Spatial Statistics 21 (2017) 241–261

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Introducing bootstrap methods to investigate coefficient non-stationarity in spatial regression models

Paul Harris [a,*], Chris Brunsdon [b], Binbin Lu [c], Tomoki Nakaya [d], Martin Charlton [b]

[a] *Sustainable Agricultural Sciences, Rothamsted Research, North Wyke, Okehampton, Devon, EX20 2SB, UK*
[b] *National Centre for Geocomputation, Maynooth University, Maynooth, Ireland*
[c] *School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China*
[d] *Department of Geography, Ritsumeiken University, 56-1, Tojin-kita-machi, Kita-ku Kyoto, 603-8577, Japan*

ARTICLE INFO

Article history:
Received 29 January 2017
Accepted 24 July 2017
Available online 29 July 2017

Keywords:
Geographically weighted regression
Spatial regression
Hypothesis testing
Collinearity
GWmodel

ABSTRACT

In this simulation study, parametric bootstrap methods are introduced to test for spatial non-stationarity in the coefficients of regression models. Such a test can be rather simply conducted by comparing a model such as geographically weighted regression (GWR) as an alternative to a standard linear regression, the null hypothesis. In this study however, three spatially autocorrelated regressions are also used as null hypotheses: (i) a simultaneous autoregressive error model; (ii) a moving average error model; and (iii) a simultaneous autoregressive lag model. This expansion of null hypotheses, allows an investigation as to whether the spatial variation in the coefficients obtained using GWR could be attributed to some other spatial process, rather than one depicting non-stationary relationships. The new test is objectively assessed via a simulation experiment that generates data and coefficients with known multivariate spatial properties, all within the spatial setting of the oft-studied Georgia educational attainment data set. By applying the bootstrap test and associated contextual diagnostics to pre-specified, area-based, geographical processes, our study

* Corresponding author.
E-mail address: paul.harris@rothamsted.ac.uk (P. Harris).

http://dx.doi.org/10.1016/j.spasta.2017.07.006

provides a valuable steer to choosing a suitable regression model for a given spatial process.

1. Introduction

Often when fitting a regression model to spatial data, it is not clear what, if any, spatial effects should be accounted for. Should we focus solely on spatial autocorrelation effects (e.g. Anselin, 1988; Cressie, 1993) or should we focus solely on spatial heterogeneity effects with respect to data relationships (e.g. Fotheringham et al., 2002). Alternatively, should we try to capture both effects (e.g. Haas, 1996; Brunsdon et al., 1998a; Mur et al., 2008; Cho et al., 2010; Harris et al., 2010a; Kim et al., 2010), or investigate ways to link (e.g. Griffith, 2003, 2008; Murakami et al., 2017), or fuse them together (e.g. Gelfand et al., 2003; Finley, 2011), and if so, which are more important? Further, should we ignore both effects altogether, and instead focus on a non-spatial model that is additionally calibrated with key spatial predictor variables, such as the sample coordinates (e.g. Beale et al., 2010)? Further still, should we consider that we are missing vital predictors and that any observed spatial effects are attributable to this omission (e.g. Cressie and Chan, 1989)—and as such, focus our attention on capturing these (likely elusive) missing variables? Unfortunately, such questions are almost always difficult to answer with any objectively, and can involve problematic analytical impasses and confounders (e.g. Anselin, 1990). For example, how to identify first- from second-order effects (e.g. Armstrong, 1984), where relationship heterogeneity is commonly modelled as the former, whist autocorrelation is modelled as the latter effect? These issues are particularly pertinent for spatial data sets, as their collection are rarely part of a statistically-designed experiment—that by definition should negate confounders.

Given such issues, it is commonplace to ignore them, and instead a regression for spatial data is often chosen following a rather subjective exploratory analysis that is itself pre-defined according to the given research hypothesis and/or sometimes biased towards the particular statistical expertise of the analyst. Thus, our study aim is to provide objectivity to a particular aspect of this model selection process, where we introduce parametric bootstrap methods to test for spatial non-stationarity in the coefficients of regression models. The tests are general and can be used to compare any spatially-varying coefficient (SVC) regression as an alternative to any set of constant coefficient regressions (with or without spatial autocorrelation effects). As demonstration, we compare geographically weighted regression (GWR) (Brunsdon et al., 1996, 1998b) as an alternative to the following four null hypotheses: (i) a multiple linear regression model (MLR), (ii) a simultaneous autoregressive error model (ERR); (iii) a moving average error model (SMA); and (iv) a simultaneous autoregressive lag model (LAG). This set of null hypotheses, allows an investigation as to whether the spatial variation in the coefficients obtained using GWR could be attributed to some other spatial process (in this case, some autocorrelation effect), rather than one depicting non-stationary relationships.

To achieve this, a bootstrapping methodology (Efron, 1979, 1981, 1982) is proposed that assesses the variability of the local coefficient estimates found from GWR under the model assumptions for each of the four null hypotheses (i.e. the MLR, ERR, SMA and LAG models). The observed values of coefficient variability are then compared against these as reference distributions. Our bootstrapping methodology complements the bootstrap methods to test for zero coefficients in a mixed GWR model (Mei et al., 2006) and constant coefficients in a basic GWR model in order to specify a mixed GWR model (Mei et al., 2016). Neither studies however, compare GWR with alternative (spatially-autocorrelated) regressions, as we do here. Our paper is structured as follows. Firstly, the study regressions are formally stated; the concept of bootstrapping is reviewed; and our spatial application of bootstrapping is outlined. Secondly, the described methodology is objectively assessed via a simulation experiment based on cokriging (Matheron, 1970) that generates data and regression coefficients, each with known multivariate spatial properties, and all within the spatial setting of the Georgia educational attainment data set (Fotheringham et al., 2002; Griffith, 2008). We complement and contextualise the bootstrap results with associated diagnostics. Thirdly, we discuss and conclude this research.

2. Methods

2.1. Study regression models

For the case where there are several predictor variables $x_{n1}, x_{n2}, \ldots, x_{nm}$ and observations indexed by $i = 1, \ldots, n$, MLR has this form for response variable y_i:

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \varepsilon_i, \tag{1}$$

where the coefficients β_j, are commonly estimated by ordinary least squares (OLS). MLR only models stationary relationships between the response and predictor variables. Where these relationships are expected to change across space, MLR can be adapted to form the GWR model as follows:

$$y_i = \beta_0 (u_i, v_i) + \sum_{j=1}^{m} \beta_j (u_i, v_i) x_{ij} + \varepsilon_i, \tag{2}$$

where (u_i, v_i) is the spatial location of the ith observation and $\beta_j (u_i, v_i)$ is a realisation of the continuous function $\beta_j (u, v)$ at point i. As with (OLS) MLR, the ε_i's in GWR are random error terms which are independently normally distributed with zero mean and common variance σ^2. For GWR, a local regression is calibrated at any location i with observations near to i given more influence than observations further away by weighting them according to some kernel weighting function.

In addition to GWR, there are a number of spatial models in which the y-variable or the error term exhibits spatial autocorrelation, although the regression coefficients remain fixed over space (Anselin, 1988; Schabenberger and Gotway, 2005). Among these models is the ERR model:

$$\left.\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \gamma_i \\ \text{where} \quad \gamma_i &= \lambda \sum_{j=1}^{n} c_{ij} \gamma_j + \varepsilon_i \end{aligned}\right\}, \tag{3}$$

where c_{ij} is the ijth element of a row-normalised connectivity matrix. The parameter λ controls the degree of autocorrelation in the error term γ_i. Alternatively, the correlation between the γ_i's could be confined to near neighbours as defined by the connectivity matrix, as in the SMA model:

$$\left.\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \gamma_i \\ \text{where} \quad \gamma_i &= \lambda \sum_{j=1}^{n} c_{ij} \varepsilon_j + \varepsilon_i \end{aligned}\right\}. \tag{4}$$

As before, λ governs the degree of spatial association. A further alternative is the LAG model:

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \lambda \sum_{j=1}^{n} c_{ij} y_j + \varepsilon_i. \tag{5}$$

In this case, each y_i depends on the neighbouring y-values directly through the connectivity matrix and λ. Although λ plays a qualitatively different role than in the previous models (since it directly connects the predictor variable rather than the error terms), it still governs the degree of autocorrelation.

2.2. The parametric bootstrap

The parametric bootstrap is a statistical technique for estimating characteristics of the sampling distribution of a wide range of test statistics when the underlying data distribution is known. As

an introduction, the univariate case is explained. Suppose we have a set of independent, identically distributed observations $\{x_1, x_2, \ldots, x_n\}$, which can be written as a column vector $\mathbf{x}$. A general test statistic (such as a median, or interquartile range) is a function of these data, and may be written as $S(\mathbf{x})$. Knowing the probability distribution of the individual x_i's sometimes makes it possible to derive the distribution of S theoretically. For example, when S is the t-statistic and the x_i's are normally distributed. However, in general it is not possible to adopt a theoretical approach. When theory cannot help, one way forward is to use simulation. If we knew the distribution of the x_i's, we could simulate a large number of samples $\left\{x_1^*, x_2^*, \ldots, x_n^*\right\}$, where the starred variates are generated randomly, and for each sample compute $S^* = S(\mathbf{x}^*)$. Repeating this for R random samples gives a set $\left\{S_1^*, S_2^*, \ldots, S_R^*\right\}$, and finding descriptive statistics for this set provides approximate information about the sampling distribution of S. For example, the standard error of S may be estimated by the sample standard deviation of $\left\{S_1^*, S_2^*, \ldots, S_R^*\right\}$:

$$\mathrm{SE}_{\mathrm{bs}} = \frac{1}{R} \sum_{i=1 \ldots R} \left(S_i^* - \hat{S}^*\right)^2 \tag{6}$$

where $\hat{S}^*$ is the sample mean of $\left\{S_1^*, S_2^*, \ldots, S_R^*\right\}$. Now, suppose the data $\mathbf{x}$ are distributed with a probability distribution function $f(\mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters for f, then $\boldsymbol{\theta}$ is estimated from the sample using, say, maximum likelihood (ML), giving an estimate $\hat{\boldsymbol{\theta}}$ and then a number of 'pseudo-samples' are drawn by simulating random $\mathbf{x}^*$'s from $f\left(\mathbf{x}, \hat{\boldsymbol{\theta}}\right)$. From these we build up a series of R simulated S^* values, as set out above. The idea is that as n increases, the simulation distribution $f\left(\mathbf{x}, \hat{\boldsymbol{\theta}}\right)$ gets closer to the 'true' distribution of the x_i's since $\hat{\boldsymbol{\theta}}$ gets closer to $\boldsymbol{\theta}$ and so, asymptotically, the approximation converges to the true distribution. This implies that there are two kinds of error in a bootstrap analysis—firstly we approximate $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\theta}}$, and secondly, we approximate the 'true' sampling distribution of S with a large number (R) of simulated values. The first kind of error is reduced by increasing n, the second by increasing R. Typically, it is easier to increase R (i.e. by running more simulations) than n. However, given the constraints on accuracy that a finite n imposes, there is often a 'diminishing returns' effect when increasing R. For most empirical studies, a value of $R = 999$ is recommended. However, given the computational burden that this study's simulation experiment imposes, and given that two competing bootstrap tests are evaluated, a value of $R = 99$ is used here.

