Rothamsted Repository Download

A - Papers appearing in refereed journals

The publisher's version can be accessed at:

- https://dx.doi.org/10.1111/gean.12163

The output can be accessed at: https://repository.rothamsted.ac.uk/item/84898.

A Simulation Study on Specifying a Regression Model for Spatial Data: Choosing between Autocorrelation and Heterogeneity Effects

Paul Harris 🆔

Rothamsted Research, SSGS, North Wyke, Okehampton EX20 2SB, United Kingdom of Great Britain and Northern Ireland

In this simulation study, regressions specified with autocorrelation effects are compared against those with relationship heterogeneity effects, and in doing so, provides guidance on their use. Regressions investigated are: (1) multiple linear regression, (2) a simultaneous autoregressive error model, and (3) geographically weighted regression. The first is nonspatial and acts as a control, the second accounts for stationary spatial autocorrelation via the error term, while the third captures spatial heterogeneity through the modeling of nonstationary relationships between the response and predictor variables. The geostatistical-based simulation experiment generates data and coefficients with known multivariate spatial properties, all within an area-unit spatial setting. Spatial autocorrelation and spatial heterogeneity effects are varied and accounted for. On fitting the regressions, that each have different assumptions and objectives, to very different geographical processes, valuable insights to their likely performance are uncovered. Results objectively confirm an inherent interrelationship between autocorrelation and heterogeneity, that results in an identification problem when choosing one regression over another. Given this, recommendations on the use and implementation of these spatial regressions are suggested, where knowledge of the properties of real study data and the analytical questions being posed are paramount.

Introduction

As outlined in Anselin (1988), and reviewed in Goodchild (2004); Anselin (2010), two core effects need to be considered when fitting a regression to spatial data, one of spatial autocorrelation (e.g., Cressie 1993; LeSage and Pace 2009) and one of spatial heterogeneity with respect to data relationships (e.g., Fotheringham, Brunsdon, and Charlton 2002; LeSage and Pace 2009). Furthermore, McMillen (1996, 2001); Basile et al. (2014) argue for a third effect in nonlinearity. The question then arises on which effect(s) should we focus on for a given analysis—

Correspondence: Paul Harris, Sustainable Agricultural Sciences North Wyke, Rothamsted Research, Okehampton, EX20 2SB, United Kingdom of Great Britain and Northern Ireland
e-mail: paul.harris@rothamsted.ac.uk

that of autocorrelation, heterogeneity, or nonlinearity? For example, apply a model designed to capture both autocorrelation and heterogeneity (e.g., Besag, York, and Mollié 1991; Fischer and Stumpner 2010; Mur, López, and Angelo 2010), or both autocorrelation and nonlinearity (e.g., Christensen, Diggle, and Ribeiro 2001; Opsomer, Wang, and Yang 2001; Lambert, Xu, and Florax 2014), or both heterogeneity and nonlinearity (e.g., Wang, Mei, and Yan 2008; Yu, Peterson, and Reid 2009; Shen, Mei, and Zhang 2011). Alternatively, choose a geo-additive model, whose inherent flexibility allows all three effects to be catered for simultaneously (e.g., Basile and Gress 2005; Su and Jin 2010; Montero, Mínguez, and Durbán 2012; Basile et al. 2014). Models also exist that indirectly use autocorrelation effects, but with the aim of capturing heterogeneity effects (Assunção 2003; Gelfand et al. 2003; Griffith 2003, 2008).

All three effects are subject to issues of scale, whether this concerns the effects of the scale (or support) of the observation units (e.g., Gotway and Young 2002; Zhang, Atkinson, and Goodchild 2014; Murakami and Tsutsumi 2015), or whether the modeling objective itself is to capture processes that operate across scales (e.g., Finley 2011; Harris, Dong, and Zhang 2013a; Dong and Harris 2014; Dong et al. 2015; Osland, Thorsen, and Thorsen 2016; Bivand et al. 2017; Fotheringham, Yang, and Kang forthcoming; Leong and Yue 2017; Murakami et al. 2017; Lu et al. forthcoming). Predictor variables also play a role. It is sometimes pragmatic to ignore all such effects, and instead focus on a nonspatial model that is calibrated with key spatial predictors, such as the sample coordinates. Similarly, key predictors may be missing and that any observed spatial effects are attributable to this omission (e.g., Cressie and Chan 1989; McMillen 2003)—attention is then focused on capturing these missing variables.

Unfortunately, all such model specification issues are almost always difficult to address with any objectively, and commonly involve analytical impasses and confounders (Anselin 1990, 204; Pace and LeSage 2004, 31; Mur, López, and Angelo 2008). Analytical issues are particularly pertinent for spatial data, as their collection are rarely part of a statistically designed experiment—that by definition should negate confounders (e.g., Stefanova, Smith, and Cullis 2009). For example, it is difficult to identify first- from second-order effects (Armstrong 1984), where relationship heterogeneity is commonly modeled as the former, whist autocorrelation is modeled as the latter effect. And observe here, capturing relationship heterogeneity (e.g., Casetti 1972; Brunsdon, Fotheringham, and Charlton 1996) is only one of many nonstationary decisions that can be considered (Atkinson 2001; Lloyd 2010), including that for autocorrelation itself (e.g., Anselin 1995; Sampson, Damian, and Guttorp 2001; Harris, Charlton, and Fotheringham 2010a; Fuglstad et al. 2015; Fouedjio forthcoming), and for nonlinearity (e.g., Harris, Brunsdon, and Charlton 2013b), and those that capture both relationship and autocorrelation heterogeneity (Haas 1996; Brunsdon, Fotheringham, and Charlton 1998; Geniaux and Martinetti forthcoming). Furthermore, spatial data commonly represent a single realization, whereas model assumptions are usually based on knowledge of the full multidimensional spatial distribution that can never be fully determined, as there are no replicate samples (Myers 1989). Thus, distributional inference can only be determined under some decision of stationarity. For nonstationary models, these decisions can be locally defined with associated bias-variance trade-offs (Brunsdon, Fotheringham, and Charlton 1996; Haas 1996), or globally defined to provide greater statistical coherency (Sampson, Damian, and Guttorp 2001; Gelfand et al. 2003; Griffith 2008).

Given such variety of interrelated model and data issues, it is useful to provide guidance when selecting a regression to model spatial data. It is not viable to study all issues here, so instead, the study aim is focused on the comparison of a regression specified with stationary

autocorrelation effects against that with relationship heterogeneity effects. This is achieved via a simulation experiment, where regressions are assessed for overall fit, coefficient accuracy, and outcomes from related tests. So as not to dilute this study's message (with a barrage of regression forms), only three core models are studied: (1) multiple linear regression (MLR)[1]; (2) a simultaneous autoregressive error model (SAR) (Anselin 1988); and (3) geographically weighted regression (GWR) (Brunsdon, Fotheringham, and Charlton 1996). The first regression is nonspatial and acts as a control, the second captures autocorrelation via the error term, while the third captures heterogeneity through the modeling of nonstationary relationships between the response and predictor variables.

This study provides a bridge between simulation studies such as Kissling and Carl (2007); Dormann (2007); Beale et al. (2010) for regressions with autocorrelation effects only, and simulation studies such as Finlay (2011); Páez, Farber, and Wheeler (2011); Mellin et al. (2014) for regressions with heterogeneity effects only. Studies using simulation to investigate both effects are rare, but include that of Harris et al. (2010b) from a prediction viewpoint and Geniaux and Martinetti (forthcoming) from an inference viewpoint, neither of which investigate the inherent identification problem in as much detail as that given here.[2] This study's simulation experiment generates data with known multivariate spatial properties (Pebesma 2004), where the design of the experiment itself is considered an advance, providing a template for future work. The spatial processes that are generated should strongly favor only one from MLR, SAR, or GWR as appropriate choices. This article is structured as follows. First, the regressions are formally stated and the simulation experiment is described. Second, the outcomes of the simulation experiment are reported and critiqued. Third, a set of discussions are given, together with an empirical case study illustrating the issues described and a possible solution.

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 fixed relationships between the response and predictors. Where these relationships are expected to vary across space, MLR can be adapted to form GWR 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 realization 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 distance-decay, kernel weighting function. There are also models in which the error term exhibits spatial

autocorrelation, although the regression coefficients remain fixed over space. Among these models is the SAR model:

$$\left.\begin{array}{c} y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \gamma_i \\[2mm] \gamma_i = \lambda \sum_{j=1}^{n} c_{ij} \gamma_j + \varepsilon_i \end{array}\right\}, \tag{3}$$

where c_{ij} is specified as the ijth element of a row-normalized connectivity matrix with "queen" contiguity (i.e., a single common point is used to define two polygons as neighbors). The parameter λ controls the degree of autocorrelation in the error term γ_i.

For MLR, OLS estimation is used to estimate the fixed coefficients β_j, that is $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, while for SAR, maximum likelihood is used to estimate both the fixed coefficients β_j and the fixed autocorrelation parameter λ (Anselin 1988). For GWR, weighted least squares estimation is used, where if a point has coordinates, (u_k, v_k) then the standard GWR estimate of $\beta(u_k, v_k)$ is given by solving:

$$\mathbf{X}^T \mathbf{W}_{(u_k, v_k)} \mathbf{X} \hat{\beta}(u_k, v_k) = \mathbf{X}^T \mathbf{W}_{(u_k, v_k)} \mathbf{y} \tag{4}$$

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, an adaptive bi-square kernel function is specified, so that the ith elements of the diagonal of $\mathbf{W}_{(u_k, v_k)}$ is:

$$w_{ii} = \left(1 - (d_{ik}/r_k)^2\right)^2 \text{ if } d_{ik} \leq r_k \quad w_{ii} = 0 \quad \text{otherwise,} \tag{5}$$

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 $\beta(u_k, v_k)$. Typically, r_k is chosen 'automatically' from the data set, and in this study, a corrected AIC approach is adopted for this purpose (Fotheringham, Brunsdon, and Charlton 2002).

The simulation experiment

The experiment generates data realizations to exhibit one of four spatial process (SP) scenarios: (1) stationary data relationships (fixed coefficients) with random error effects (call this SP1), (2) stationary relationships with spatially autocorrelated error effects (SP2), and (3) nonstationary relationships (varying coefficients) with random error effects (where SP3 and SP4 are for low and high coefficient variability, respectively), as detailed in Table 1. The study objective is to assess to what degree is each regression-type able to accurately model its designated process (i.e., MLR for SP1, SAR for SP2, and GWR for SP3-4).

As an overview, the experiment generates four regression coefficients, $\beta_0, \beta_1, \beta_2, \beta_3$ with three levels of nonstationarity; and then independently, the predictor data, x_1, x_2, x_3, are generated. The coefficient and predictor realizations 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. The error itself, is specified as either a random or spatially autocorrelated process. For each of the four processes (SP1-4), 100 data realizations are generated and the three study regressions are fitted and their model fit diagnostics reported. Realizations are

Table 1. Summary of the Four Spatial Processes Generated, Each with Their Different Spatial Characteristics

Spatial process no.	Intercept and coefficients $\beta_0, \beta_1, \beta_2, \beta_3$	Error term ε_i	Mean to error ratio (%)
SP1	Stationary	Random	99.9:0.01
SP2	Stationary	Autocorrelated	90:10
SP3	Non-stationary (low variability)	Random	99.9:0.01
SP4	Non-stationary (high variability)	Random	99.9:0.01

generated to the ($n = 159$) centroids of the "counties of Georgia for the United States"; an educational attainment data set routinely used to demonstrate GWR (Fotheringham, Brunsdon, and Charlton 2002; Griffith 2008).[3] The geostatistical-based experiment is such, that many useful subcategories can be defined, where this design flexibility is key to why a second-order effects simulation approach is taken, in preference to other, more deterministic approaches (e.g., Wang, Mei, and Yan 2008). The approach is novel and provides useful stochasticity, enabling nuanced differences to each realization, generated from the same initial specifications.

