domains. Furthermore, Garson's algorithm showed evidence that as attendance at preschool increased, the health vulnerabilities decreased, based on SA2 level data.

Figures (4) and (5) show the values of the $iMSE$ for the two vulnerabilities for which attendance at preschool was found to be important. It is apparent that attendance at preschool was the most important variable for physical health and wellbeing domain vulnerability, followed closely by IRSD, and second most important (after IRSD) for the emotional maturity domain vulnerability. These two variables, atten-

dance at preschool and IRSD, were substantially more important than any of the other variables considered.

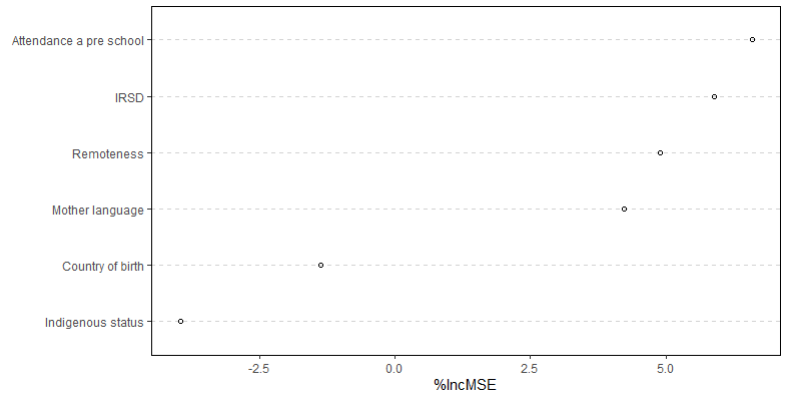

Figure 4 The importance plot using GRF for physical health and well-being domain vulnerability.

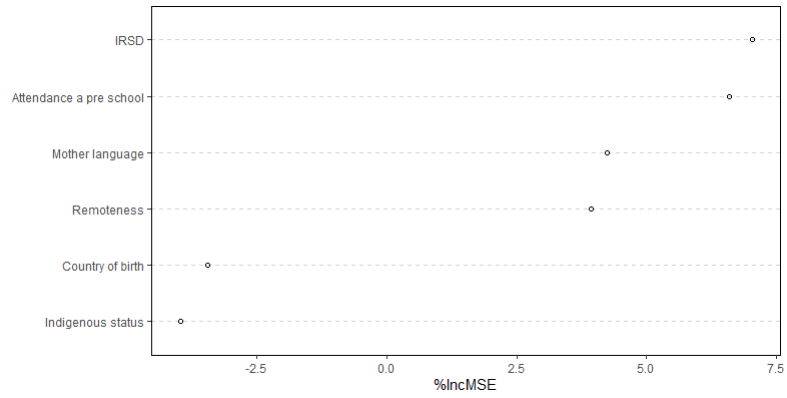

Figure 5 The importance plot using GRF for emotional maturity domain vulnerability.

Table (4) shows the values of spatial autocorrelation (Moran's I) for the residuals.

Table 4 Moran's I (P-values) for the residuals form the different models.

Responses	RF	GRF	SRF	SNN	GWANN	CAR
PHD	1e-04	1e-04	0.944	0.014	1e-04	0.001
SCD	0.012	0.002	0.942	0.0.231	0.001	0.796
EMD	1e-04	0.686	0.944	0.151	0.003	0.891
LCS	1e-04	0.008	0.944	0.0.027	1e-04	0.003
CS	0.104	0.0.002	0.942	0.489	0.003	0.003
VOD	1e-04	0.004	0.943	7e-04	0.005	0.004
VTD	0.003	0.021	0.943	0.024	0.004	0.101

According to this table the new spatial random forest was the only model to adequately fully explain spatial variation in the data, for all the health outcomes vulnerabilities. None of the GRF, GWANN or CAR models captured as much of the spatial variation.

Finally, Table (5) shows that the posterior median is substantively different from

zero, since the credible interval does not include zero. The negative value indicates as the proportion of attendance at preschool increases the proportion of vulnerabilities between children decreases.

Table 5 The posterior median and credible intervals for attendance at preschool from different types of vulnerabilities.

Responses	Posterior median	95% Credible intervals
PHD	-0.071	[-0.139,0.001]
SCD	-0.071	[-0.134,-0.004]
EMD	-0.046	[-0.102, 0.012]
LCS	-0.058	[-0.115,-0.002]
CS	-0.081	[-0.141,-0.022]
VOD	-0.066	[-0.148, 0.018]
VTD	-0.093	[-0.168,-0.026]

The actual data and model estimates are presented as maps in the Appendix, for the SRF, GRF and CAR models.