Bootstrapping and regression

In regression, the situation is similar. Assume a model such as (1) holds. The x_{ij}'s are not random as the random variables are the e_i's and by implication the y_i's, as they are a function of these. An MLR therefore considers the statistical distribution of y_i conditional on the values of the x_{ij}'s. Each y_i is independently normally distributed with mean $\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}$ and standard deviation σ. The approach to the bootstrap here, is then to fix the x_{ij} values and simulate the y_i's in each sample by generating bootstrap y_i^* values, based on ML estimates of the coefficients $\left\{\hat{\beta}_j\right\}$ and the standard deviation σ. Although the underlying bootstrap principal here is the same as above, the multivariate and conditional dependence nature of MLR modelling implies two notable practical differences: (i) in a univariate bootstrap the entire data set is simulated, whereas the bootstrap sample of a MLR data set is only simulated in the y_i values—the x_{ij}'s are fixed at the actual sample values, so the bootstrap data set is $\left\{x_{ij}, y_i^*\right\}$; and (ii) in a univariate bootstrap, each x_i is identically distributed, so ordering does not matter, whereas in MLR, each y_i has a different distribution, depending on x_{ij} so if $i_1 \neq i_2$ then y_{i1}^* is independent of y_{i2}^*, but not necessarily identically distributed. As before, provided the distributional parameters are estimated consistently (using ML, say), then asymptotically the bootstrap distribution will approach the true distribution (Efron and Tibshirani, 1986).

Up until now, we have assumed that a given MLR model holds—and is fully specified except for the values of the β_i's and the variance of the y_i's, and also that the motivation for computing $S(\mathbf{X}, \mathbf{y})$ is to estimate some descriptive statistic related to this model (where $\mathbf{X}$ is a matrix composed of the predictor variables and $\mathbf{y}$ is the vector of response variables). However, $S(\mathbf{X}, \mathbf{y})$ could also be a *test statistic* to assess the hypothesis that the model holds. Thus, the data model used is treated as a *null hypothesis*—and the sampling distribution of the test statistic is found by bootstrap simulation.

Following Davison and Hinkley (1997), it is then possible to compute bootstrap-based p-values. If R values of S^* were generated, under the null hypothesis the actual value S would be a further sample from the same distribution. This can then be added to the bootstrap sample. For example, for a one-tailed test, the significance of S is $\Pr\{S \geq S^*\}$, which is just:

$$p = \frac{\#(S^* \geq S)}{R+1}. \tag{7}$$

Bootstrapping the ERR, SMA and LAG models

Bootstrapping the MLR model in (1) has been described. For bootstrapping the ERR model in (3), the residuals (the difference between the predicted and observed y_i values) correspond to the estimated γ_i values—in vector form $\hat{\boldsymbol{\gamma}}_i$, where the individual elements of this vector are not independent. However, if λ is known (or at least estimated reliably by a ML estimate, $\hat{\lambda}$), we can express $\boldsymbol{\gamma}$ in terms of $\boldsymbol{\varepsilon}$— a random vector whose elements are independent:

$$\boldsymbol{\gamma} = \left(\mathbf{I} - \hat{\lambda}\mathbf{C}\right)^{-1}\boldsymbol{\varepsilon} \tag{8}$$

assuming $\mathbf{I} - \hat{\lambda}\mathbf{C}$ is invertible, where $\mathbf{C}$ is the connectivity matrix. The mean of each ε_i is zero, and if an estimate of the standard deviation of the ε_i's, $\hat{\sigma}$ can be obtained, one can create bootstrap samples $\boldsymbol{\varepsilon}^*$ and from these, create simulated residual samples $\boldsymbol{\gamma}^* = \left(\mathbf{I} - \hat{\lambda}\mathbf{C}\right)^{-1}\boldsymbol{\varepsilon}^*$. From these, simulated $\mathbf{y}$ vectors, $\mathbf{y}^*$ can be created by:

$$\mathbf{y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\gamma}^* \tag{9}$$

so that bootstrap data sets $(\mathbf{X}, \mathbf{y}^*)$ can be generated. In turn, bootstrap samples of test statistics can be simulated. In the simulations, ML estimates for λ and $\boldsymbol{\beta}$, and a bias corrected estimate of σ are used. For the SMA model in (4), the assumption of independence of residuals again does not hold. However, a similar approach to the ERR model may be used. As before, assume we have an estimate of λ, then we can express $\boldsymbol{\gamma}$ in terms of $\boldsymbol{\varepsilon}$- again a random vector with independent, identically normally distributed elements:

$$\boldsymbol{\gamma} = \left(\mathbf{I} - \hat{\lambda}\mathbf{C}\right)^{-1}\boldsymbol{\varepsilon}. \tag{10}$$

As before, $\hat{\sigma}$ can be obtained, and bootstrap samples $\boldsymbol{\varepsilon}^*$ and $\boldsymbol{\gamma}^* = \left(\mathbf{I} - \hat{\lambda}\mathbf{C}\right)^{-1}\boldsymbol{\varepsilon}^*$ are created. From these, simulated $\mathbf{y}$ vectors, $\mathbf{y}^*$ can be created by:

$$\mathbf{y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\gamma}^* \tag{11}$$

to generate $(\mathbf{X}, \mathbf{y}^*)$ and as usual, bootstrap samples of test statistics are created. As for the ERR model, we again compute ML estimates for λ and $\boldsymbol{\beta}$, and a bias corrected estimate of σ. For the LAG model in (5), the ε_i's are independently normally distributed with zero mean and common variance σ^2. This can be expressed in matrix notation as:

$$\mathbf{y} = X\boldsymbol{\beta} + \lambda\mathbf{C}\mathbf{y} + \boldsymbol{\varepsilon} \tag{12}$$

which may be re-arranged to:

$$\mathbf{y} = (\mathbf{I} - \lambda\mathbf{C})^{-1}(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \tag{13}$$

which provides an approach to the bootstrap simulation, where in this case, ML estimates of λ and $\boldsymbol{\beta}$, and a bias corrected estimate of σ are estimated, and bootstrap samples $\boldsymbol{\varepsilon}^*$ are created from these. This in turn gives bootstrap samples of $\mathbf{y}$:

$$\mathbf{y}^* = \left(\mathbf{I} - \hat{\lambda}\mathbf{C}\right)^{-1}\left(\mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}^*\right) \tag{14}$$

and as before, these may be used as a basis for bootstrap estimation of standard errors of the parameter estimates, significance tests and so on.

2.3. Bootstrap test statistics for coefficient non-stationarity

For this study, it is intended to create a test statistic for spatial non-stationarity, as characterised for example by the GWR model (2) and to use this to test against a number of null hypotheses with globally-fixed regression coefficients, such as models (1)–(5). In previous work, GWR models have been considered as alternative hypotheses to null models of type (1), but here further null models of type (3)–(5) are considered, that each have some kind of spatial structure, although the regression coefficients still do not vary spatially. Thus, a key study aim is to check whether the observed spatial variation when calibrating a GWR model could be attributed to the spatial structure contained in the autoregressive models. In this respect, a test statistic is defined with the intention of detecting non-stationarity in regression coefficients. The GWR model assumes that the vector of regression coefficients, $\boldsymbol{\beta}$, varies geographically, but that is not the case for any of the null models considered. Therefore, a test statistic measuring spatial variability of regression coefficients is proposed.

A basic test statistic

One such statistic is the standard deviation of a number of local estimates of a given regression coefficient over a number of distinct sampling points. Here, GWR is used to obtain the local coefficient estimates. If a point has coordinates, (u_k, v_k) then the GWR estimate of $\boldsymbol{\beta}(u_k, v_k)$ is given by solving:

$$\mathbf{X}^T\mathbf{W}_{(u_k,v_k)}\mathbf{X}\hat{\boldsymbol{\beta}}(u_k, v_k) = \mathbf{X}^T\mathbf{W}_{(u_k,v_k)}\mathbf{y} \tag{15}$$

where $\mathbf{W}_{(u_k,v_k)}$ is a diagonal matrix whose diagonal entries are the geographical weighting of each observation for the regression point k. In this study, we specify an adaptive (by distance) bi-square kernel function, so that the ith elements of the diagonal of $\mathbf{W}_{(u_k,v_k)}$ is:

$$w_{ik} = \left(1 - (d_{ik}/r_k)^2\right)^2 \quad \text{if } d_{ik} \le r_k \; w_{ik} = 0 \text{ otherwise,} \tag{16}$$

where d_{ik} is the distance between the location of observation i and (u_k, v_k); and r_k is a bandwidth parameter controlling the size of the local window used to calibrate $\boldsymbol{\beta}(u_k, v_k)$. In this study, r_k is chosen 'automatically' from the data using a *corrected* AIC approach (Fotheringham et al., 2002). Each coefficient $\beta_j(u_k, v_k)$ is calibrated for a number of different locations comprising the set $L = \{(u_k, v_k)\}$, and the standard deviation of these values gives a test statistic q_j for each coefficient:

$$q_j^2 = \frac{\sum_{k=1\ldots L}\left(\beta_j(u_k, v_k) - \breve{\beta}_j\right)^2}{L} \tag{17}$$

where

$$\breve{\beta}_j = \frac{\sum_{k=1\ldots L}\beta_j(u_k, v_k)}{L}. \tag{18}$$

A bootstrap approach, as outlined above, will be used to test a hypothesis of stationarity for each coefficient. Typically, L is simply the set of locations of the observations. Our approach is similar to that used in Brunsdon et al. (1996, 1998b) and Leung et al. (2000) for testing for spatial non-stationarity against a MLR null. In Brunsdon et al. (1996, 1998b) the approach is again simulation-based, but uses a randomisation test in which random permutations of the predictor and response variable list for each individual are assigned to the locations. The bootstrap approach here, differs in that it does not condition on the exact values of the variables observed in the data.