Regression coefficient co-simulation

Nonstationary coefficient surfaces for $\beta_0, \beta_1, \beta_2, \beta_3$ are generated using an unconditional sequential Gaussian co-simulation (e.g., Wackernagel 2003), where un-conditional means that the realizations are not conditioned to any data. This procedure simultaneously generates coefficients that are spatially dependent and spatially co-dependent with each other, which is to be expected. A linear model of co-regionalisation (LMC) is specified with Matérn models (Matérn 1986), where relatively short and relatively long correlation range parameters are specified to generate surfaces with high and low levels of coefficient nonstationarity, respectively.[4] The smoothing parameter of the Matérn model is also varied and similarly influences coefficient nonstationarity, where a value of 0.75 produces a coefficient surface with a low level of smoothing, while a value of 1.5 produces a high level of smoothing. Coefficient variability (or nonstationarity) increases in this order: (1) coefficients generated with a long range coupled with high smoothness (for SP3); and (2) coefficients generated with a short range coupled with low smoothness (for SP4). The remaining parameters of the LMC, such as the nugget variances/cross-variances, the structural variances/cross-variances, and the (simple cokriging) means of each realization, are chosen with care, so to reduce unnecessary confounders in our understanding of the outcomes (Table 2). Example stationary and nonstationary coefficient realizations are given in Fig. 1.

Predictor variable co-simulation

The three predictor variables, x_1, x_2, x_3 are also generated using an unconditional sequential Gaussian co-simulation, where again, a LMC is specified with Matérn models. Again, spatially dependent/co-dependent predictors are entirely expected, as they themselves are spatial variables. The co-simulation parameters are given in Table 2, chosen to provide neutral levels of collinearity. Example realizations for the three predictors are given in Fig. 2a. The generated predictors together with their corresponding coefficients should ensure a broadly similar

Table 2. LMC Parameters Specified for the Co-Simulation of: (1) The Four Regression Coefficients and (2) the Three Predictor Variables. ME is the Maximum Extent of the Study Area

	Cokriging mean	Nugget variances and cross-variances	Structural variances and cross-variances	Correlation range	Smoothing parameter
(1) Regression coefficient					
β_0	2	0	7.0	$0.95 \times ME$ (long) or $0.15 \times ME$ (short)	0.75 (low) or 1.5 (high)
β_1	5	0	9.0	$0.95 \times ME$ (long) or $0.15 \times ME$ (short)	0.75 (low) or 1.5 (high)
β_2	5	0	9.5	$0.95 \times ME$ (long) or $0.15 \times ME$ (short)	0.75 (low) or 1.5 (high)
β_3	3	0	7.5	$0.95 \times ME$ (long) or $0.15 \times ME$ (short)	0.75 (low) or 1.5 (high)
All six cross pairs	–	0	0	$0.95 \times ME$ (long) or $0.15 \times ME$ (short)	0.75 (low) or 1.5 (high)
(2) Predictor variable					
x_1	0.45	0	0.91	$0.40 \times ME$	0.5
x_2	0.55	0	0.95	$0.40 \times ME$	0.5
x_3	0.50	0	0.93	$0.40 \times ME$	0.5
x_1, x_2	–	0	0.20	$0.40 \times ME$	0.5
x_1, x_3	–	0	0.30	$0.40 \times ME$	0.5
x_2, x_3	–	0	0.10	$0.40 \times ME$	0.5

influence on the response variable (i.e., multiply the respective (cokriging) means in Table 2, where the intercept β_0 can be taken to correspond to predictor x_0 with a mean of 1).

Error simulation: mean to error ratio

The mean to error ratio strongly controls the outcomes of the experiment, where as it tends to 0 : 100, worthwhile insights on model behavior reduce (as all models will essentially tend to their intercept). The chosen ratios reflect this tendency. A 99.9 : 0.01 ratio is used for processes generated with nonstationary relationships (SP3-4), as a nonstationary intercept term is also representative of the errors. This commonality leads to a difficult identification problem,[5] which is further complicated in that a nonstationary intercept tends to reflect autocorrelated errors. With these issues in mind, the random errors generated for these processes are relatively small; and clearly, the consideration of spatially 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 (SP1), whilst for stationary relationship processes with autocorrelated errors (SP2), the narrower 90 : 10 ratio is used.

Error simulation: error term

Thus with the constraints set in place for the ratio of the mean to error component for variation in the response, for each of the four spatial processes, either: (1) a random error term (for SP1,
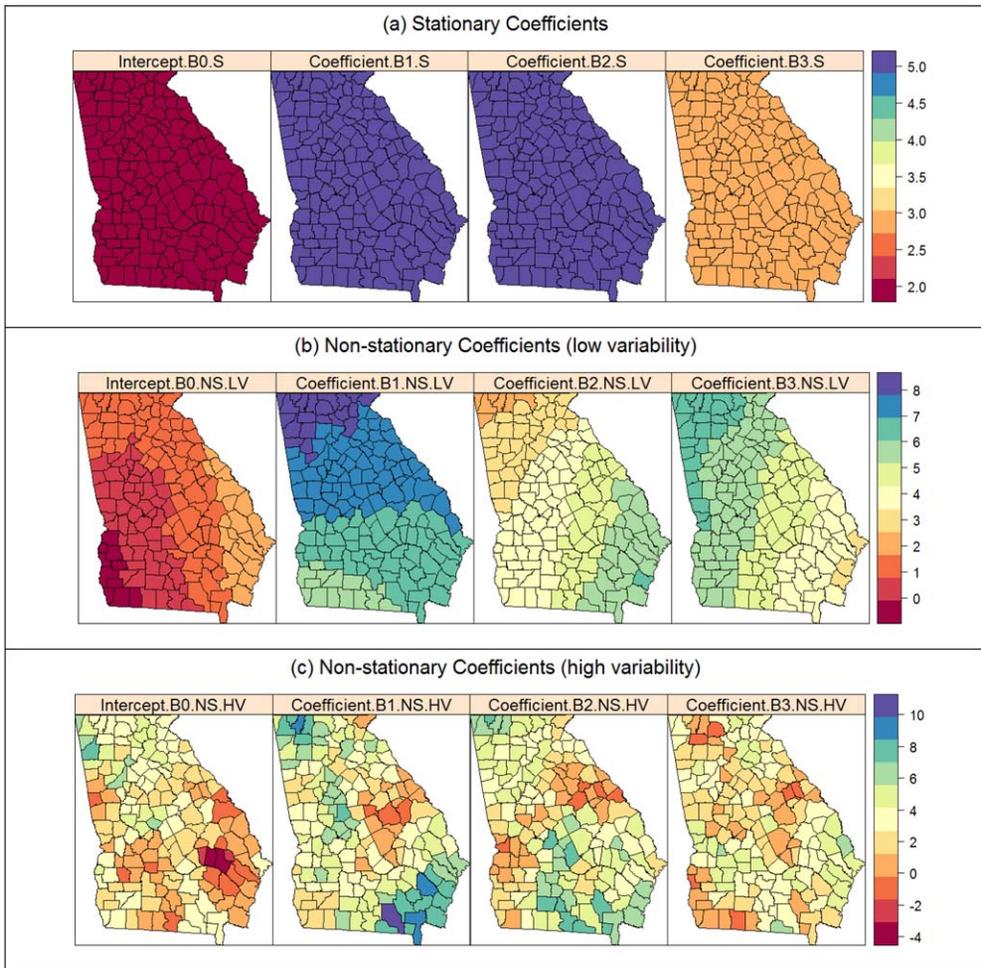
Figure 1. Example realizations for nonstationary regression coefficients: (a) zero variability; (b) low variability; and (c) high variability.

SP3-4) is generated via independent draws from a normal distribution, or (2) a spatially auto-correlated error term (for SP2) is generated via an un-conditional sequential Gaussian simulation, specified with an exponential variogram model, with a nugget variance of 0.2 and a correlation range set at three fifths of the maximum extent of the study area (such settings ensure a highly autocorrelated error term). All error terms are generated with a mean of zero. Fig. 2b,c present example realizations for the response and error terms, for SP1-4. On application of the simulation experiment, moderate to strong response to predictor relationships were generated, as would be hoped for.

Further points of interest

The steps of the simulation algorithm are such, that many of the (4 x 100 = 400) realizations will have common elements. For example, the first realizations for processes SP1 and SP2 will share the same regression coefficients and predictor data but differ in their error and response data; and this type of commonality pervades for all realizations. An area-based simulation is
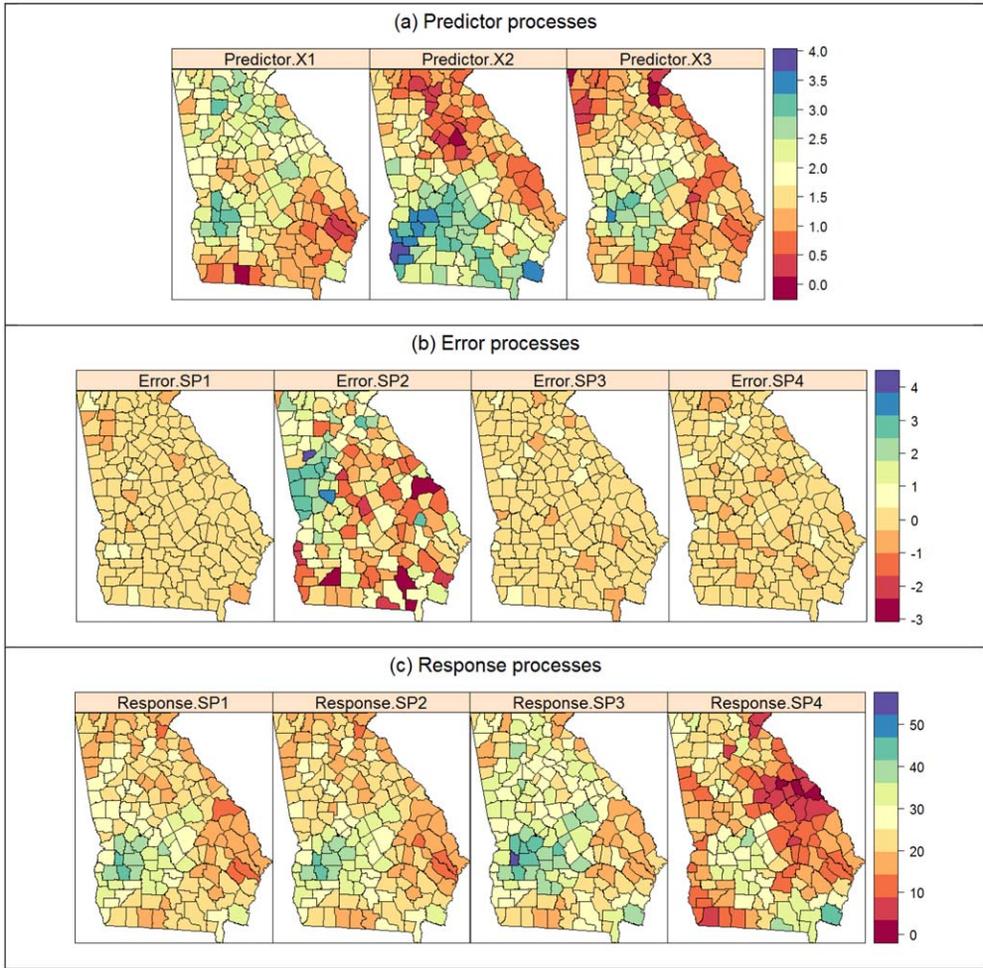
Figure 2. Example realizations for: (a) the predictors, (b) the errors for SP1–4, and (c) the responses for SP1–4.

preferred to grid-based approaches (e.g., Wang, Mei, and Yan 2008; Beale et al. 2010), as the latter are a poor reflection of geographical reality. The Georgia counties geographical system is similarly used in the GWR simulation study of Páez, Farber, and Wheeler (2011). Clearly, the simulation experiment has the potential to be highly involved, and only a few of the many (regression model and simulation) specifications are varied. In this respect, the outcomes are dependent on those specifications chosen. It is stressed that all specification decisions are made with the view that process SP1 should favor MLR, SP2 should favor SAR, and SP3-4 should favor GWR. Conversely, MLR should be a poor choice for SP2-4, SAR a poor choice for SP1, SP3-4, and GWR a poor choice for SP1-2.