Discussion

This study achieved different aspects related to population health data. First, a new spatial random forest model has been proposed. Second, the model has been integrated in a broader review of aspatial and spatial statistical machine learning models. Third, these models have been applied to a substantive case study which aimed to evaluate the relative importance of attendance at preschool on health vulnerabilities of children in Queensland, based on aggregated small area data (statistical area level SA2 data).

It is acknowledge that the results of the case study depend on the choice of covariates considered in the models and on the scale of data aggregation. In this case, we used publicly available data from AEDC and ABS aggregated at small area SA2. It is assumed that by increasing the spatial resolution of data and adding other variables affecting the health vulnerabilities, other insights will arise. Other covariates that could be considered relate to religion, ethnic background, parent's education, period of pregnancy and the child's weight at birth.

When comparing the models, it was apparent that including spatial features resulted in a better fit. The drawback for GRF was that it needs more time to run in comparison with SRF and SNN: GRF required around 6.25 minutes to run for each type of vulnerability with 400 bandwidth, while the SRF, NN needed 4.3 and 5.6 seconds respectively. Bayesian spatial linear modelling nedded 2.6 minutes to run.

In this study, the spatial neighbourhood was simply defined based on shared boundaries. Other options can be considered [62]. For example, considering the average distances between neighbors for each region might work well to explain spatial autocorrelation for the random forest model.

Finally, in this case study the data are analysed at SA2 level of aggregation. Care must therefore be taken in making inferences at other level of aggregation or about individuals due to biases such as Simpson's paradox [63] and the modifiable areal unit problem [64].

Conclusion

In this study, the performance of different statistical machine learning algorithms and their corresponding predictions confirmed that it is important to take into account the spatial nature of data when fitting statistical machine learning model,

when analysing population health data in SA2 level. A new spatial random forest model was introduced and was shown to explain more of the spatial variation and provide a better model fit in the case study of health vulnerabilities among children in Queensland. In this case study, attendance at preschool was found to have the highest percentage of importance for vulnerability in the physical health and wellbeing domain and the emotional maturity domain.

Acknowledgements
The authors thank the Children's Health Queensland staff for their support during the research.

Funding
This research was supported by Children's Health Queensland (CHQ) and (QUT) Center for Data Science, Queensland, Australia.

Availability of data and materials
All the data used in this study are available to the public from the Australian Bureau of Statistic and Australian Early Development Census.

Declaration
"This paper uses data from the Australian Early Development Census (AEDC). The AEDC is funded by the Australian Government Department of Education, Skills and Employment. The findings and views reported are those of the author and should not be attributed to the Department or the Australian Government."

Competing interests
All authors have read and approved this version of the article, and declared that they have no competing financial or non-financial interests to disclose.

Consent for publication
Not applicable.

Authors' contributions
Study design and setting: Arnett, Areed, Price and Mengersen. Data analysis and interpretation: Areed, Price, Mengersen and Arnett. Manuscript drafting: Areed. Critical revision of the manuscript: Areed, Price, Arnett and Mengersen.

Author details
[1]School of Mathematical Science, Center for Data Science, Queensland University of Technology, Queensland, Australia. [2]Children's Health Queensland , Queensland, Australia.