A modified test statistic

One issue when considering the estimation of coefficients in GWR is that of local collinearity amongst the predictor variables (Wheeler and Tiefelsdorf, 2005; Wheeler, 2007). For the global regression case, if there is a high degree of correlation between $\mathbf{X}$ variables, then problems calibrating the regression model can follow. In particular, if a pair of predictors are *exactly* related by a linear equation, say $x_1 = ax_2 + b$ for some a and b, then the determinant of $\mathbf{X}^T\mathbf{X}$ is zero and equations such as (15) in their global form, cannot be solved. A similar problem occurs if any of the predictors are

constant, and there is an intercept term. In practice, such *exact* linear relationships rarely occur, but situations in which they very nearly occur are more frequent. In this case, the above equation may be solved, but in the words of Farrar and Glauber (1967) '*the elements of* $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ *explode*'. This issue can be particularly important in GWR, since a near-linear relationship between predictors need only hold in a particular geographical region, rather than in the data set as a whole, to cause such problems.

In Eq. (17), the q_j test statistic for non-stationarity is based on the standard deviation of local coefficient estimates. If some individual local coefficients behave erratically due to say, local collinearity, they could affect the reliability of the test statistic. Thus, we modify q_j by working with standard deviations of the local *normalised* coefficient estimates, which down-weights the effects of any outlying coefficient estimates. That is, estimates of $\beta_i(u_k, v_k)$ divided by their estimated standard errors are used instead of just the estimates of $\beta_i(u_k, v_k)$, alone. Call this modified statistic $q_j^{\#}$, then:

$$\left(q_j^{\#}\right)^2 = \frac{\sum_{k=1\ldots L}\left(\beta_j^{\#}(u_k, v_k) - \breve{\beta}_j^{\#}\right)^2}{L} \tag{19}$$

where

$$\beta_j^{\#}(u_k, v_k) = \frac{\beta_j(u_k, v_k)}{\mathrm{SE}\left(\beta_j(u_k, v_k)\right)} \tag{20}$$

and

$$\breve{\beta}_j^{\#} = \frac{\sum_{k=1\ldots L}\beta_j^{\#}(u_k, v_k)}{L} \tag{21}$$

where $q_j^{\#}$ is assumed more robust to the effects of local collinearity than the basic test statistic q_j. Observe that local coefficients and local standard errors can vary over space not only for the effects of local collinearity, but for other reasons as well. For example, if we use a fixed bandwidth for the geographical weighting function, the local estimates around the edge of the geographical extent under study are obtained from a relatively small number of samples compared to those located in the centre of the study area, and this can cause unusually high local standard errors. Outlying observations can have a similar effect on the local estimates, where such observations need only be locally-outlying (Harris et al., 2010b). For these reasons, the use of $q_j^{\#}$ is preferred to q_j, in general.

Remarks and interpretation

Thus, each of the four fixed coefficient regressions are calibrated in the usual way and associated bootstrap data sets are generated (where only the response varies, whilst the predictors are always fixed). For each bootstrap data set, GWR estimates of the local coefficients are generated, where the S^* values from before, will be one of q_j or $q_j^{\#}$ computed accordingly. The MLR null acts as the benchmark, whilst the other nulls are spatial. As q_j or $q_j^{\#}$ are measures of variability of the local estimates, the tests are easy to interpret. Since each null regression model is a random process, even when coefficients do not vary geographically one would not expect the local coefficient estimates to be identical in different locations. The aim of the bootstrap analysis is to determine how much variability in coefficient estimates one might expect to encounter due to the random variation in a model, and to compare the level of variability in the observed data set against this. In particular, if the null model is geographically fixed in the regression coefficients, but exhibits spatial autocorrelation in either the response variable or the error term, one might expect the degree of variation in local calibrations to be greater. For example, if one of the predictor variables takes on a higher value in a certain region, and autocorrelation in the response values gives rise to a cluster of relatively high levels in the same region, this could lead to a higher local estimate of the regression coefficient associated with the predictor, if the regression window contains this region. This variability should be entirely explained by factors other than geographical variation in regression coefficients. The aim in this study is to test whether the degree of variability in local estimates exceeds the amount expected due to situations such as that described.

Table 1
Summary of the six spatial processes generated, each with their different spatial characteristics.

Spatial process no.	Intercept β_0	Coefficient β_1	Coefficient β_2	Coefficient β_3	Predictor data collinearity for x_2, x_3	Error term ε_i	Mean to error ratio (%)
SP1	Stationary	Stationary	Stationary	Stationary	Weak	Random	99.9:0.01
SP2	Stationary	Stationary	Stationary	Stationary	Strong	Random	99.9:0.01
SP3	Stationary	Stationary	Stationary	Stationary	Weak	Autocorrelated	90:10
SP4	Stationary	Stationary	Stationary	Stationary	Strong	Autocorrelated	90:10
SP5	Non-stationary	Non-stationary	Non-stationary	Non-stationary	Weak	Random	99.9:0.01
SP6	Non-stationary	Non-stationary	Non-stationary	Non-stationary	Strong	Random	99.9:0.01

2.4. The simulation experiment

The geostatistical-based simulation experiment generates realisations to exhibit one of three core scenarios: (i) stationary data relationships with random (or independent) error effects, (ii) stationary data relationships with spatially-autocorrelated (or dependent) error effects, and (iii) non-stationary data relationships with random (or independent) error effects. Intuitively, the spatial processes that result should strongly favour: MLR for scenario (i), ERR/SMA/LAG for scenario (ii), and GWR for scenario (iii). The experiment provides useful stochasticity, enabling nuanced differences to each realisation, generated from the same initial specifications. Six spatial processes are specified, two for each scenario, see Table 1.

As an overview, the experiment generates three predictors, x_1, x_2, x_3, with two levels of collinearity (weak and strong) between x_2, x_3 only; and then independently, the regression coefficients, $\beta_0, \beta_1, \beta_2, \beta_3$ are simulated. The predictor and coefficient realisations are then directly used to generate the response variable y_i, and the error data ε_i, where 90% or 99.9% of the variation in the response is explained by the mean component of the spatial process. Thus, respective ratios of 90:10 and 99.9:0.01 are specified for the mean to error components of the spatial process. The error itself, is specified as either a random or spatially-autocorrelated process. The resultant response to predictor variable correlations would range from moderately weak to moderately strong.

The mean to error ratio strongly controls the outcomes of the experiment, where as it tends to 0:100, worthwhile insights on model behaviour reduce. The chosen ratios reflect this tendency. The 99.9:0.01 ratio is used for processes generated with non-stationary relationships, as a non-stationary intercept term is also representative of the errors. This commonality leads to a difficult identification problem, which is further complicated in that a non-stationary intercept tends to reflect autocorrelated errors. With these issues in mind, the random errors generated for these particular processes are relatively small; and clearly, the consideration of autocorrelated error effects would only further complicate. The same ratio of 99.9:0.01 is also used for stationary relationship processes with random errors, whilst for stationary relationship processes with autocorrelated errors, the narrower 90:10 ratio is used.

For each of the six different processes (SP1 to SP6), 30 data realisations are generated and the bootstrapping methodology applied, together with contextual model fits and diagnostics. Thus $6 \times 30 = 180$ realisations are generated in total. The steps of the simulation algorithm are such, that many realisations will have common elements. For example, for the first realisations for processes SP1 and SP3 in Table 1, they will share the same predictor data and regression coefficients, but differ in their error and response data; and this type of commonality pervades for all 30 realisations. Realisations are generated to the centroids of the Georgia counties educational attainment data set for the United States. Area-based simulations are considered a realistic improvement to grid-based ones (e.g. Wang et al., 2008. The Georgia counties are similarly used in the GWR simulation study of Paez et al. (2011), where its sample size ($n = 159$) is considered too small to reliably use GWR. This is not viewed as a concern and instead, presents a critical challenge to the bootstrap tests. Details of the simulation experiment are given in supporting information, section S1.

Basic and modified bootstrap tests are applied (with $R = 99$), to each of the 30 data realisations, stemming from each of the six different spatial processes. The bootstrap outputs will only consist of p-values, testing each of the four null hypotheses. Hence for each test statistic, a sequence of p-value distributions is found (each of size 30) and these are summarised with boxplots, where instances of $p \leq 0.05$ (as a one-tailed test) would indicate significant coefficient non-stationarity at the 95% level. It is also useful to place in context the bootstrap results, with an existing alternative. In this respect, we find p-values from the permutation test (described before) for significant variation in each GWR coefficient, with MLR as the null hypothesis. Here $R = 99$ random permutations are specified for this test; and the resulting p-value distributions are summarised with boxplots, where instances of $p \leq 0.05$ (as also one-tailed) would indicate significant coefficient non-stationarity at the 95% level. Further contextual information is found via various regression diagnostics, for all 180 realisations, where we report: (i) spatial autocorrelation tests for the response and residual data from a MLR fit; (ii) GWR bandwidths for spatial heterogeneity effects; (iii) global design matrix condition numbers (CNs) for collinearity effects; (iv) ranked (corrected) AIC values for relative model fit; (v) coefficient estimation accuracy; and (vi) coefficient estimation confidence interval (ECI) accuracy. These diagnostics are detailed in supporting information, section S2. On a medium specs laptop (Intel® core ™ i7-4600U CPU @ 2.10 GHz to 2.70 GHz with 16.0 GB using a 64-bit OS), the full simulation experiment took 3 days, 1 h and 30 min to run.

2.5. Local patterns

Having applied the described bootstrap tests, where some or all of the regression's coefficients are considered non-stationary, it is next useful to map each local coefficient set and determine in which areas they differ significantly from zero, or some other quantity. Since we are assessing a non-stationary hypothesis against several stationary null hypotheses, a test as to whether local regression coefficients differ from the value of the hypothesised global coefficient is appropriate. This is a useful map-based test, as it visually explores the geographical consequences of incorrectly using a stationary model by mapping those places likely to have notably different regression coefficients.