Diagnostics reported

For each realization, Moran's *I* spatial autocorrelation tests are conducted on both the response and the residual from a MLR fit (Cliff and Ord 1981), using the same weights matrix as that defined for the SAR model, above. Hence for each test statistic, a sequence of *P*-value

distributions is found (each of size 100) and these are presented using boxplots. For a guide to spatial heterogeneity effects, the automatically found GWR bandwidths are reported (as percentage of sample data), for each realization. For model fit, (corrected) AIC values are found for MLR, SAR, and GWR fits, to each realization. AIC rankings are then found based on the ordering of the AIC results, where an AIC ranking of one, indicates the 'best fitting' regression, for that particular realization. Boxplots are again used to present the resultant bandwidth and ranked AIC distributions, for each of the four processes.

Statistics that reflect each regression's accuracy in estimating the simulated (actual) coefficients are also found. Due to the nature of the simulation experiment, a *relative* accuracy statistic is calculated for each of the study regressions, as this enables an objective comparison across all realizations. In particular, a relative root mean squared error (RMSE) is found, defined as relRMSE$=$RMSE$_{\text{Model}}$/RMSE$_{\text{Mean}}$. The further this ratio falls below unity, the greater the improvement in the accuracy of the model coefficient estimations to that found using the mean of the simulated coefficients, as the estimator. Coefficient estimation confidence interval (ECI) accuracy is assessed using coverage probabilities (Goovaerts 2001). For example, if symmetric 95% ECIs were found at each simulation point for a given coefficient, say the intercept, that is, $\hat{\beta}_0 \pm 1.96$ SE $\left(\hat{\beta}_0\right)$, then a correct modeling of local uncertainty would entail that there is a 0.95 (expected coverage) probability that the actual (simulated) value β_0, falls within the interval. If a coverage probability is found for a range of symmetric ECIs (say from a 1% to a 99% ECI in increments of 1%) and the results plotted against the probability interval p, then an accuracy plot is found which should follow the 45°, $x=y$ line. In a simulation study, it is unrealistic to present accuracy plots, so instead they are summarized via the G-statistic, defined as:

$$\text{G}-\text{STAT}=1-\int_0^1 [3a\,(p)-2]\left[\bar{\bar{\xi}}(p)-p\right]\,\mathrm{d}p \tag{6}$$

where $\bar{\bar{\xi}}$ is the fraction of actual values falling in the ECI; and G-STAT $= 1$, is ideal. The indicator function $a\,(p)$ is defined as $a(p)=\begin{cases} 1 & if\ \bar{\bar{\xi}}(p) \geq p \\ 0 & \text{otherwise} \end{cases}$, which implies that twice the importance is given to deviations when $\bar{\bar{\xi}}(p) < p$. For cases where two regressions provide similar G-STAT values for a given coefficient, one can be preferred if its ECI widths containing the actual coefficient are narrower. Here, corresponding ECI width plots can be constructed, but to act as a rough summary of this plot, a mean ECI width (M-ECI-W) for all p is found, which should be as small as possible. Thus, a regression model's G-STAT and M-ECI-W values for a given coefficient, should always be viewed in conjunction, as a strong G-STAT value is of little use if it is coupled with a poor M-ECI-W value (and vice versa).[6]

The relRMSE, G-STAT and M-ECI-W values are reported for the coefficient set, $\beta_0, \beta_1,$ β_2, β_3 considered as a whole, and not individually. This decision to aggregate the results is not ideal as $\beta_0, \beta_1, \beta_2, \beta_3$ are not independent (and can be strongly dependent due to collinearity) but is taken so as not to be over-burdened when reporting the results. For each of the four processes, the relRMSE values are summarized using boxplots, whilst the paired G-STAT and M-ECI-W values are presented using scatterplots that are spatially smoothed via kernel density estimation. Here, the M-ECI-W values are also re-scaled to lie between zero and one. Observe that in many instances of the experiment, an estimated coefficient is set the same at all locations (i.e., for MLR and SAR), as is a simulated coefficient (i.e., for SP1-2).

Finally, it is worthwhile to test for significant coefficient nonstationarity in a GWR fit. In this respect: (A) permutation tests are conducted for coefficient nonstationarity against a null hypothesis of stationary coefficients of a MLR model (Brunsdon, Fotheringham, and Charlton 1996); and (B) parametric bootstrap tests are conducted for coefficient nonstationarity, where GWR is an alternative to two null hypotheses, MLR and SAR models (Harris et al. 2017). For the bootstrap test, the SAR null allows an investigation as to whether the spatial variation in the coefficients obtained using GWR could be attributed to some other spatial process (i.e., some autocorrelated process), rather than one depicting nonstationary relationships. Bootstrap tests differ to permutation tests in that they do not condition on the exact values of the variables observed (Good 2005). Tests are applied with $R = 99$ randomizations or random samples, to each of the 100 data realizations, stemming from each of the four processes. A preferred setting of $R = 999$ was not followed as it would entail considerable computational burden within the simulation experiment.[7] The permutation test will result in a P-value distribution of size 100 for each coefficient, whilst the bootstrap test will consist of two P-value distributions (each of size 100) for each coefficient, and for each null. All such distributions are summarized with boxplots, where instances of $P \leq 0.05$ (as all one-tailed tests) would indicate significant coefficient nonstationarity (or variation) at the 95% level.

Results

Tests for autocorrelation

Significant spatial autocorrelation is found in the response variable (with $P \ll 0$), for all four processes for all 400 realizations (Fig. 3a). This is expected and considered a consequence of generating predictor data that are spatially autocorrelated, regardless of how the regression coefficients or error terms are generated. Significant autocorrelation is also found for the residual data sets (Fig. 3a), but this time, for only three of the four processes (with $P \ll 0$). Process SP1, does not tend to provide significant autocorrelated residuals from an MLR fit, as the error term was specified as such. Thus, for processes SP3-4, an inappropriate MLR fit will produce autocorrelated residuals. This behavior is not unexpected and confirms an inherent interrelationship between autocorrelation and heterogeneity. Essentially, it creates an identification issue that has a key bearing on many outcomes to follow.

GWR bandwidths

GWR bandwidth distributions are given in Fig. 3b. For process SP1, bandwidths tend to the maximum of 100%, strongly indicating GWR to be an inappropriate model choice, as would be expected. However, there are a few instances when much smaller bandwidths are found (as low as 46.5%), indicating that GWR will sometimes suggest spatial pattern in data relationships when no spatial effects are present. Bandwidths with an average size of 43.4%, result for process SP2, indicating that GWR will nearly always (and inappropriately) suggest a moderate spatial pattern in data relationships, when the response and residual data exhibit autocorrelation. This behavior is the reversal of that found above, with MLR fits to processes SP3-4, and is therefore, the same identification issue.

Bandwidths with an average size of 18.5%, result for all processes SP3-4, and thus provide the strongest evidence for choosing GWR, which would be expected. Coefficient processes that have a high variability (SP4) tend to provide larger bandwidths than those with low variability (SP3). This is somewhat counter-intuitive, but suggests that GWR will under-fit the former, whilst over-fit the latter. Over-fitting is a known consequence of applying a minimum AIC
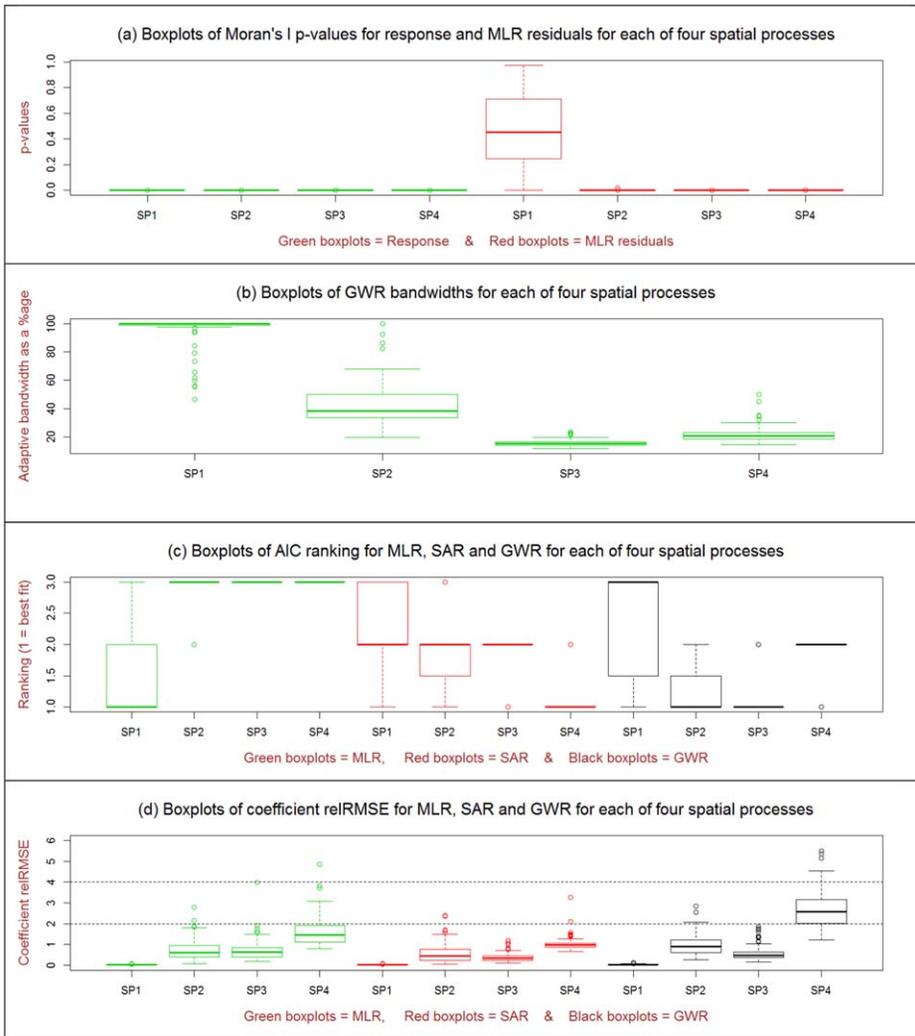
Figure 3. Boxplots for (a) Moran's *I* *P*-values (response and MLR residuals); (b) GWR bandwidths; (c) AIC rankings for MLR, SAR, and GWR model fit; and (d) relRMSE data for coefficient estimation accuracy.

criterion (Loader 2004), resulting in smaller bandwidths than required (Jetz, Rahbek, and Lichstein 2005). Assuming the generation of different levels of coefficient nonstationarity is reasonable, results reflect a direct limitation of GWR itself, and lend weight against the practice of automatic ("black-box") bandwidth selection (see Lu et al. 2014). Results also reflect that AIC-based bandwidth selection is based on prediction accuracy of the response, rather than coefficient estimation accuracy (which is not possible, of course).