References

1. Skinner, D., Franz, B., Taylor, M., Shaw, C., Kelleher, K.: How us children's hospitals define population health: a qualitative, interview-based study. BMC Health Services Research **18**(1), 1–10 (2018)
2. Kindig, D., Stoddart, G.: What is population health? American Journal of Public Health **93**(3), 380–383 (2003)
3. McGinnis, M., Williams-Russo, P., Knickman, J.R.: The case for more active policy attention to health promotion. Health Affairs **21**(2), 78–93 (2002)
4. Allison, M., Attisha, E., et al.: The link between school attendance and good health. Pediatrics **143**(2) (2019)
5. Wang, Y., Li, J., Gu, J., Zhou, Z., Wang, Z.: Artificial neural networks for infectious diarrhea prediction using meteorological factors in Shanghai (China). Applied Soft Computing **35**, 280–290 (2015)
6. Goldfeld, S., Sayers, M., Brinkman, S., Silburn, S., Oberklaid, F.: The process and policy challenges of adapting and implementing the early development instrument in Australia. Early Education and Development **20**(6), 978–991 (2009)
7. Lo, C.: Population estimation using geographically weighted regression. GIScience & Remote Sensing **45**(2), 131–148 (2008)
8. Hengl, T., Nussbaum, M., Wright, M., Heuvelink, G., Gräler, B.: Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables, PeerJ, 6, 5518 (2018)
9. Fotheringham, S., Yang, W., Kang, W.: Multiscale geographically weighted regression (MGWR). Annals of the American Association of Geographers **107**(6), 1247–1265 (2017)
10. Du, Z., Wang, Z., Wu, S., Zhang, F., Liu, R.: Geographically neural network weighted regression for the accurate estimation of spatial non-stationarity. International Journal of Geographical Information Science **34**(7), 1353–1377 (2020)
11. Bailey, T., Gatrell, A.: Interactive spatial data analysis: Longman scientific and technical essex. Geographical Information System (1995)
12. Behrens, T., Schmidt, K., Viscarra Rossel, R., Gries, P., Scholten, T., MacMillan, R.: Spatial modelling with Euclidean distance fields and machine learning. European Journal of Soil Science **69**(5), 757–770 (2018)
13. Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., Kalogirou, S.: Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto International, 1–16 (2019)
14. Gaspard, G., Kim, D., Chun, Y.: Residual spatial autocorrelation in macroecological and biogeographical modeling: a review. Journal of Ecology and Environment **43**(1), 1–11 (2019)
15. Parkin, A.: The states, federalism and political science: a fifty-year appraisal. Australian Journal of Public Administration **62**(2), 101–112 (2003)
16. Sultana, S., Baumgartner, J., Dominiak, B., Royer, J., Beaumont, L.: Potential impacts of climate change on habitat suitability for the Queensland fruit fly. Scientific Reports **7**(1), 1–10 (2017)
17. Clark, P., Stuart, K., Leggett, B., Crawford, D., Boyd, P., Fawcett, J., Whiteman, D., Baade, P.: Remoteness, race and social disadvantage: disparities in hepatocellular carcinoma incidence and survival in Queensland, Australia. Liver International **35**(12), 2584–2594 (2015)
18. Moran, P.: The interpretation of statistical maps. Journal of the Royal Statistical Society: Series B (Methodological) **10**(2), 243–251 (1948)
19. Goovaerts, P., *et al.*: Geostatistics for Natural Resources Evaluation. Oxford University Press on Demand, ??? (1997)