For example, given a predictor variable shows strong signs of departure from coefficient stationarity, it would be useful to find the sampling distribution of its local coefficient estimate (or a test statistic based on this coefficient) in a number of locations, and measure its deviation from a global coefficient. An approach based on the bootstrap analysis of the null hypotheses here, is now proposed. Essentially, one estimates a local regression coefficient, as in Eq. (15), and then compute its standard error according to:

$$\mathrm{SE}_{h_0}\left(\hat{\boldsymbol{\beta}}\right) = \mathrm{diag}\left[\hat{\sigma}_{h_0}\left(\mathbf{X}^T\mathbf{W}_{(u_k,v_k)}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}^2_{(u_k,v_k)}\mathbf{X}\left(\mathbf{X}^T\mathbf{W}_{(u_k,v_k)}\mathbf{X}\right)^{-1}\right] \tag{22}$$

where $\mathrm{SE}_{h_0}\left(\hat{\boldsymbol{\beta}}\right)$ is a vector of coefficient standard errors based on the null hypothesis; and $\hat{\sigma}_{h_0}$ is an estimate of the error term standard deviation for the null model. The local estimate has the global estimate subtracted, and the result is then divided by the standard error from Eq. (22), to obtain a *pseudo* t-statistic at location (u_k, v_k):

$$t_{j,h_0}(u_k, v_k) = \frac{\beta_j(u_k, v_k) - \beta_{j,h_0}}{\mathrm{SE}_{h_0}\left(\beta_j(u_k, v_k)\right)} \tag{23}$$

where β_{j,h_0} denotes the *global* regression coefficient under the hypothesis h_0. As this statistic is no longer assumed to have a t-distribution, a more appropriate name might be a *local Wald statistic* after (Wald, 1943), which similarly down-weights the effects of extreme-valued local coefficients, as does the modified test statistic, $q_j^{\#}$ from before. Thus R bootstrap samples can be created, based on one of the four possible null hypotheses, and a number of local coefficient-based *pseudo* t-statistics can be computed for each bootstrap sample. Thus for each local statistic, a bootstrap p-value is found and these are mapped to identify where the local regression coefficients significantly differ from the global one.

3. Results

The bootstrap results of the simulation experiment are given in Figs. 1–4, where they can be placed in context of: (i) the simulation specifications in Table 1 and Table S1 in supporting information, and

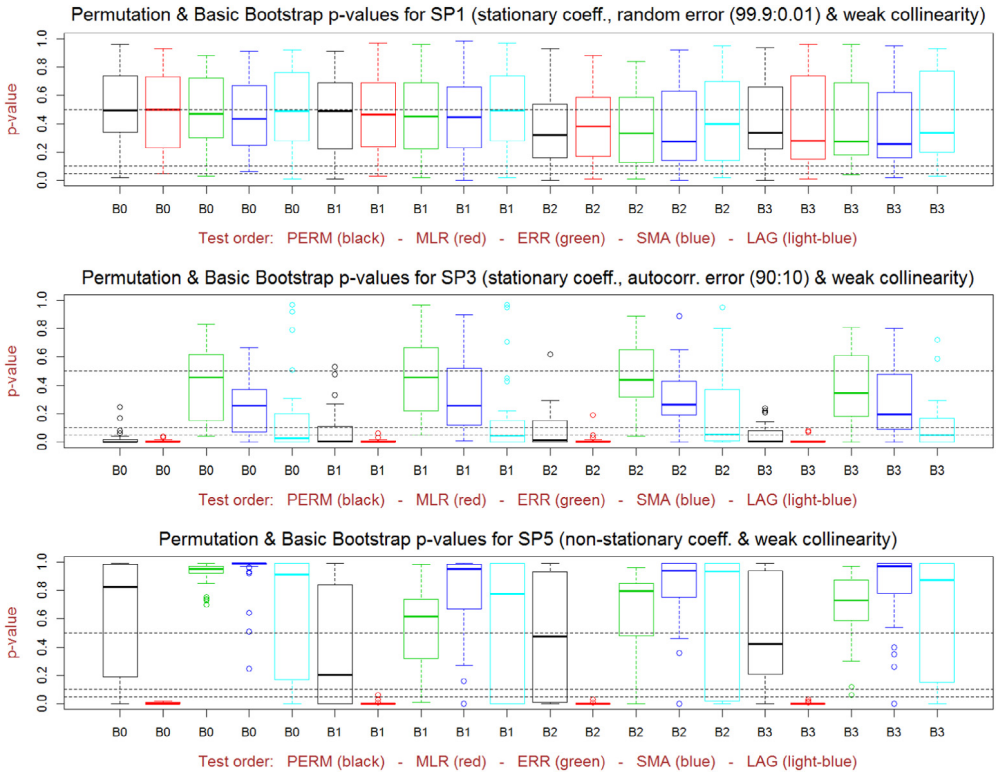


Fig. 1. All 90 realisations for three spatial processes (SP1, SP3 and SP5): boxplots of permutation test and basic bootstrap test *p*-values. Plots are shown with *p*-value thresholds at 0.05, 0.10 and 0.50.

example realisations in Figures S1 to S3 in supporting information; (ii) the model diagnostic outputs in Figures S4 to S9 in supporting information; and (iii) the example single realisation outputs in Tables 2–4 and Figs. 5 and 6, from Section 3.3. Clearly, it is quite a challenge to interpret the diversity of the inputs and outputs created.

3.1. All 180 realisations: contextual diagnostics

A full presentation of the contextual results is given in supporting information (section S3), but the following key observations are listed here, together with a critical commentary:

i. ***Autocorrelation tests***: For all six processes, significant autocorrelation is always found in the response variable. Significant autocorrelation is also found for the residual data sets from an MLR fit, but only for processes SP3 to SP6. Thus for non-stationary coefficient processes (SP5, SP6), inappropriate MLR fits will directly produce autocorrelated residuals.

ii. ***GWR bandwidths***: For random error and stationary coefficient processes (SP1, SP2), bandwidths tend to the maximum of 100%, indicating GWR to be an inappropriate model choice, as would be expected. Bandwidths with an average size of 41.7%, result for autocorrelated error and stationary coefficient processes (SP3, SP4), indicating that GWR will tend to inappropriately suggest spatial pattern in data relationships, when the response and residual data exhibit autocorrelation (as confirmed in (i)). Bandwidths with an average size of 15.9%, result for SP5 and SP6, providing the strongest evidence for choosing GWR. However, as GWR with bandwidth selection through minimum AIC criterion is known to over-fit (Jetz et al., 2005; Paez et al.,

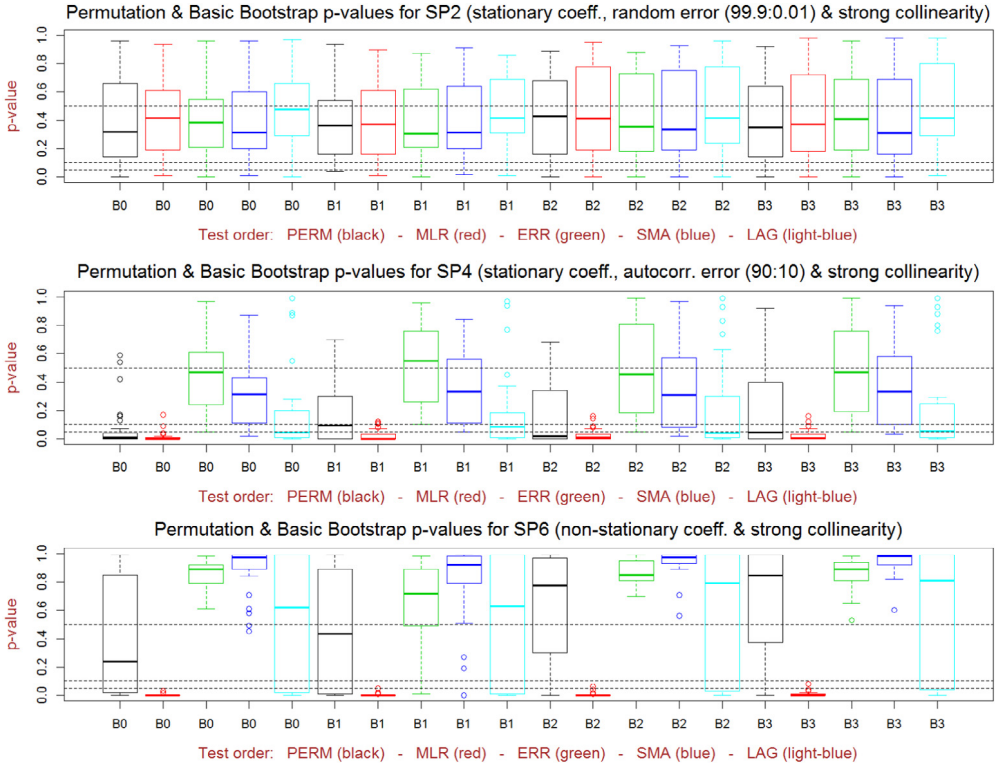


Fig. 2. All 90 realisations for three spatial processes (SP2, SP4 and SP6): boxplots of permutation test and basic bootstrap test p-values. Plots are shown with p-value thresholds at 0.05, 0.10 and 0.50.