Ranked AIC-based model fits
From Fig. 3c, MLR tends to provide the best fits to process SP1, which is expected, while GWR provides the poorest fits. On average, GWR provides the best fits to process SP2, which is contrary to what is expected. Here, SAR tends to provide the next best fits. GWR also

provides the best fits to process SP3 (low coefficient variability), whilst SAR provides the best fits to process SP4 (high coefficient variability). Given the roots of GWR lie in local regression smoothing (Loader 2004), this suggests that GWR is most suited to processes with a low coefficient variability. In summary, a regression with nonstationary coefficients can fit an autocorrelated error and stationary coefficient process, better than a regression designed to do so. Conversely, a stationary coefficient regression with autocorrelation effects can fit a nonstationary coefficient process, better than a regression designed to do so. Given this, model fit (at least via AIC) is not a particularly good discriminator across the process groups.

Coefficient accuracy

Boxplots of relRMSE values reflecting each regression's accuracy in estimating the actual coefficients are shown in Fig. 3d. All three models provide similar levels of coefficient estimation accuracy for process SP1, which is expected, as SAR and GWR will *tend* (or *default*) to the MLR calibration in these instances. Similarly, all models provide similar levels of coefficient estimation accuracy for process SP2, with SAR marginally the most accurate. This reflects that (stationary) coefficient estimation tends to not be adversely effected by whether or not autocorrelation is accounted for, whereas coefficient estimation uncertainty commonly is (see below). Coefficient estimation accuracy diverges when considering processes SP3-4. For low coefficient variability (SP3), model performance is broadly similar, but where SAR performs the best, while MLR performs the poorest. However, for high coefficient variability (SP4), GWR clearly performs the poorest, with SAR the best-performing. Thus, GWR's performance depends on the nature of coefficient variation, where relatively accurate estimates result only for coefficients with low levels of variability. Again, this reflects GWR's smoothing roots, and again, SAR can out-perform GWR when not expected to.

Although results suggest little value in applying GWR, it is stressed that the results are "on average." It is not unexpected for GWR to perform poorly in this respect and is a direct consequence of a *bias-variance* trade-off that is employed in these nonstationary models. Here, the size of the local data subset used to estimate each local coefficient needs to be small enough to be representative of the locality (i.e., *minimize bias*), but at the same time large enough so the coefficient is estimated reliably (i.e., *minimise variance*). Similarly, the "on average" results do not relay whether the GWR coefficient surfaces broadly reflect the actual coefficient patterns generated. This is the key remit of GWR after all—the spatial exploration of relationship heterogeneity via the mapping of the coefficients (see below).

Coefficient ECI accuracy

Coefficient ECI accuracy is assessed using paired G-STAT and M-ECI-W statistics in Fig. 4. As a promising G-STAT value takes preference to a promising M-ECI-W value, ideal results should cluster in the lower right quadrant of each plot, with the next best clustering in the upper right quadrant, while the poorest results are considered those that cluster in the bottom left quadrant. Results are not entirely as expected, where GWR performs relatively well for all four processes (SP1-4), and not just SP3-4.

For process SP1, MLR and SAR perform similarly, but GWR clearly performs best, as its results cluster in the lower right quadrant (although there are many anomalies). As before, MLR, SAR, and GWR would be expected to perform similarly for SP1 (as SAR and GWR should *default* to MLR). Of the four processes, the results for SP2 display the greatest scatter, where MLR is the poorest choice, whilst GWR outperforms SAR. For SP3 and SP4, MLR is
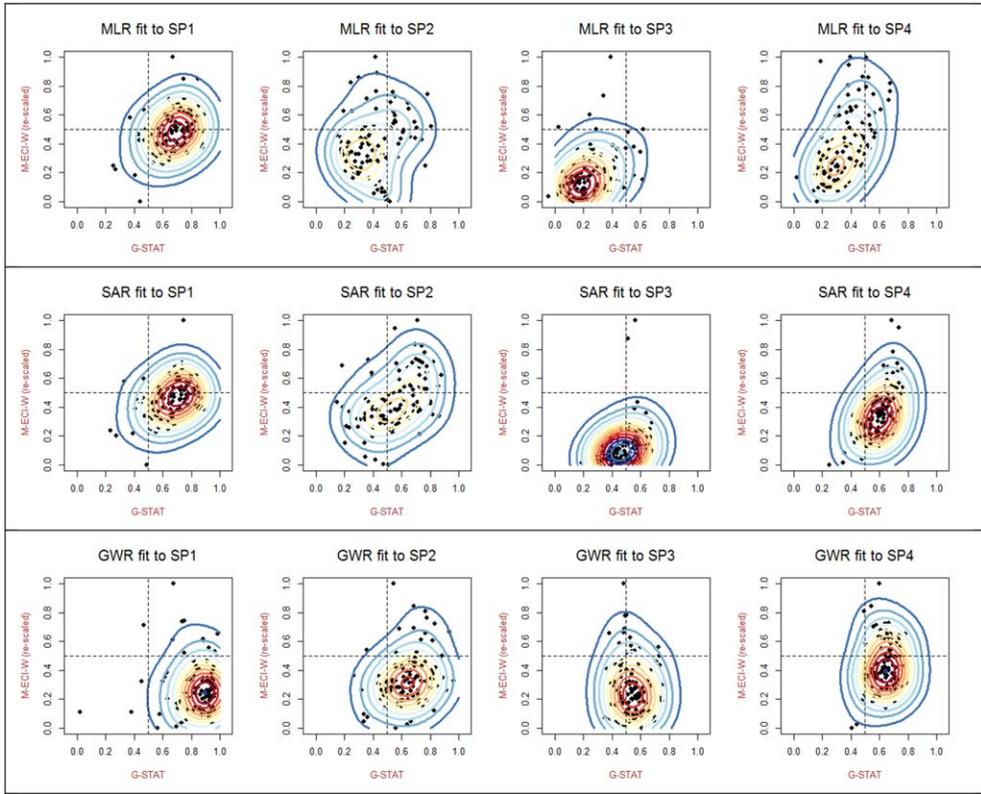
Figure 4. Scatterplots with spatial kernel densities for coefficient ECI accuracy via G-STAT and M-ECI-W data.

again the poorest model, whilst GWR marginally out-performs SAR (as GWR's outputs tend more to the two right-hand side quadrants). Thus, GWR outperforms SAR for autocorrelated error and stationary coefficient processes, whilst SAR can, on occasion, outperform GWR for nonstationary coefficient processes. These contrary results are now not so surprising.

Promising results indicate that uncertainties in a model's coefficient estimation are, in general, accurately accounted for in their coefficient standard errors. Thus, for GWR its relatively poor coefficient estimation accuracy performance (from above) is largely off-set by its relatively strong performance here. Maps depicting localized ECI accuracy may ascertain further worth of GWR in this respect (Harris, Brunsdon, and Charlton 2013b). Given these results, it is recommended that any GWR analysis should include an assessment of local coefficient estimation uncertainty, just as that routinely done in any stationary coefficient analysis. This is not so commonplace with GWR, often due to multiple hypothesis testing issues (da Silva and Fotheringham 2016) and limited inferential properties (e.g., Finlay 2011). Thus, reporting such uncertainties still requires caveats. Fotheringham and Oshan (2016) similarly argue for such assessments in relation to poor GWR coefficient estimation due to collinearity.

Tests for coefficient nonstationarity
On viewing the *P*-value boxplots in Fig. 5a, for process SP1, all tests (permutation and bootstrap) perform as they should do. There is little evidence for coefficient nonstationarity for all

nulls, and all response to predictor data relationships are correctly viewed as fixed across space. For process SP2 (Fig. 5b), the permutation tests rarely perform as expected. These tests consistently and erroneously indicate significant coefficient nonstationarity, for its MLR null, for all four coefficients. This is a direct reflection of GWR incorrectly finding spatial pattern to these autocorrelated processes, as observed before; and that a SAR null is not considered. For the bootstrap tests for SP2, the results depend on the null hypothesis, where MLR nulls, erroneously indicate coefficient nonstationarity, when none is present. Again, this is a consequence of GWR erroneously finding pattern to these processes, but now coupled with relatively small variability in the bootstrapped response variable with MLR nulls (as random errors are incorrectly assumed). This has the joint effect that the (false) variability seen in the GWR coefficients is viewed as significant. For SAR nulls, this is not the case, and these models tend to provide bootstrapped response variables that have sufficiently large enough variability, for the same false levels of variability seen in the GWR coefficients to be perfectly acceptable considering a process that *is* spatially autocorrelated with stationary relationships. Thus, SAR nulls are accepted and data relationships are correctly viewed as fixed across space.

For processes SP3-4 (Fig. 5c,d), the test outcomes depend on the variability of the coefficients generated. If the tests are working as expected, they would tend to provide low P-values, indicating coefficient nonstationarity, for all nulls, and for all coefficients. Permutation tests, only tend to work as they should for coefficients generated with high variability (SP4), whereas they appear inefficient as coefficient variation decreases. This behavior is expected, given the premise for this test. The bootstrap tests often perform in reverse to the permutation tests for the SAR null, but not the MLR null. Bootstrap tests only tend to work as they should for coefficients generated with low variability (SP3). Of the four coefficients, the intercept appears the most enigmatic, and most likely to be incorrectly viewed as stationary. This behavior is not surprising, given this term will reflect poor estimation in any of the three predictor coefficients. For a MLR null, the bootstrap test always correctly indicates that data relationships vary across space, regardless.

Overall, the permutation test has little value, whilst the bootstrap test has merit, but with limitations (see Harris et al. 2017). Confounding results are in part, a consequence of GWR's tendency to under- and over-fit giving rise to false levels of coefficient variability; and in part, a consequence of an MLR fit to a nonstationary coefficient process producing autocorrelated residuals. The SAR model specifications may also provide confounding issues but is not directly assessed here.