20. Kalogirou, S., Hatzichristos, T.: A spatial modelling framework for income estimation. Spatial Economic Analysis **2**(3), 297–316 (2007)
21. Tango, T.: A test for spatial disease clustering adjusted for multiple testing. Statistics in medicine **19**(2), 191–204 (2000)
22. Song, C., Kulldorff, M.: Tango's maximized excess events test with different weights. International Journal of Health Geographics **4**(1), 1–7 (2005)
23. Tango, T.: A class of tests for detecting 'general' and 'focused' clustering of rare diseases. Statistics in Medicine **14**(21-22), 2323–2334 (1995)
24. Song, Y., Ying, L.: Decision tree methods: applications for classification and prediction. Shanghai Archives of Psychiatry **27**(2), 130 (2015)
25. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
26. Panov, P., Džeroski, S.: Combining bagging and random subspaces to create better ensembles. In: International Symposium on Intelligent Data Analysis, pp. 118–129 (2007). Springer
27. Arfiani, A., Rustam, Z.: Ovarian cancer data classification using bagging and random forest. In: AIP Conference Proceedings, vol. 2168, p. 20046 (2019). AIP Publishing LLC
28. Martínez-Muñoz, G., Suárez, A.: Out-of-bag estimation of the optimal sample size in bagging. Pattern Recognition **43**(1), 143–152 (2010)
29. Zahedi, P., Parvandeh, S., Asgharpour, A., McLaury, B., Shirazi, S., McKinney, B.: Random forest regression prediction of solid particle erosion in elbows. Powder Technology **338**, 983–992 (2018)
30. Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A., Pereira, J.M.: Modeling spatial patterns of fire occurrence in mediterranean Europe using multiple regression and random forest. Forest Ecology and Management **275**, 117–129 (2012)
31. Albert, J., Aliu, E., Anderhub, H., Antoranz, P., Armada, A., Asensio, M., Baixeras, C., Barrio, J., Bartko, H., Bastieri, D., *et al.*: Implementation of the random forest method for the imaging atmospheric cherenkov telescope magic. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **588**(3), 424–432 (2008)
32. Pal, M.: Random forest classifier for remote sensing classification. International Journal of Remote Sensing **26**(1), 217–222 (2005)
33. Ao, Y., Li, H., Zhu, L., Ali, S., Yang, Z.: The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. Journal of Petroleum Science and Engineering **174**, 776–789 (2019)
34. Li, J., Heap, A., Potter, A., Daniell, J.: Application of machine learning methods to spatial interpolation of environmental variables. Environmental Modelling & Software **26**(12), 1647–1659 (2011)
35. Chiles, J., Delfiner, P.: Geostatistics: Modeling Spatial Uncertainty vol. 497. John Wiley & Sons, ??? (2009)
36. Fotheringham, S., Crespo, R., Yao, J.: Geographical and temporal weighted regression (GTWR). Geographical Analysis **47**(4), 431–452 (2015)
37. Brunsdon, C., Fotheringham, S., Charlton, M.: Geographically weighted regression. Journal of the Royal Statistical Society: Series D (The Statistician) **47**(3), 431–443 (1998)
38. Kalogirou, S.: Destination choice of Athenians: an application of geographically weighted versions of standard and zero inflated poisson spatial interaction models. Geographical Analysis **48**(2), 191–230 (2016)
39. Sharma, H., Park, J., Mahajan, D., Amaro, E., Kim, J.K., Shao, C., Mishra, A., Esmaeilzadeh, H.: From high-level deep neural models to FPGAs. In: 2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), pp. 1–12 (2016). IEEE
40. Hagenauer, J., Helbich, M.: A geographically weighted artificial neural network. International Journal of Geographical Information Science, 1–21 (2021)
41. Bengio, Y.: Practical recommendations for gradient-based training of deep architectures. Neural networks: Tricks of the trade, 437–478 (2012)
42. Garson, D.: Interpreting neural network connection weights. Computer Science (1991)
43. Olden, J., Jackson, D.: Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecological modelling **154**(1-2), 135–150 (2002)
44. López-Moreno, J., Nogués-Bravo, D.: A generalized additive model for the spatial distribution of snowpack in the spanish pyrenees. Hydrological Processes: An International Journal **19**(16), 3167–3176 (2005)
45. Whittle, P.: On stationary processes in the plane. Biometrika, 434–449 (1954)
46. Burden, S., Cressie, N., Steel, D.: The SAR model for very large datasets: a reduced rank approach. Econometrics **3**(2), 317–338 (2015)
47. Kazar, B., Celik, M.: Spatial autoregression (sar) model: parameter estimation techniques (2012)
48. Huang, J.: The autoregressive moving average model for spatial analysis. Australian Journal of Statistics **26**(2), 169–178 (1984)
49. Besag, J., York, J., Mollié, A.: Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics **43**(1), 1–20 (1991)
50. Leroux, B., Lei, X., Breslow, N.: Estimation of disease rates in small areas: a new mixed model for spatial dependence. Statistical Models in Epidemiology, the Environment, and Clinical Trials, 179–191 (2000)
51. Rampaso, R., de Souza, A., Flores, E.: Bayesian analysis of spatial data using different variance and neighbourhood structures. Journal of Statistical Computation and Simulation **86**(3), 535–552 (2016)
52. Ishwaran, H.: The effect of splitting on random forests. Machine learning **99**(1), 75–118 (2015)
53. Bivand, R., Pebesma, E., Gomez-Rubio, V., Pebesma, E.: Applied Spatial Data Analysis with R vol. 2. Springer, ??? (2013)
54. Bivand, R.: Implementing spatial data analysis software tools in R. Geographical Analysis **38**(1), 23–40 (2006)
55. Team, C., et al.: R: A language and environment for statistical computing. The R Project for Statistical Computing (2013)
56. ColorBrewer, S., Liaw, M.: Package 'randomforest'. University of California, Berkeley: Berkeley, CA, USA (2018)

57. Wickham, H., Chang, W., Wickham, M.: Package 'ggplot2'. Create Elegant Data Visualisations Using the Grammar of Graphics. Version **2**(1), 1–189 (2016)
58. Kuhn, M.: The caret package. Journal of Statistical Software **28**(5) (2009)
59. Kalogirou, S., Georganos, S.: SpatialML, R package. Geocarto International (2019)
60. Fritsch, S., Guenther, F., Guenther, F.: Package 'neuralnet'. Training of Neural Networks. (2019)
61. Lee, D.: Carbayes: An R package for spatial areal unit modelling with conditional autoregressive priors. Journal of Statistical Software, 1–24 (2013)
62. Abedi, M.: Non-Euclidean distance measures in spatial data decision analysis: investigations for mineral potential mapping. Annals of Operations Research, 1–22 (2020)
63. Rojanaworarit, C.: Misleading epidemiological and statistical evidence in the presence of simpson's paradox: An illustrative study using simulated scenarios of observational study designs. Journal of Medicine and Life **13**(1), 37 (2020)
64. Hennerdal, P., Nielsen, M.: A multiscalar approach for identifying clusters and segregation patterns that avoids the modifiable areal unit problem. Annals of the American Association of Geographers **107**(3), 555–574 (2017)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ArticleAppendix.pdf