2011), it is likely that smaller bandwidths have been found than what is ideally required; see also Loader (2004) for the local regression case.

iii. ***Ranked AIC-based model fits***: In general, the model fit via AIC is found to be a poor discriminating diagnostic across the three core process groups. Here GWR can often provide the best fit to SP3 and SP4, while ERR can on occasion provide the best fit to SP5 and SP6.

iv. ***Regression coefficient accuracy***: All models under-perform due to predictor variable collinearity, for the processes they are designed to suit (i.e. MLR for SP2; ERR/SMA/LAG for SP4; and GWR for SP6). For GWR, this re-affirms what has been widely reported (e.g. Paez et al., 2011), together with remediation strategies to counter it (Wheeler, 2007, 2009; Gollini et al., 2015). Results show however, that collinearity is a generic problem. On average, all models provide similar levels of coefficient estimation accuracy, to SP1 and SP2, which is expected as all study models will *tend* to the MLR calibration. Similarly, all models provide similar levels of coefficient estimation accuracy to SP3 and SP4. This is because stationary coefficient estimation tends to be robust to spatial effects, whereas coefficient estimation uncertainty commonly is not. GWR does not provide the best results for SP5 and SP6 as ERR and SMA do. It is not unexpected for GWR to perform relatively poorly in this respect, and results do not relay whether or not the GWR coefficient surfaces broadly reflect the actual non-stationary coefficient patterns simulated. This is the remit of GWR after all—that is the spatial exploration of relationship heterogeneity via maps.

v. ***Regression coefficient ECI accuracy***: Results are broadly as expected, with each model performing well for the process group, it is designed to suit. Again, GWR can outperform ERR for SP3 and SP4, although it is not designed to do so. For SP5 and SP6, GWR and ERR perform the best,

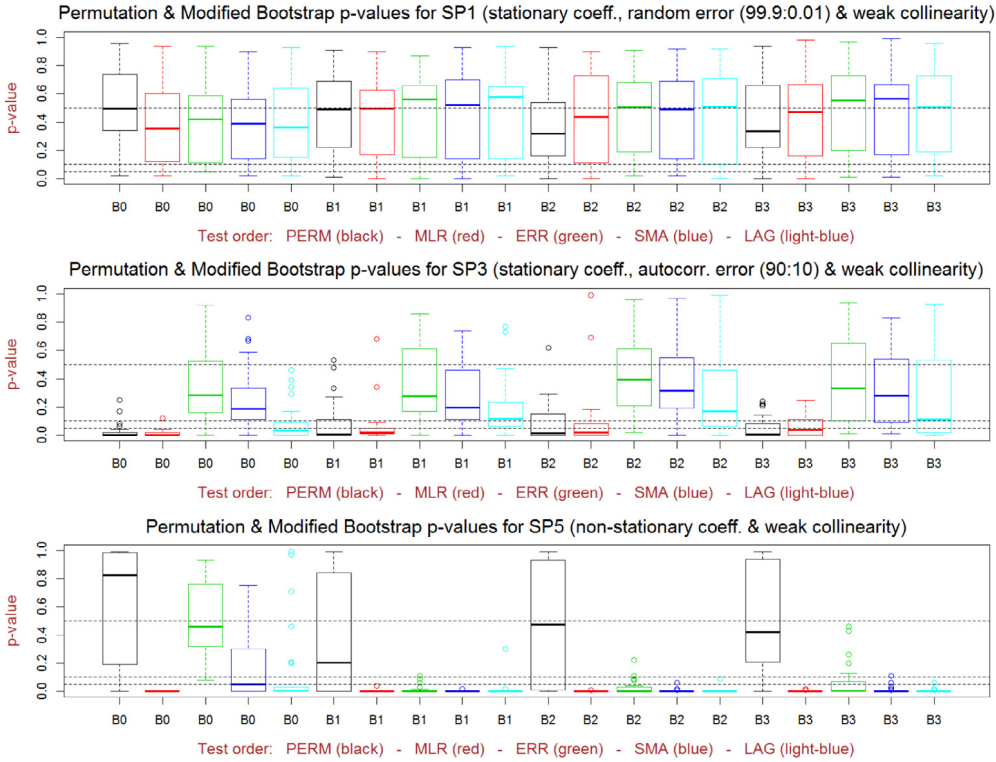


Fig. 3. All 90 realisations for three spatial processes (SP1, SP3 and SP5): boxplots of permutation test and modified bootstrap test *p*-values. Plots are shown with *p*-value thresholds at 0.05, 0.10 and 0.50.

indicating that uncertainties in their coefficient estimation are, in general, accurately accounted for in their coefficient standard errors. Thus for GWR, its relatively poor coefficient estimation accuracy performance is off-set by its relatively strong performance here. Given these results, it is evident that any GWR analysis should include an assessment of coefficient estimation uncertainty, just as that routinely done in any stationary coefficient analysis. Fotheringham and Oshan, 2016 similarly argue for such assessments, but in relation to poor GWR coefficient estimation due to local collinearity.

3.2. All 180 realisations: bootstrap test results

On viewing the *p*-value boxplots in the top panels of Figs. 1–4, for SP1 and SP2, it is clear that all tests (permutation, basic and modified bootstrap) perform as they should do. On average, there is little evidence for coefficient non-stationarity for all nulls and all response to predictor data relationships are correctly viewed as fixed across space. There is also no evidence to indicate that the level of collinearity compromises these results.

On viewing the *p*-value boxplots in the middle panels of Figs. 1–4, for SP3 and SP4, it is clear that the permutation tests rarely perform as they are expected to do. These tests consistently and erroneously indicate significant coefficient non-stationarity, for its MLR null, for all four coefficients. This is a direct reflection of GWR's tendency to incorrectly find coefficient spatial pattern to these processes, as observed above. For the bootstrap tests (basic and modified), the results are dependent on the null hypothesis, where MLR and LAG nulls rarely perform as expected. Both will tend to erroneously

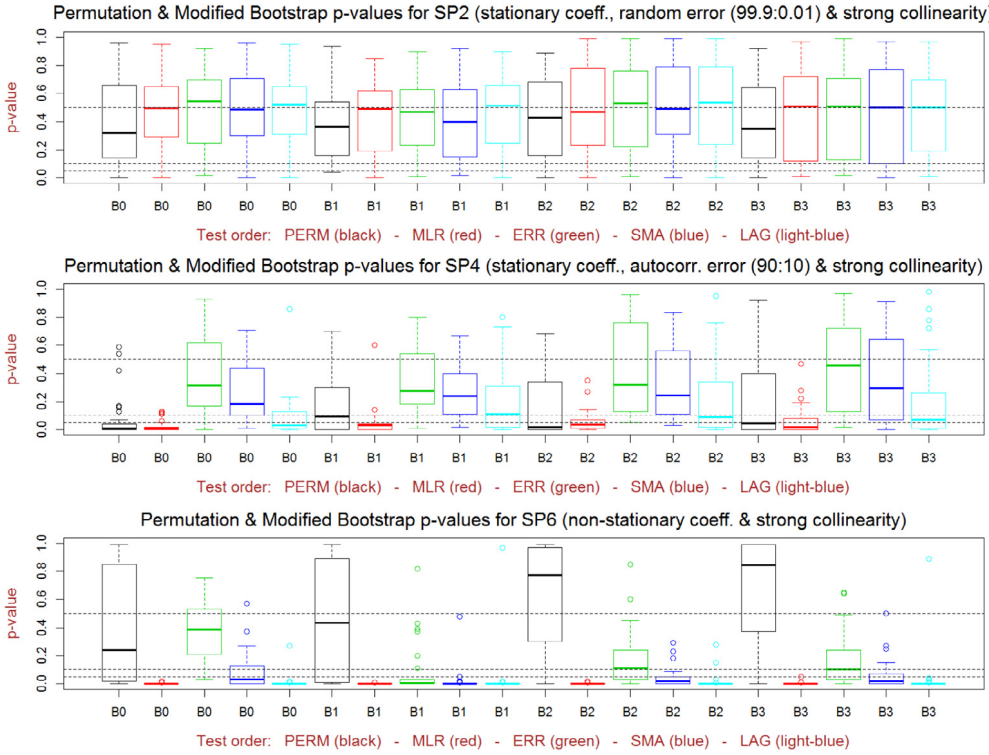
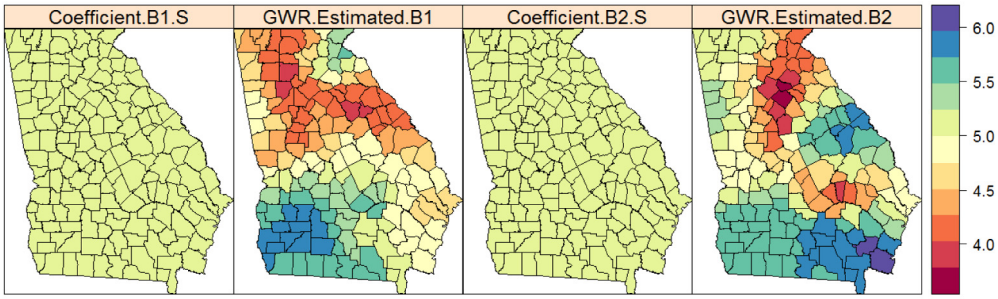


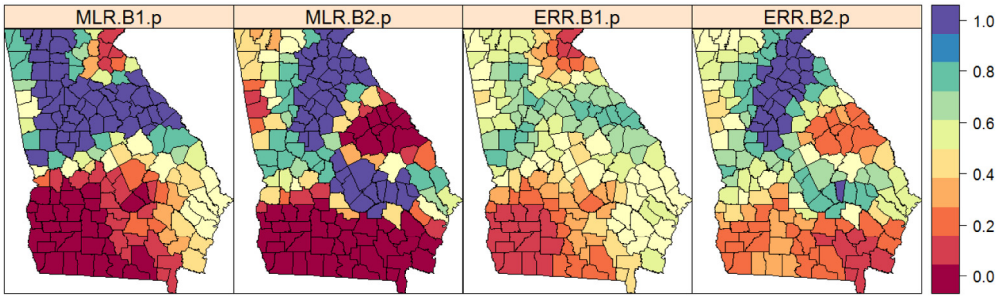
Fig. 4. All 90 realisations for three spatial processes (SP2, SP4 and SP6): boxplots of permutation test and modified bootstrap test *p*-values. Plots are shown with *p*-value thresholds at 0.05, 0.10 and 0.50.

indicate significant coefficient non-stationarity when none is present. This is again considered a direct consequence of GWR erroneously finding pattern to these processes, but now coupled with relatively small variability in the bootstrapped response variable with the MLR and LAG nulls. This has the joint effect that the (false) variability seen in the GWR coefficients is viewed as significant for these nulls. For ERR and SMA nulls, this is not the case and these models tend to provide bootstrapped response variables that have sufficiently large enough variability, for the (same erroneous) level of variability seen in the GWR coefficients to be perfectly acceptable considering a process that *is* spatially-autocorrelated with stationary relationships. Thus, ERR and SMA nulls are accepted and data relationships are correctly viewed as fixed across space. Again, there is little evidence to indicate that the level of collinearity compromises these results.