Spatial patterns from single realizations

For GWR, it is important that its outputs are interrogated spatially, but only example outputs from a single realization in turn can be presented, which is limiting given the stochastic nature of the experiment. That said, spatial patterns of actual and estimated coefficients are now compared and contextualized by an associated measure of coefficient uncertainty. For the latter, one approach is to map the local *pseudo t*-values from a GWR output, noting that values in the range [−1.96, +1.96] correspond to coefficients that do not significantly differ from zero, or some other quantity, at the 95% level (e.g., Harris, Fotheringham, and Juggins 2010c). Alternatively, a related approach is adopted, but now under any of the two null models, MLR or SAR, together with bootstrapping to determine significance levels, rather than a *t*-approximation. Here, $R = 99$ bootstrap samples are created, based on each null hypothesis, and local bootstrap P-values are mapped to identify where the varying GWR estimated coefficients significantly differ from the single, fixed one, estimated by the chosen null (for details, see Harris et al. 2017).
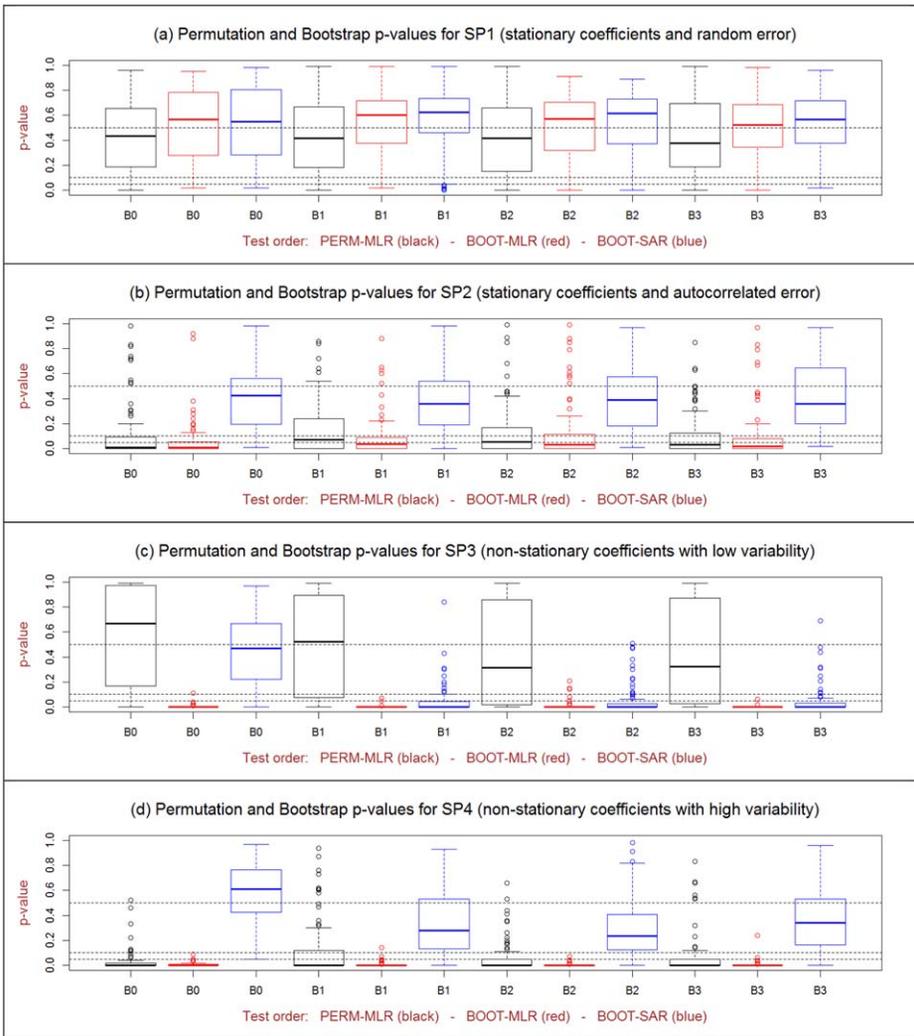
Figure 5. Boxplots of permutation test and bootstrap test *P*-values for coefficient nonstationarity, for SP1 to SP4. Plots are shown with *P*-value thresholds at 0.05, 0.10, and 0.50.

Example paired sets of surfaces are given for processes SP2 (Fig. 6) and for SP3-4 (Fig. 7). Surfaces are given for actual coefficients, estimated coefficients, and bootstrap *P*-values for MLR and SAR nulls, for β_1, β_2 only. As demonstration, localized coefficients that significantly differ from the corresponding global coefficient of the given null are those with *P*-values > 0.975 or *P*-values < 0.025 (respectively colored "dark blue" or "dark red" in the maps).[8] For context, *P*-values > 0.95 and *P*-values < 0.05 are also mapped. Surfaces for SP1 are not shown as in practise GWR would usually be dismissed from the outset, given the (mostly) large bandwidth results from before (Fig. 3b).

From Fig. 6a for SP2, it is clear that GWR will find spatial pattern in data relationships when none are present, with a proviso that spatial autocorrelation is present (which of course, is commonly the case). As a result, the bootstrap *P*-values (Fig. 6b) for a MLR null falsely indicate regions of significant coefficient nonstationarity (i.e., those colored "dark blue" or "dark
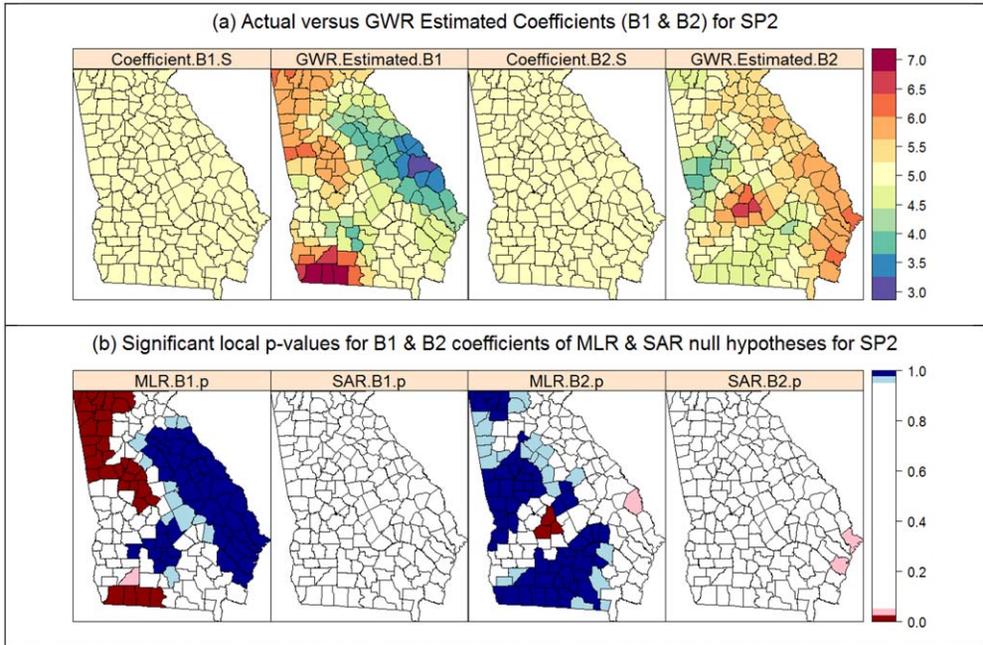
Figure 6. Example realizations for SP2 with (a) actual versus GWR estimated coefficients; and (b) significant local bootstrap *P*-values at 95% (and 90%) level for MLR and SAR null hypotheses. Given for β_1, β_2 only.

red" for respectively, much smaller or much larger coefficients than that estimated globally). Reassuringly however, the bootstrap *P*-values (Fig. 6b) for a SAR null gives no evidence for significant areas of coefficient nonstationarity, implying that the underlying spatial process is suited to a SAR-type model, as is expected.

From Fig. 7a for SP3, GWR will correctly find spatial pattern in data relationships when they are indeed present. However, for coefficient processes that are relatively low in variability, as is the case here, GWR will tend to over-fit, giving a heightened sense of relationship nonstationarity. As a result of the over-fit, the bootstrap *P*-values for the MLR null (Fig. 7b), indicates much larger regions of significant coefficient nonstationarity than is likely. Conversely, the SAR null (Fig. 7b) appears more conservative, indicating much smaller regions of significant coefficient nonstationarity. As an example, the estimated β_2 coefficients in the extreme south east (colored "blue" in far right map of Fig. 7a) are strongly considered to result from a nonstationary relationship between the response and predictor x_2, as the corresponding bootstrap *P*-values are significant (colored "dark blue" in both β_2 maps of Fig. 7b), indicating these coefficients to be significantly smaller than the MLR and SAR estimates of β_2.

From Fig. 7c for SP4 (high coefficient variability), GWR will again find spatial pattern in data relationships when they are present. However, GWR will now tend to under-fit, giving a reduced sense of relationship nonstationarity than that which exists; and unlike the low coefficient variability case, spatial patterns between actual, and estimated coefficients can be markedly dissimilar. The significant bootstrap *P*-values for the MLR null (Fig. 7d) simply reflect the patterns of the poorly estimated coefficients, and as such, suggest false regions of coefficient nonstationarity. There are very few significant bootstrap *P*-values for the SAR null (Fig. 7d), entailing that regions of true coefficient nonstationarity are unidentified.
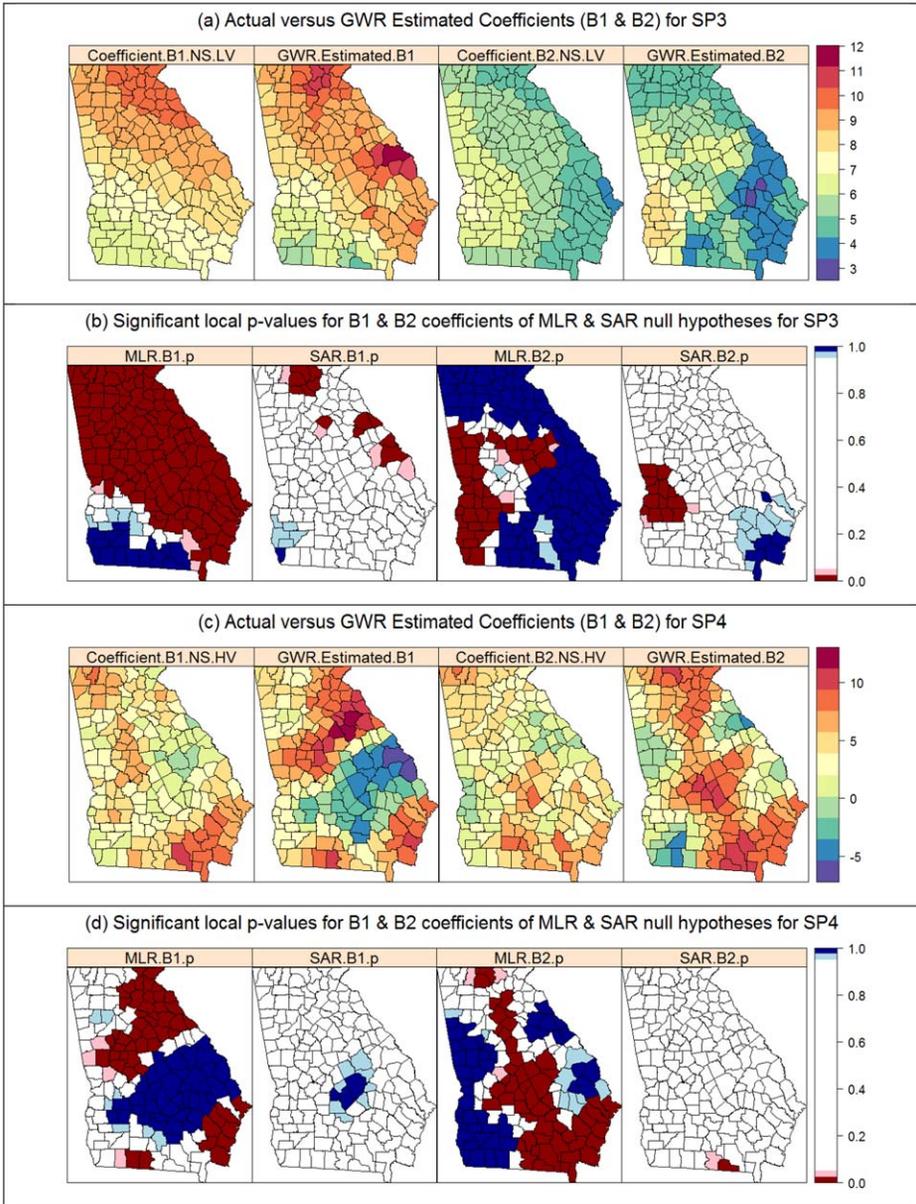
Figure 7. Example realizations for SP3 and SP4 with (a, c) actual versus GWR estimated coefficients; and (b, d) significant local bootstrap *P*-values at 95% (and 90%) level for MLR and SAR null hypotheses. Given for β_1, β_2 only.

Discussion

In summary, three routinely applied regressions to geographical data were assessed when applied to simulated data with known spatial characteristics. Given the specified characteristics are reasonable, results objectively confirm a strong interrelationship between autocorrelation and relationship heterogeneity (as extensively discussed in texts such as Anselin 1988, 2010). This can result in an identification problem when choosing one regression over another. An

MLR fit to a nonstationary coefficient process will tend to produce autocorrelated residuals, thus ensuring a SAR fit to be a reasonable, but incorrect model choice. Conversely, GWR will often find spatial pattern in the coefficients when none exists, provided the error is autocorrelated. On occasion, GWR will also find pattern in the coefficients when none exists, when the error is random—but in most cases, the bandwidth is large enough to strongly suggest a stationary coefficient process (this concurs with Páez, Farber, and Wheeler 2011). Furthermore, GWR's performance depends on the coefficient process itself, where it can both over- and under-fit depending on the nature of coefficient variability. It is well-known for GWR to over-fit (e.g., Jetz, Rahbek, and Lichstein 2005; Páez, Farber, and Wheeler 2011), but it is not so well-known that GWR also has the capacity to under-fit. All such identification problems can result in ineffectual model fit comparisons together with ineffectual tests for significant spatial effects. To provide valuable context to this simulation study, a discussion on its sensitivity to: (1) the chosen regression specifications and (2) the simulation design itself is given in the Appendix. For the latter, a sensitivity analysis is provided in respect of: (1) predictor collinearity, and (2) the mean to error ratio. Much of this additional work has a certain resonance for future study.

Empirical case study

To illustrate the issues described and provide a possible solution, a case study is given that uses the actual educational attainment data for the 159 counties of Georgia, USA. The response is the percentage of the adult population with a Bachelor's degree or higher qualification (*PctBach*), whose variance, is in part, explained by: (1) the total population in 1990 (*TotPop90*); (2) percentage of total population deemed to be living in a rural area (*PctRural*); (3) percentage of total population aged 65 or above (*PctEld*); (4) percentage of total population born outside of the US (*PctFB*); (5) percentage of total population living on or below the poverty level (*PctPov*); and (6) percentage of total population who are African Americans (*PctBlack*). A fourth-root transform was applied to *TotPop90*, entailing that nonlinear effects were immediately included. Initial regressions were again limited to that of MLR, SAR, and GWR only, together with a similar set of diagnostics to that used above. SAR and GWR use the same weight specifications as that defined above.

All predictors, aside from *PctBlack*, correlate well with *PctBach*. Both the response and the MLR residual term exhibit significant spatial autocorrelation, with Moran's *I* *P*-values of < 0.000 and < 0.001, respectively. The predictors provide similarly significant Moran's *I* *P*-values of $< 0.000, < 0.020, < 0.000, < 0.000, < 0.000$, and < 0.000, for *TotPop90, PctRural, PctEld, PctFB, PctPov*, and *PctBlack*, respectively.[9] Summaries of the MLR and SAR fits are given in Table 3, where in both cases, *Totpop90* and *PctFB* are the only significant predictors of *PctBach*. There is little difference between the MLR and SAR results, although, the SAR autocorrelation parameter (estimated at $\lambda = 0.34$) is significant with $P < 0.006$. For GWR, an AIC-defined optimum bandwidth at 88% ($n = 141$) suggests moderate relationship nonstationarity. Model AIC values decreased by 6.4 units from MLR to SAR and by 15.7 units from MLR to GWR. Model residual sum of squares (RSS) decreased by 125.1 units from MLR to SAR and by 375.4 units from MLR to GWR. Thus, GWR provides the best fit.

Bootstrap tests for GWR were run with respect to coefficient nonstationarity for MLR and SAR nulls (now with $R = 999$). In each case, the 95% points of the bootstrap samples were computed, and significance levels were found for upper one-tailed hypothesis tests. Results are given in Table 4, where, *PctFB* yields significant bootstrap *P*-values for both nulls. Thus, its

Table 3. Georgia Data Analysis: Result Summaries for MLR, SAR, GWR, and FBGWR Model Fits

	MLR				SAR			
	Estimate	Std. error	t Value	Prob. >|t|	Estimate	Std. error	z Value	Prob. >|z|
Intercept	2.61	3.40	0.767	0.444	2.67	3.19	0.838	0.402
TotPop90	0.703	0.143	4.902	0.000	0.745	0.140	5.340	0.000
PctRural	−0.0137	0.0171	−0.803	0.423	−0.0132	0.0160	−0.821	0.412
PctEld	0.00594	0.120	0.049	0.961	−0.0201	0.124	−0.163	0.870
PctFB	1.49	0.295	5.043	0.000	1.42	0.291	4.881	0.000
PctPov	−0.105	0.0742	−1.412	0.160	−0.0978	0.0780	−1.254	0.210
PctBlack	0.0347	0.0249	1.395	0.165	0.0231	0.0288	0.802	0.422

	GWR				FBGWR			
	Varying estimates by 1st, 2nd, and 3rd Quartiles			Number 'Significant' at 95% level	Fixed estimate or varying estimates by 1st, 2nd, and 3rd Quartiles			Number 'Significant' at 95% level
	Q_1	Q_2	Q_3		Q_1	Q_2	Q_3	
Intercept	3.25	4.49	5.30	22		4.04 (Fixed)		'Significant'
TotPop90	0.508	0.666	0.715	152		0.596 (Fixed)		'Significant'
PctRural	−0.0273	−0.0178	−0.00507	0		−0.00871 (Fixed)		'Significant'
PctEld	−0.136	−0.0838	−0.0351	0		−0.00806 (Fixed)		Not 'Significant'
PctFB	0.923	1.79	2.48	135	0.697	1.86	1.96	118
PctPov	−0.115	−0.0558	−0.00431	0		−0.151 (Fixed)		'Significant'
PctBlack	−0.00967	0.0132	0.0443	27	0.0507	0.0524	0.0556	159

Table 4. Georgia Data Analysis: Results of Bootstrap Tests for MLR and SAR nulls. Actual SD is the Standard Deviation of the GWR Coefficients Divided by Their Standard Errors, Resulting from the GWR Fit to the Data

	Intercept	TotPop90	PctRural	PctEld	PctFB	PctPov	PctBlack
Actual SD	0.568	0.774	0.588	0.389	2.135	0.717	1.184
MLR 95%	0.741	0.733	0.789	0.922	0.932	0.742	0.832
MLR *P*-value	0.153	0.041	0.154	0.569	0.000	0.061	0.003
SAR 95%	0.884	1.087	0.891	1.145	1.269	1.028	1.215
SAR *P*-value	0.252	0.133	0.231	0.702	0.001	0.180	0.057

coefficient is viewed as nonstationary, whose spatial variation exceeds that expected due to spatial effects such as that caused by autocorrelation. Weaker evidence of coefficient nonstationarity is shown for *Totpop90* and *PctBlack*, as each only provide significant bootstrap *P*-values for the MLR null at the 95% level. All other predictors appear to have stationary relationships with *PctBach*, including the *intercept*. For all predictors, the 95% point of the distribution of the test statistic, is always lower for MLR compared with SAR. Thus, the degree to which one might expect local coefficients to vary when a regression with fixed coefficients holds, is strongest with the SAR model.

Local bootstrap *P*-values for each coefficient are mapped for a SAR null only (Fig. 8a), with the same levels of significance as that used in the simulation experiment. The maps confirm the findings from Table 4 but provide spatial detail. For example, the coefficient for *PctFB* is clearly nonstationary, it being much larger than the fixed SAR estimate in the north of Georgia, whilst much smaller in the south. The higher *PctFB* coefficients in the north could be linked to the international make-up of the population living near to a number of universities in the region. The coefficient for *PctBlack* also appears nonstationary, as it is much larger than the SAR estimate in the south-west, while it is much smaller in the north-east. The coefficient for *Totpop90* is smaller than the SAR estimate in the north. The remaining coefficients are broadly stationary, but never entirely so. This completes the initial analysis.

Likely identification issues aside, it appears that the main outcome of the initial analysis is that both heterogeneity and autocorrelation effects are present but weak, as depicted by the relatively large bandwidth for GWR and the strong similarity between the MLR/SAR coefficients (and standard errors), respectively. Only *TotPop90* and *PctFB* are significant predictors globally, but where the relationship between *PctFB* to *PctBach* is likely to be nonstationary. The relationships between *TotPop90* and *PctBach* and between *PctBlack* and *PctBach* also appear nonstationary, but weakly so.

Given these findings, it is possible that the single bandwidth of basic GWR is effectively set too large for the nonstationary processes, as it was compromised (diluted) by the stationary processes. Thus, an obvious next step would be to replace basic GWR with: (1) a mixed GWR (Brunsdon, Fotheringham, and Charlton 1999), specified with the local relationships identified, while all others are fixed globally; or better still, (2) a flexible bandwidth GWR (FBGWR) (Leong and Yue 2017; Fotheringham, Yang, and Kang forthcoming; Lu et al. forthcoming). FBGWR relaxes the unrealistic assumption of the same degree of spatial smoothness for each coefficient in basic GWR (or those that are set as local, in mixed GWR). That is, FBGWR allows each relationship to vary at its own spatial scale, by finding coefficient-specific bandwidths.
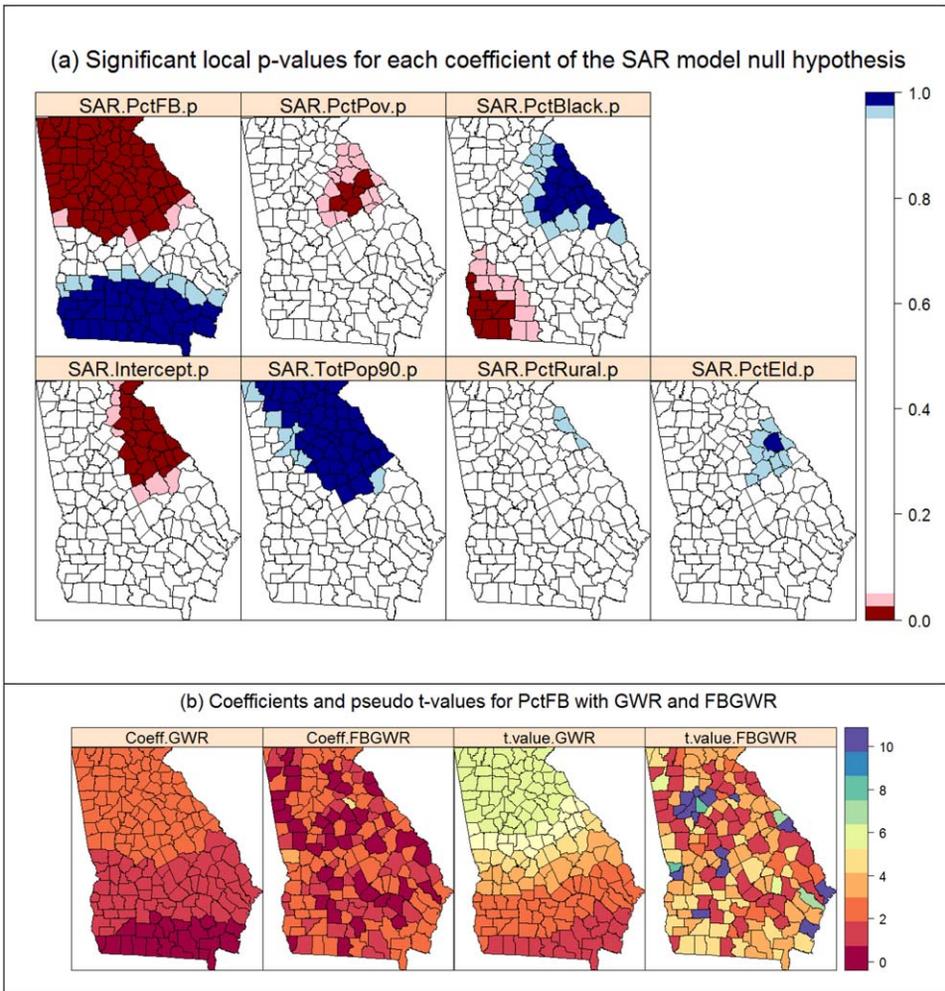
Figure 8. Georgia data analysis: (a) significant local bootstrap *P*-values at 95% (and 90%) level for SAR null hypothesis, and (b) coefficient and *pseudo t*-values for *PctFB* from GWR and FBGWR fits.