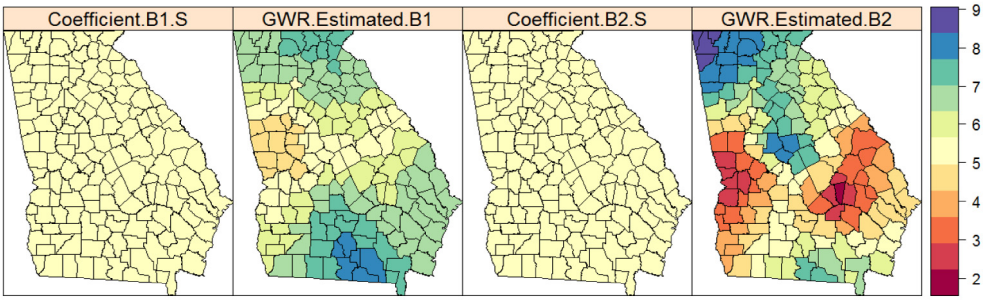
On viewing the *p*-value boxplots in the bottom panels of Figs. 1–4, for SP5 and SP6, it is clear that not all tests work as they are expected to, as they should all tend to low *p*-values indicating significant coefficient non-stationarity, for all nulls and for all coefficients. Here the permutation tests are consistently in error, possibly a consequence of the relatively low coefficient variability in the generated processes (as deliberately specified by the LMC parameters, see section S1 in supporting information). Conversely for the basic bootstrap test with the same MLR null, all tests perform as they should. However for the same basic bootstrap test with ERR, SMA or LAG nulls, all tests are again consistently in error. For the modified bootstrap tests, results clearly improve where all tests tend to perform correctly for all four (MLR, ERR, SMA and LAG) nulls, except that is for the intercept term with ERR (and possibly, SMA) nulls (i.e. the intercept is the most likely to be incorrectly viewed as stationary). The behaviour for intercept is not surprising given this term will directly reflect poor estimation in any of the three predictor coefficients, and is also likely to relate to the



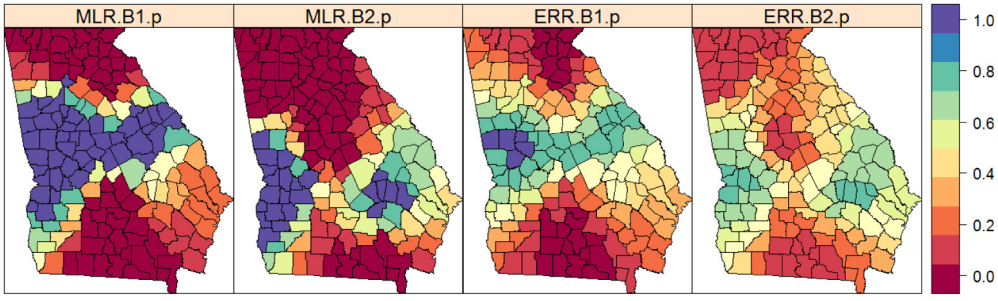
(a) Actual versus GWR Estimated Coefficients (B1 & B2) for SP3.



(b) Local p-values for B1 & B2 coefficients of MLR & ERR model null hypothesis for SP3.

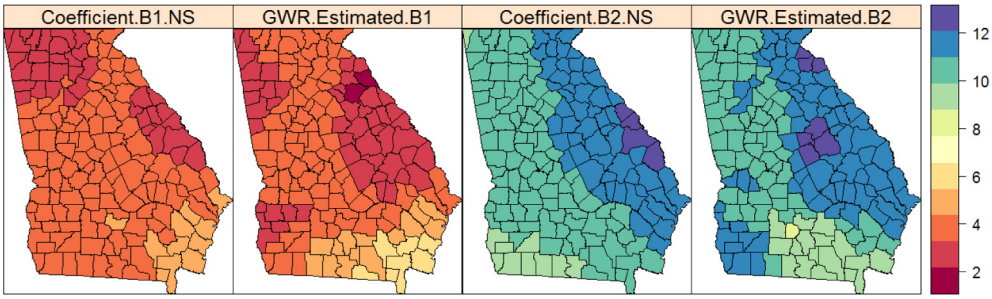


(c) Actual versus GWR Estimated Coefficients (B1 & B2) for SP4.

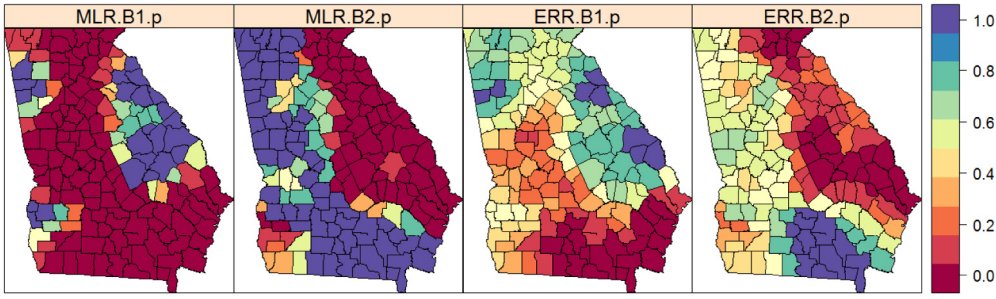


(d) Local p-values for B1 & B2 coefficients of MLR & ERR model null hypothesis for SP4.

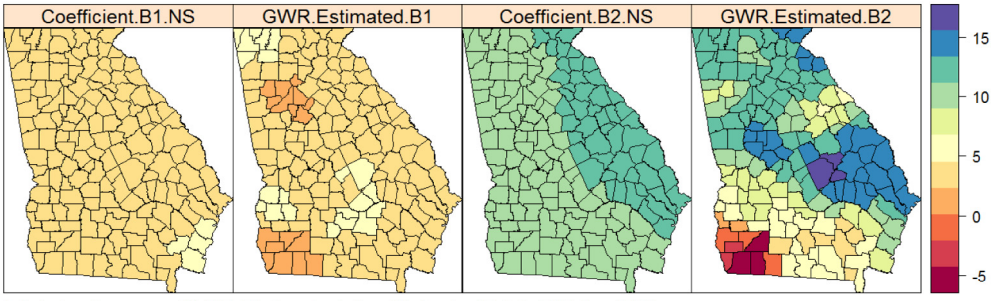
Fig. 5. Example realisations for SP3 and SP4 with (a, c) actual versus GWR estimated coefficients; and (b, d) local bootstrap p-values for MLR and ERR null model hypotheses. Results are given for β_1, β_2 only, and each using the same stationary coefficient process. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



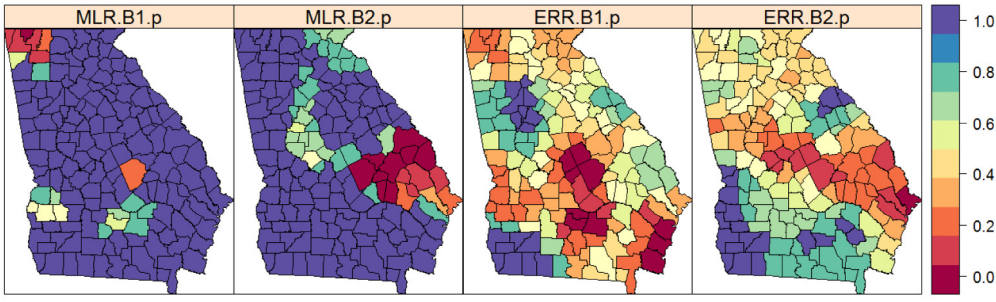
(a) Actual versus GWR Estimated Coefficients (B1 & B2) for SP5.



(b) Local p-values for B1 & B2 coefficients of MLR & ERR model null hypothesis for SP5.



(c) Actual versus GWR Estimated Coefficients (B1 & B2) for SP6.



(d) Local p-values for B1 & B2 coefficients of MLR & ERR model null hypothesis for SP6.

Fig. 6. Example realisations for SP5 and SP6 with (a, c) actual versus GWR estimated coefficients; and (b, d) local bootstrap p-values for MLR and ERR null model hypotheses. Results are given for β_1, β_2 only, and each using the same non-stationary coefficient process. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Results for modified bootstrap tests, for a single realisation—one from SP3 to SP6. Actual SD is the standard deviation of the GWR coefficients divided by their standard errors, resulting from the GWR fit to the single realisation. Both SP3 and SP4 use the same coefficient process, with actual correlations between predictors x_2 and x_3 of $r = -0.24$ for weak collinearity and $r = 0.94$ for strong collinearity. Both SP5 and SP6 use the same coefficient process, with actual correlations between predictors x_2 and x_3 of $r = 0.15$ for weak collinearity and $r = 0.94$ for strong collinearity.

	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
	Weak collinearity (SP3)				Strong collinearity (SP4)			
Modified bootstrap test								
Actual SD	1.955	2.118	2.857	2.666	1.096	1.964	2.159	1.311
MLR 95%	0.861	2.956	2.726	2.328	0.974	2.103	1.073	0.993
MLR *p-value*	**0**	0.11	**0.03**	**0.03**	**0.03**	0.07	**0**	**0.02**
ERR 95%	2.216	3.066	3.419	3.872	2.055	3.097	2.638	1.999
ERR *p-value*	0.12	0.52	0.26	0.59	0.58	0.46	0.21	0.52
SMA 95%	2.082	2.802	3.338	3.917	1.632	2.993	2.314	1.911
SMA *p-value*	0.08	0.37	0.26	0.47	0.53	0.36	0.13	0.38
LAG 95%	1.264	2.962	2.618	2.838	1.066	2.813	1.418	1.147
LAG *p-value*	**0**	0.18	**0.02**	0.07	**0.04**	0.18	**0**	**0.02**
	Weak collinearity (SP5)				Strong collinearity (SP6)			
Modified bootstrap test								
Actual SD	2.866	3.525	9.103	10.688	3.236	2.719	4.909	2.777
MLR 95%	1.019	1.547	2.663	2.450	1.537	1.246	1.492	1.140
MLR *p-value*	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
ERR 95%	3.700	3.827	6.831	7.319	4.231	2.688	3.626	3.002
ERR *p-value*	0.19	0.13	**0**	**0**	0.35	**0.04**	**0**	0.13
SMA 95%	2.731	3.050	5.382	6.565	3.448	2.083	3.382	2.575
SMA *p-value*	**0.03**	**0**	**0**	**0**	0.14	**0**	**0**	**0.02**
LAG 95%	1.760	3.028	4.737	3.950	2.140	2.588	2.630	2.490
LAG *p-value*	**0**	**0**	**0**	**0**	**0**	**0.03**	**0**	**0**

Table 3
Exploratory diagnostics for a single realisation, one from SP3 to SP6: Moran's I p-values for the response and for the residual from MLR fit; GWR bandwidth and global CN from MLR fit.