Thus, FBGWR was run using the algorithm of Lu et al. (forthcoming), where all seven bandwidths were found automatically via a corrected AIC-based back-fitting procedure. The results are in good agreement of that suggested above, where the bandwidths for the *intercept*, *TotPop90*, *PctRural*, *PctEld*, and *PctPov* were all set as global, while the bandwidths for *PctFB and PctBlack* were locally set at 34% ($n = 54$) and 92% ($n = 146$), respectively. Thus, the coefficients for *PctFB* are re-affirmed as nonstationary while the coefficients for *PctBlack* are re-affirmed as weakly nonstationary. All other coefficients are stationary or fixed, including that for *TotPop90*. Interestingly, the *intercept* is estimated as being fixed. Thus, autocorrelation effects[10] have been entirely dispensed with, since any GWR model with a locally varying intercept term is likely to indirectly account for residual autocorrelation.

The model AIC values decreased by 20.8 units from GWR to FBGWR, whilst the model RSS decreased by 208.8 units from GWR to FBGWR. FBGWR clearly improves model fit.

Summaries of the GWR and FBGWR fits are given in Table 3. As examples, coefficient and *pseudo t*-value surfaces for *PctFB* found from GWR and FBGWR are given in Fig. 8b, where the nature of *PctFB* coefficient nonstationarity clearly differs between the two models. Space precludes a more detailed assessment. Thus, in summary, the case study has demonstrated a roadmap to an entirely plausible final regression in FBGWR, through careful investigations of both autocorrelation and relationship heterogeneity effects.[11]

Final thoughts

Based on the results from a simulation experiment, this study has objectively confirmed an inherent interrelationship between spatial autocorrelation and spatial heterogeneity (with respect to data relationships), that results in an identification problem when choosing one regression over another. Although it should never be expected to entirely remove or solve the identification problem, useful guidelines on the implementation of regressions to spatial data have been provided both via the simulation experiment and via an empirical case study, so to reduce the chances of regression misspecification due to this clearly insidious issue. Intriguingly, as a method for dealing simultaneously with autocorrelation and heterogeneity, FBGWR appears a reasonable choice at least empirically. In this respect, future work could objectively assess FBGWR via a simulation experiment, benchmarking its performance with alternatives such as the hybrid mixed GWR-SAR models of Geniaux and Martinetti (forthcoming), and by natural extension, hybrids of both.

Acknowledgements

Notes

1 MLR is commonly referred to as "OLS" in the literature in respect of its OLS estimation procedure.

2 Unlike this study, Geniaux and Martinetti (forthcoming) has a core remit of introducing a new class of models (termed MGWR-SAR), where each regression coefficient together with an autocorrelation parameter can be specified as either fixed or locally varying—essentially a hybridization of mixed GWR (Brunsdon, Fotheringham, and Charlton 1999) with a GWR-SAR hybrid (Brunsdon, Fotheringham, and Charlton 1998). In this respect, their simulation experiment, although assessing MGWR-SAR to MLR, SAR, and GWR, is not as detailed as that presented here. In particular, their study does not investigate global and local coefficient uncertainty, say with confidence interval assessments and bootstrap tests, nor does it investigate the impacts and reasons of over- and under-fitting with GWR-based models.

3 In the simulation study of Páez, Farber, and Wheeler (2011), a sample size of around $n = 160$ is considered too small to reliably apply GWR. This is not viewed as a concern here, and if anything, presents a critical challenge to the value of this study's results. Furthermore, this recommendation should not be viewed as authoritative, as Páez, Farber, and Wheeler (2011) acknowledge with this statement: "However, an important step would be to determine whether an appropriate minimum sample size for the use of GWR exists, and what this value would be under various conditions." Fotheringham and Oshan (2016) also call for understanding on this issue with this statement: "Finally, a clearer understanding of the relationship between the sample size, the bandwidth, and the robustness of GWR would be useful." It is considered that a minimum sample size for GWR depends on the characteristics of the coefficient process, where small samples will suffice if this process is considered well-behaved (i.e., is

generated from a relatively large bandwidth, has no collinearity issues, has no outliers, etc.). Conversely, studies with (perceived) large samples may suffer from insufficient information if the coefficient process is not well-behaved (i.e., is generated from a relatively small bandwidth, has collinearity issues, has outliers, etc.).

4 Observe that coefficient nonstationary is simulated according to a fixed distance parameter (i.e., the correlation range), but an adaptive distance kernel is specified for GWR. Further work could investigate the use of both fixed and adaptive distance kernels.

5 The interpretation of a locally varying intercept together with locally varying errors is a subject beyond the scope of this study. Geo-additive models (Kammann and Wand 2003) may provide some direction in this respect.

6 Ultimately the coverage probabilities depend on the variance estimated for the corresponding estimators and that M-ECI-W depends directly on the variance.

7 On a medium specs laptop (Intel® core[TM] i7–4600U CPU @ 2.10–2.70 GHz with 16.0 GB using a 64-bit OS), the simulation experiment (including that reported in the Appendix) took nine days, three hours to run. This study primarily relied on functions provided by GWmodel (Lu et al. 2014; Gollini et al. 2015), gstat (Pebesma 2004), spgwr (Bivand and Yu 2014), and spdep (Bivand and Piras 2015) R packages.

8 Observe that the local bootstrap tests are two-tailed, whilst the global bootstrap test is one-tailed (Harris et al. 2017).

9 This empirical result of autocorrelated predictors lends weight to the likewise predictor specifications used in the simulation experiment, and suggested consequences thereafter. Murakami et al. (2017) and Geniaux and Martinetti (forthcoming) also acknowledge the influence of autocorrelated predictors and the identification difficulties that may result.

10 In practice, it would be prudent to re-investigate with further Moran's I tests with different weights structures, together with more informative measures such as variograms. In this respect, the Georgia data MLR residual variogram depicted no spatial structure.

11 Observe there is evidence of global and local collinearity between PctPov and PctBlack, which may result in regions of spurious negative estimates in PctBlack on applying GWR. However, the application of FBGWR alleviates this concern. Geniaux and Martinetti (forthcoming) similarly promote the use of mixed GWR forms, as a way of addressing collinearity.

Appendix

Sensitivity to regression specifications

A refinement to this research would be to conduct a sensitivity analysis for regression specification choices, while keeping the simulation design the same. Only basic MLR, SAR, and GWR models were specified, where SAR used a standard spatial weights structure, while GWR used a standard calibration procedure for its weights. Alternative specifications (i.e., with different weighting structures) exist for SAR (e.g., Kissling and Carl 2007; LeSage and Pace 2014) and similarly for GWR (Fotheringham, Brunsdon, and Charlton 2002; Gollini et al. 2015). For GWR this can be quite involved, with many kernels to choose from (Boxcar, Gaussian, Exponential, etc.), the choice of fixed or adaptive bandwidths and different bandwidth selection procedures (AIC or leave-one-out cross-validation, say). Related models could also have been chosen, such as a SAR lagged model where autocorrelation is modeled through the response rather than the error (Cressie 1993) or a locally linear GWR (Wang, Mei, and Yan 2008). There are also alternative models that have the same aims. For example, a geostatistical regression, where autocorrelation in the error term is measured continuously for all distances via a variogram function (e.g., Lark, Cullis, and Welham 2006), provides an alternative to SAR (see also Wall 2004). For example, a Bayesian spatially varying coefficient model (Gelfand et al. 2003) provides an alternative to GWR. In this respect, this study's outcomes must be placed in context of these rudimentary model choices
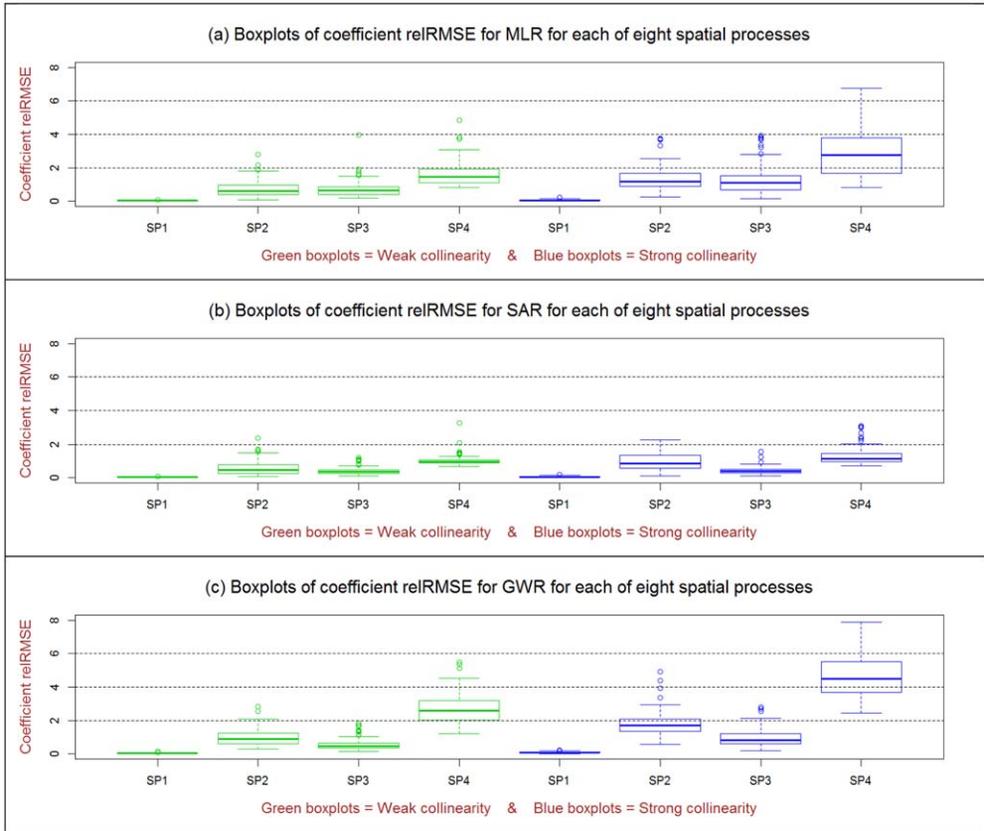
Figure A1. Boxplots of relRMSE data for coefficient estimation accuracy for weak versus strong levels of collinearity.

and specification decisions therein. Expanding the simulation experiment to include: (1) different specification forms, (2) model variants, and (3) model alternatives would swamp study objectives. That said, a few experiments were conducted. GWR was trialed with an exponential kernel function (a continuous function) but provided similar results to the (discontinuous) bi-square kernel. A SAR lagged model and a spatial moving average error model were trialed but gave similar results to the SAR error model.