Spatial process	Moran's I for response (p-value)	Moran's I for MLR residual (p-value)	GWR Bandwidth (%)	Global CN
SP3	0	0	29.56	10.00
SP4	0	0	37.89	22.48
SP5	0	0	16.98	9.45
SP6	0	0	16.22	34.45

identification problem discussed before. Collinearity levels have a marginally adverse effect on these tests, although the modified test is in part, designed to deal with such issues. Observe from before, that an inappropriate MLR fit to a non-stationary coefficient processes will directly produce autocorrelated residuals. Thus all unexpected results for the ERR, SMA and LAG nulls need to be viewed in this context.

3.3. Single realisations

In summary, the permutation test appears to have little to no value, the basic bootstrap test only has value for MLR nulls, whilst the modified bootstrap has value for all four nulls (MLR, ERR, SMA and LAG). To complete this demonstration of the bootstrap methodology, relevant realisation-specific and localised outputs are interrogated. Modified bootstrap test outputs are shown where in each case the 95% points of the bootstrap samples are computed and significance levels are found using Eq. (7) for upper single-tailed hypothesis tests. Contextual diagnostics are given in Tables 3 and 4. Spatial patterns of the simulated (actual) and estimated GWR coefficients are compared, and contextualised by the local bootstrap p-values. Presenting outputs from a single realisation can be limiting given the

Table 4
Model accuracy diagnostics for a single realisation, one from SP3 to SP6: model AIC, coefficient relRMSE, coefficient G-STAT and coefficient M-ECI-W values.

Spatial process	Statistic	MLR	ERR	SMA	LAG	GWR
SP3	AIC	545.53	523.38	527.61	545.6	**521.98**
	relRMSE	0.29	0.19	**0.16**	0.44	1.21
	G-STAT	0.46	0.76	**0.78**	0.53	0.67
	M-ECI-W	0.29	0.58	**0.57**	0.37	1.16
SP4	AIC	621.91	589.47	596.5	620.91	**578.39**
	relRMSE	1.02	**0.95**	0.97	1.10	1.73
	G-STAT	**0.65**	0.63	0.60	0.53	0.50
	M-ECI-W	**0.96**	1.37	1.21	0.97	2.02
SP5	AIC	556.63	317.54	370.38	539.01	**269.51**
	relRMSE	0.16	**0.13**	**0.13**	0.18	0.20
	G-STAT	0.41	0.49	0.32	0.29	**0.58**
	M-ECI-W	0.55	1.44	0.42	0.47	**0.72**
SP6	AIC	664.41	382.28	472.78	619.7	**318.85**
	relRMSE	0.52	**0.14**	0.22	0.62	0.64
	G-STAT	−0.12	**0.78**	0.47	0.16	0.41
	M-ECI-W	0.64	**1.20**	0.78	0.49	1.56

stochastic nature of the simulation experiment, but as will be seen, the results tend to reflect that already observed for all realisations.

In the first instance, bootstrap results for example realisations from SP3 and SP4 are presented where all coefficients should be viewed as stationary (Table 2). Here at the 95% level, evidence for coefficient non-stationarity appears with respect to MLR and (less so for) LAG null hypotheses, for both realisations, and for β_0, β_2, β_3, only. However, given there is no evidence for coefficient non-stationarity for ERR and SMA nulls, it is reasonable to assume that spatial variation in all coefficients should be (correctly) considered a consequence of autocorrelation effects rather than relationship non-stationarity. For all regression terms, the 95% point of the distribution of the test statistic tends to increase in the following model order: MLR, LAG, SMA and ERR. Thus, the degree to which one might expect local coefficients to vary when a regression with fixed coefficients holds, increases in this order. In other words, ERR is the 'best' at addressing any perceived coefficient non-stationarity by the specification of a spatial autocorrelation effect instead. Collinearity levels have not adversely influenced these results. For the same realisations, surfaces are presented for β_1, β_2 only and the local bootstrap p-values are only found for MLR and ERR nulls. It is evident how GWR can erroneously find pattern in coefficients when none exist (Fig. 5(a)) and that such behaviour can be exacerbated in the presence of collinearity (Fig. 5(c)). It is just such spatial patterns that the (unsuspecting) analyst can be most struck with, and without conducting a fuller analysis as that suggested here, can lead to entirely false interpretations. Thus, not only do the global bootstrap results rein in such false perceptions, but the local bootstrap results do so too. In Fig. 5(b) and (d), localised coefficients that significantly differ from the corresponding global coefficient of the MLR or ERR model are those with p-values > 0.95 or p-values < 0.05, say (i.e. sample units respectively coloured 'dark blue' or 'dark red'). It is evident that such areas decrease on viewing the ERR null in relation to the MLR null, resulting in only a few areas considered to have a non-stationary relationship for the given regression term—areas which incidentally are a direct reflection of the erroneous GWR fit in the first place.

In the second instance, bootstrap results for example realisations from SP5 and SP6 are presented (Table 2). Here evidence for coefficient non-stationarity is strong with respect to MLR and LAG null hypotheses, for both realisations, and for all coefficients. Although such evidence is marginally less convincing for ERR and SMA nulls (as remember, non-stationary coefficient processes will tend to have autocorrelated errors), the test results as a whole strongly (and correctly) indicate that all coefficients should indeed be viewed as non-stationary. Again for the same realisations, surfaces are presented for β_1, β_2 only, and the local bootstrap p-values are only found for MLR and ERR nulls. It is now evident how GWR can over-fit (Fig. 6(a)) and that such behaviour can be exacerbated in the presence of collinearity (Fig. 6(c)). Again, our fuller analysis helps rein in likely misconceptions

about the true nature of relationship non-stationarity, where reassuringly the results of Table 2 do not appear adversely influenced by collinearity. It is clear in Fig. 6(b) and (d), that although areas of significant coefficient non-stationarity decrease on viewing the ERR null in relation to the MLR null, they do not decrease to the extent that ERR would be favoured over GWR. Finally, note that the bootstrap results are unable to discriminate true coefficient non-stationarity from that due to collinearity, as the modified and local tests are only designed to be robust to such effects. In this respect, it is recommended that a thorough analysis of collinearity is always conducted in practise (e.g. Gollini et al., 2015).

4. Discussion and concluding remarks

This study has set out a bootstrap methodology to test for coefficient non-stationarity in a regression model against four models with globally-fixed coefficients—the most basic of these having no spatial component, the others involving spatial autocorrelation effects. In particular, we tested for coefficient non-stationarity for GWR against four nulls: MLR, ERR, SMA and LAG models. The methodology provides three test statistics: (i) a basic test statistic; (ii) a modified test statistic, that is robust to unusual local coefficients and local standard errors; and (iii) a localised test statistic that is similarly robust to that defined in (ii). The methodology was objectively assessed via a simulation experiment and was not assessed in isolation, as it was important to provide contextual analyses. We found the basic bootstrap test statistic to only have value for MLR nulls, whilst the superior, modified bootstrap test statistic was found to have value for all four study nulls—MLR, ERR, SMA and LAG.

There are however caveats to this research, especially considering that an MLR fit to a non-stationary coefficient process will directly produce autocorrelated residuals, whilst GWR will commonly find spatial pattern in the coefficients when none exists provided the response/error is autocorrelated. Little can be done to address these unwanted (identification) artefacts, where it should also be noted that ERR will on occasion appear a good model choice to a non-stationary coefficient process. Although ensuring a broad range of stationary coefficient null models are used, as done here, is one way of mitigating against these confounders. Study results are also inter-dependent on: (a) the characteristics of the simulated data; (b) the properties and assumptions of the chosen non-stationary coefficient model; and (c) the properties and assumptions of the chosen null models. These caveats and dependencies are considered tolerable given the study's core aim is to introduce the bootstrap methodology and demonstrate its potential.

The six spatial processes specified were kept as simple as possible, so not to detract from the bootstrap methodology, but to also provide a level of objectivity so that the new methodology could be meaningfully assessed (something not viable in any empirical setting, see section S4 in supporting information). Furthermore, only basic GWR has been studied using a standard calibration procedure (adaptive bi-square kernel with AIC-defined bandwidth), and the results only relate to this use, together with the use of basic null models with standard spatial weights structures. However, as the bootstrap tests are general the next logical steps would be a conduct a regression specification sensitivity analysis for their effects on the bootstrap results, whilst keeping the simulation design the same as that used here.

The simulation experiment does however have the potential to be highly involved by varying the many parameters of the LMC. This flexibility of design is key to why a second-order effects simulation approach was taken in preference to other, more deterministic approaches found in the literature (e.g. Wang et al., 2008). Thus, by extending the simulated experiment to generate a more varied set of spatial processes, say with different mean to error ratios, graded levels of coefficient non-stationarity, different scales of coefficient non-stationarity, different levels of residual autocorrelation, different levels of predictor variable autocorrelation (e.g. Hughes and Haran, 2013), different levels of collinearity, the introduction of anomalies—the full spectrum of extended GWR models could be assessed together with nulls that are similarly extended. Many early GWR models spring to mind, for example, robust and heteroskedastic GWR (Fotheringham et al., 2002).

In a similar vein, the Bayesian SVC model of Gelfand et al. (2003) could be trialled, as this is commonly viewed as a more accurate alternative to GWR (Wheeler and Calder, 2007; Finley, 2011). Interestingly, coefficient estimation with the Bayesian SVC model is directly based on a LMC, as are

the simulated coefficients generated here. Although a difficulty arises, in that (unlike this study) it is not ideal to assess a statistical model using a simulation experiment that the model itself is based on. Furthermore, as a direct result of using a LMC, the Bayesian SVC model suffers computationally and as such, is not commonly applied. In this respect, a revised eigenvector spatial filtering (ESF) based SVC model (Griffith, 2008) that provides a variant of the Bayesian SVC model (Murakami et al., 2017), holds much promise as it does not suffer computationally.

Given that basic GWR can over-fit the coefficient processes generated in this study, spatial processes may also exist when GWR is likely to under-fit, respectively giving rise to unduly high and unduly low levels of coefficient variability, which in turn would compromise the validity of the bootstrap tests. Such issues warrant further study and directly relate to bandwidth selection in GWR. Different approaches to bandwidth selection can be investigated via the simulation experiment, such as the robust cross-validation procedure of Farber and Paez (2007), the procedure of Cho et al. (2010) that provides a bandwidth which ensures GWR residuals display random variation, or the dual bandwidths of an anisotropic GWR model (Paez, 2004). Scale GWR models should also be considered. For example, mixed GWR where some coefficients are fixed whilst others vary (Brunsdon et al., 1999) or flexible bandwidth GWR (FB-GWR) where each coefficient varies at its own spatial scale (Yang et al., 2012; Lu et al., 2017). For these models, a hierarchical testing strategy could be employed where MLR, basic GWR and mixed GWR are viewed as subsets of FB-GWR—effectively the strategy used in Nakaya et al. (2005) but with the testing being achieved via a bootstrap approach. FB-GWR counters a problem with most GWR models, in that they assume the same degree of spatial smoothness for each coefficient, which is unrealistic. Here Bayesian SVC and ESF-based SVC models provide comparison to FB-GWR, as they are also not so constrained. Furthermore, mixed GWR with only the intercept term varying, provides a route to better understand the behaviour of this term, with respect to its relationship to other coefficients and to associated first- and second-order identification issues.

Alternative stationary coefficient nulls can also be considered, such conditional autoregressive nulls, in which the random part of the model is specified in terms of conditional probability distributions (Cliff and Ord, 1973). Here, advances on the bootstrap method itself may be useful, such that found in Fingleton and Legallo (2008), Lahiri (2010), Burridge and Fingleton (2010), Monchuk et al. (2011), Han and Lee (2012), Herrera et al. (2013). Finally moving forward to empirical studies, where the true value of this bootstrap method will be realised, the results from this study (and possible extensions) can help guide the selection of a given SVC model over an associated stationary coefficient null, that suits the properties of the real study data and the analytical questions being posed.

Acknowledgements

Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by the Science Foundation Ireland under the National Development Plan. Work was also supported by a UK Biotechnology and Biological Sciences Research Council grant (BBSRC BB/J004308/1). The authors gratefully acknowledge this support. and projects from the National Natural Science Foundation of China [NSFC: 41401455; NSFC: U1533102].

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.spasta.2017.07.006. This material includes: (i) details of the simulation experiment, (ii) details of the contextual model diagnostics for the simulation experiment, (iii) simulation results for the contextual model diagnostics, (iv) empirical case study and (v) code.

References

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht.
Anselin, L., 1990. Spatial dependence and spatial stucture instability in applied regression analysis. J. Reg. Sci. 30, 185–207.
Armstrong, M., 1984. Problems with universal kriging. Math. Geol. 16, 101–108.
Beale, C.M., Lennon, J.J., Yearsley, J.M., Brewer, M.J., Elston, D.A., 2010. Regression analysis of spatial data. Ecol. Lett. 13, 246–264.

Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1996. Geographically weighted regression: A method for exploring spatial nonstationarity. Geogr. Anal. 28, 281–298.
Brunsdon, C., Fotheringham, A.S., Charlton, M., 1998a. Spatial nonstationarity and autoregressive models. Environ. Plann. A 30, 957–973.
Brunsdon, C., Fotheringham, S., Charlton, M., 1998b. Geographically weighted regression-modelling spatial non-stationarity, J. Roy. Statist. Soc. Ser. D (Statist.) 47, 431–443.
Brunsdon, C., Fotheringham, A.S., Charlton, M., 1999. Some notes on parametric significance tests for geographically weighted regression. J. Reg. Sci. 39, 497–524.
Burridge, P., Fingleton, B., 2010. Bootstrap inference in spatial econometrics: the J-test. Spat. Econ. Anal. 5, 93–119.
Cho, S.-H., Lambert, D.M., Chen, Z., 2010. Geographically weighted regression bandwidth selection and spatial autocorrelation: an empirical example using Chinese agriculture data. Appl. Econ. Lett. 17, 767–772.
Cliff, A.D., Ord, J.K., 1973. Spatial Autocorrelation. Pion, London.
Cressie, N., 1993. Statistics for Spatial Data. John Wiley and Sons, New York, USA.
Cressie, N., Chan, N.H., 1989. Spatial modeling of regional variables. J. Amer. Statist. Assoc. 84, 393–401.
Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and their Application. Cambridge University Press, Cambridge, UK.
Efron, B., 1979. Bootstrap methods: Another look at the jackknife. Ann. Statist. 7, 1–26.
Efron, B., 1981. Nonparametric estimates of standard error: the jackknife, the boostrap and other methods. Biometrika 68, 589–599.
Efron, B., 1982. The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvannia.
Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statist. Sci. 1, 54–75.
Farber, S., Paez, A., 2007. A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. J. Geogr. Syst. 9, 371–396.
Farrar, D.E., Glauber, R.R., 1967. Multicollinearity in regression analysis: The problem revisited. Rev. Econ. Stat. 49, 92–107.
Fingleton, B., Legallo, J., 2008. Estimating spatial models with endogenous variables, a spatial lag and spatially dependent disturbances: finite sample properties. Pap. Reg. Sci. 87, 319–339.
Finley, A.O., 2011. Comparing spatially-varying coefficient models for analysis of ecological sata with non-stationary and anisotropic residual dependence. Ecol. Evol. 2, 143–154.
Fotheringham, A.S., Brunsdon, C., Charlton, M., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester.
Fotheringham, A.S., Oshan, T.M., 2016. Geographically weighted regression and multicollinearity: dispelling the myth. J. Geogr. Syst. 18, 303–329.
Gelfand, A.E., Kim, H.-J., Sirmans, C.F., Banerjee, S., 2003. Spatial modeling with spatially varying coefficient processes. J. Amer. Statist. Assoc. 98, 387–396.
Gollini, I., Lu, B., Charlton, M., Brunsdon, C., Harris, P., 2015. GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. J. Stat. Softw. 63, 1–50.
Griffith, D.A., 2003. Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization. Springer, Berlin.
Griffith, D.A., 2008. Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). Environ. Plann. A 40, 2751–2769.
Haas, T.C., 1996. Multivariate spatial prediction in the presence of non-linear trend and covariance nonstationarity. Environmetrics 7, 145–165.
Han, X., Lee, L., 2012. Model selection using J-test for the spatial autoregressive model vs. the matrix exponential spatial model. Reg. Sci. Urban Econ. 43, 250–271.
Harris, P., Fotheringham, A.S., Crespo, R., Charlton, M.E., 2010a. The use of geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets. Math. Geosci. 42, 657–680.
Harris, P., Fotheringham, A.S., Juggins, S., 2010b. Robust geographically weighted regression: A technique for quantifying spatial relationships between freshwater acidification critical loads and catchment attributes. Ann. Assoc. Am. Geogr. 100, 286–306.
Herrera, M., Ruiz, M., Mur, J., 2013. Detecting dependence between spatial processes. Spat. Econ. Anal. 8, 469–497.
Hughes, J., Haran, M., 2013. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models, J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (1), 139–159.
Jetz, W., Rahbek, C., Lichstein, J.W., 2005. Local and global approaches to spatial data analysis in ecology. Global Ecol. Biogeogr. 14, 97–98.
Kim, S., Cho, S.-H., Lambert, D., Roberts, R., 2010. Measuring the value of air quality: application of the spatial hedonic model. Air Qual. Atmos. Health 3, 41–51.
Lahiri, S., 2010. Resampling Methods for Dependent Data. Springer, Berlin.
Leung, Y., Mei, C.L., Zhang, W.X., 2000. Statistical tests for spatial nonstationarity based on the geographically weighted regression model. Environ. Plann. A 32, 9–32.
Loader, C., 2004. Smoothing: Local Regression Techniques. Center for Applied Statistics and Economics (CASE), Humboldt-Universit, Berlin.
Lu, B., Brunsdon, C., Charlton, M., Harris, P., 2017. Geographically weighted regression with parameter-specific distance metrics. Int. J. Geogr. Inf. Sci..

Matheron, G., 1970. La Theorie Des Variables Regionalisees et Ses Applications: Fascicule 5, Fontainebleau, Centre de Grostatistique, Paris School of Mines.
Mei, C.-L., Wang, N., Zhang, W.-X., 2006. Testing the importance of the explanatory variables in a mixed geographically weighted regression model. Environ. Plann. A 38, 587–598.
Mei, C.-L., Xu, M., Wang, N., 2016. A bootstrap test for constant coefficients in geographically weighted regression models. Int. J. Geogr. Inf. Sci. 30, 1622–1643.
Monchuk, D., Hayes, D., Miranawski, J., Lambert, D., 2011. Inference based on alternative bootstrapping methods in spatial models with an application to county income growth in the United States. J. Reg. Sci. 51, 880–896.
Mur, J., Lopez, F., Angelo, A., 2008. Symptoms of instability in models of spatial dependence. Geogr. Anal. 40, 189–211.
Murakami, D., Yoshida, T., Seya, H., Griffith, D.A., Yamagata, Y., 2017. A Moran coefficient-based mixed effects approach to investigate spatially varying relationships. Spat. Stat. 19, 68–89.
Nakaya, T., Fotheringham, A.S., Brunsdon, C., Charlton, M., 2005. Geographically weighted Poisson regression for disease association mapping. Stat. Med. 24, 2695–2717.
Paez, A., 2004. Anisotropic variance functions in geographically weighted regression models. Geogr. Anal. 36, 299–314.
Paez, A., Farber, S., Wheeler, D., 2011. A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. Environ. Plann. A 43, 2992–3010.
Schabenberger, O., Gotway, C.A., 2005. Statistical Methods for Spatial Data Analysis. Chapman and Hall/CRC, London, UK.
Wald, A., 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Amer. Math. Soc. 54, 426–482.
Wang, N., Mei, C.L., Yan, X.D., 2008. Local linear estimation of spatially varying coefficient models: an improvement on the geographically weighted regression technique. Environ. Plann. A 40, 986–1005.
Wheeler, D.C., 2007. Diagnostic tools and a remedial method for collinearity in geographically weighted regression. Environ. Plann. A 39, 2464–2481.
Wheeler, D.C., 2009. Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso. Environ. Plann. A 41, 722–742.
Wheeler, D.C., Calder, C.A., 2007. An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. J. Geogr. Syst. 9, 145–166.
Wheeler, D., Tiefelsdorf, M., 2005. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. J. Geogr. Syst. 7, 161–187.
Yang, W., Fotheringham, A.S., Harris, P., 2012. An extension of geographically weighted regression with flexible bandwidths. In: GISRUK 2012. Lancaster.

Further reading

Hurvich, C.M., Simonoff, J.S., Tsai, C-L., 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J. R. Stat. Soc. Ser. B Stat. Methodol. 60, 271–293.
Sen, A., Srivastava, M., 1997. Regression Analysis: Theory. Methods, and Applications. Springer, Berlin.