Sensitivity to simulation design

Similarly, the simulation design could be extended to generate a more varied set of processes. For example, simulate processes with: (1) varying levels of predictor collinearity, (2) varying number of predictors, (3) varying levels of response to predictor correlations, (4) varying mean to error ratios, (5) varying levels of coefficient nonstationarity (i.e., not just low and high), (6) different scales of coefficient nonstationarity per relationship, (7) anisotropic coefficient nonstationarity, (8) varying strengths of residual autocorrelation, (9) different sample configurations (or areal units), (10) different sample sizes, (11) the introduction of global and local anomalies—all of which, would in turn, allow a full spectrum of related or alternative regressions to be assessed (e.g., ridge regressions for collinearity, robust
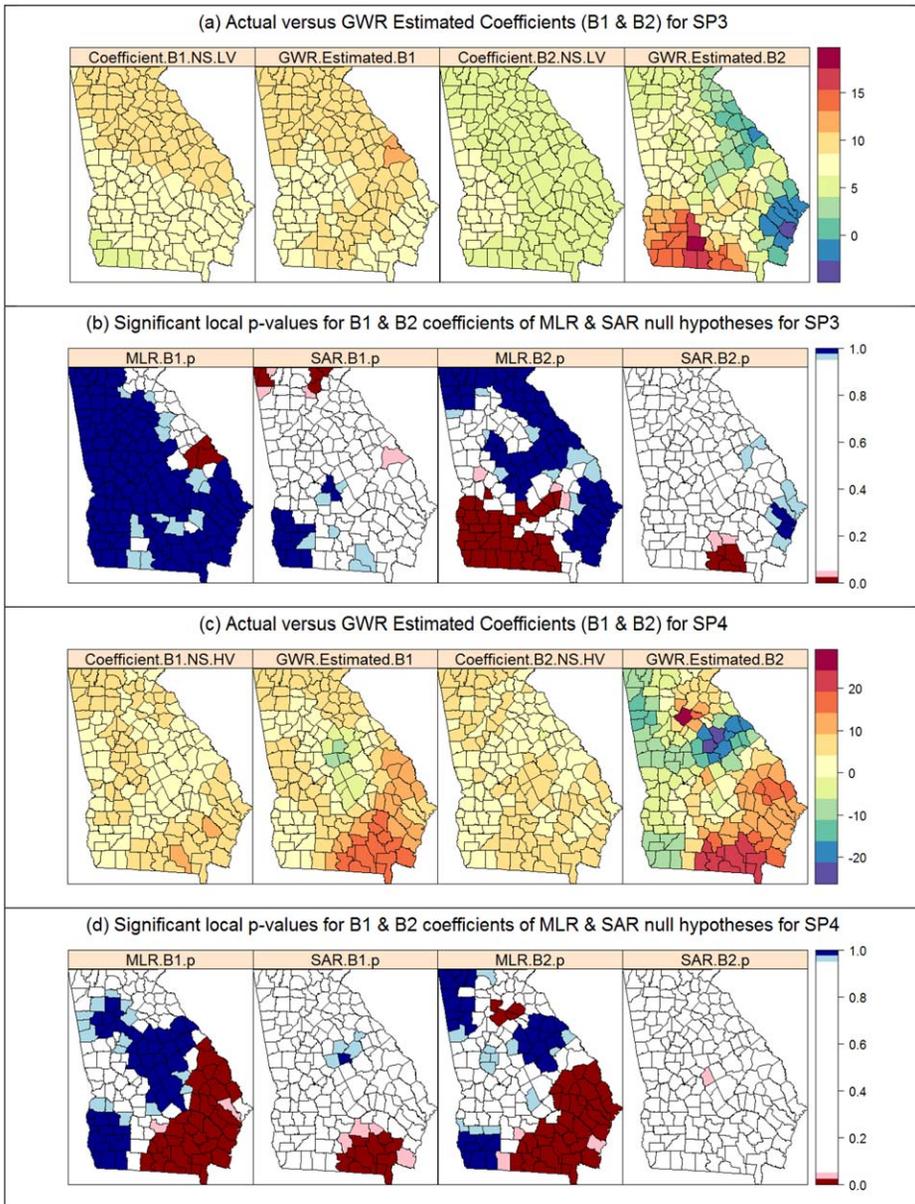
Figure A2. Example realizations for SP3 and SP4 with (a, c) actual versus GWR estimated coefficients; and (b, d) significant local bootstrap *P*-values at 95% (and 90%) level for MLR and SAR null hypotheses. Given for β_1, β_2 only. Both SP3-4 (main text with weak collinearity) and SP3–4 (here with strong collinearity) use *exactly the same* coefficient process.

regressions for anomalies). Again, this is not viable in a single presentation, as the experiment becomes highly factorial with many interaction levels (e.g., the interaction between the level of coefficient nonstationarity and sample size). However, two (noninteractive) experiments were conducted and presented below: (1) sensitivity to predictor collinearity, and (2) sensitivity to the set mean to error ratios. Finally, there are also sensitivities associated with
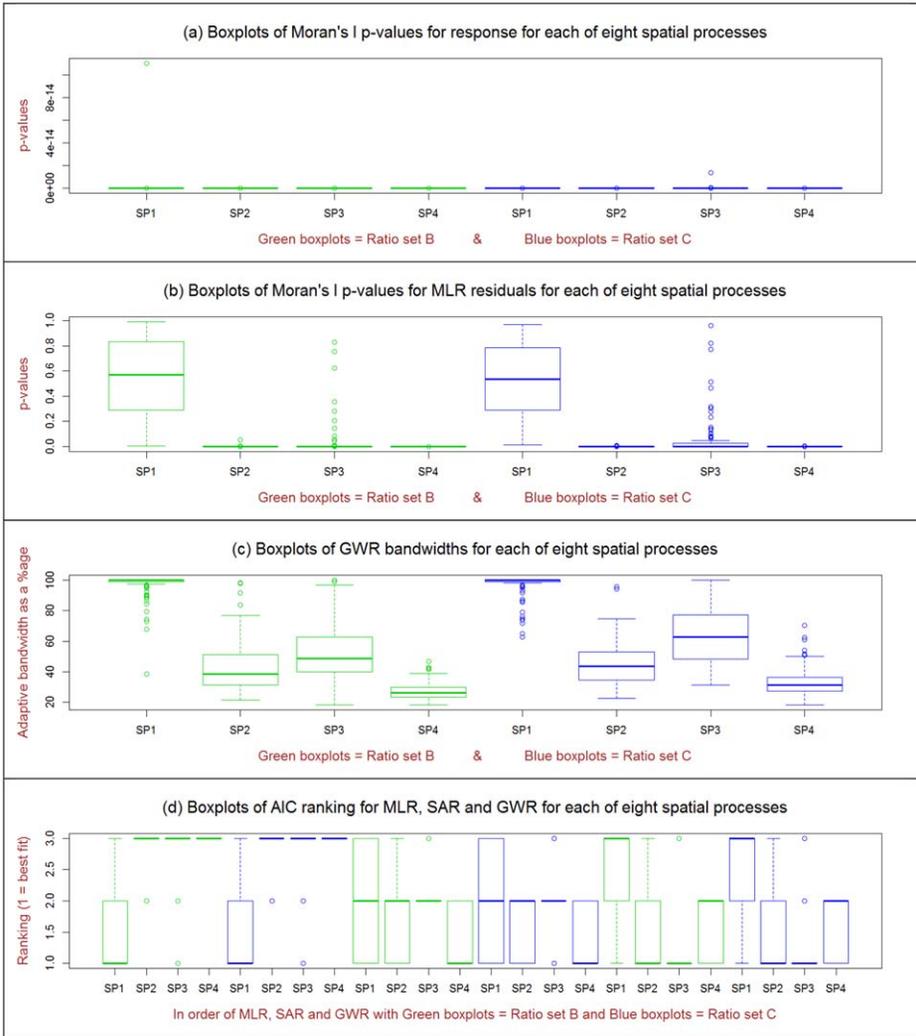
Figure A3. Boxplots for (a) Moran's *I* *P*-values (response); (b) Moran's *I* *P*-values (MLR residuals), (c) GWR bandwidths, and (d) AIC rankings for model fit—all for different mean to error ratios.

the diagnostics used. For example, the choice of weights matrix in the Moran's *I* tests, the choice of corrected AIC rather than some alternative measure of fit, setting $R = 99$ in the permutation and bootstrap tests rather than $R = 999$.

Sensitivity to simulation design: predictor variable collinearity

One issue when considering the estimation of coefficients in GWR is that of local collinearity among the predictors (Wheeler and Tiefelsdorf 2005; Wheeler 2007). For the global case (i.e., MLR and SAR), if there is a high degree of correlation between the predictors, then problems calibrating the model can follow (e.g., Dormann, Elith, and Bacher 2013). 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. Considering this, the simulation experiment was run with two levels of collinearity (weak and strong) between x_2, x_3 only. Here, the LMC cross-variances between x_2 and x_3 were specified to produce: (1) weak levels of collinearity between this predictor pair resulting in correlation coefficients of around $r = 0.1$ (i.e., as reported in the main text in Table 2), and (2) strong levels of collinearity resulting in correlations of around $r = 0.9$ (i.e., a cross-covariance of 0.9 is specified). On application of the simulation experiment, response to predictor relationships were slightly stronger with the strong levels of collinearity, as would be expected.

Results are reported in the same order as that found in the main text. First, collinearity was found to have little to no effect on: (1) the outcomes of the autocorrelation tests; (2) the determination of the GWR bandwidths; and (3) the AIC-based ranking of model fits. Such results are expected, as collinearity is primarily expected to influence coefficient accuracy and coefficient uncertainty accuracy. In this respect, boxplots of relRMSE values reflecting each regression's accuracy in estimating the actual coefficients are shown in Fig. A1. MLR, SAR, and GWR all under-perform due to strong collinearity, for the processes they are designed to suit. GWR can be the most effected, especially for process SP4. For coefficient ECI accuracy, strong collinearity tended to heighten coefficient estimation uncertainty, as it produced more scattered outputs throughout, although general patterns remained unchanged. For the tests for coefficient nonstationarity, strong collinearity had a similar effect in that the P-value distributions often displayed a wider interquartile range, but not to the extent of a different interpretation of the overall results. For spatial patterns, example surfaces are given for processes SP3-4 in Fig. A2, for comparison with Fig. 7 for corresponding weak collinearity surfaces. Given GWR is expected to perform relatively well for the low variability coefficient process (SP3) (Fig. 7a,b), its behavior is clearly compromised when fitted to *exactly the same* coefficient realization, but now with strong levels of predictor collinearity (Fig. A2a,b). GWR's poor behavior for the high coefficient variability case (SP4) with weak collinearity (Fig. 7c,d), is simply exasperated in the presence of strong collinearity (Fig. A2c,d).

Sensitivity to simulation design: the mean to error ratio

A second sensitivity analysis was conducted to highlight issues raised with respect to the mean to error ratio, where now: (1) SP1 and SP3-4 were generated with the 90 : 10 ratio, while SP2 was generated with an 80 : 20 ratio (call this "Ratio set B"); and (2) SP1 and SP3-4 were generated with the 80 : 20 ratio, while SP2 was generated with a 70 : 30 ratio (call this "Ratio set C"). The results presented in Fig. 3a-c (call this, "Ratio set A") are thus twice repeated in Fig. A3 using these extra realizations. Comparing Fig. 3a,c with Fig. A3, it is clear and entirely expected, that the outcomes of the simulation experiment become less distinct and less interpretable as this ratio narrows. This endorses the insidious nature of process identification.

References

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

Anselin, L. (1990). "Spatial Dependence and Spatial Structure Instability in Applied Regression Analysis." *Journal of Regional Science* 30, 185–207.

Anselin, L. (1995). "Local Indicators of Spatial Association." *Geographical Analysis* 27, 93–115.

Anselin, L. (2010). "Thirty Years of Spatial Econometrics." *Papers in Regional Science* 89(1), 3–25.

Armstrong, M. (1984). "Problems with Universal Kriging." *Mathematieal Geology* 16, 101–8.

Assunção, R. M. (2003). "Space Varying Coefficient Models for Small Area Data." *Environmetrics* 14, 453–73.

Atkinson, P. M. (2001). "Geographical Information Science: Geocomputation and Nonstationarity." *Progress in Physical Geography* 25, 111–22.

Basile, R., M. Durban, R. Minguez, J. M. Montero, and J. Mur. (2014). "Modeling Regional Economic Dynamics: Spatial Dependence, Spatial Heterogeneity and Non-Linearities." *Journal of Economic Dynamics and Control* 48, 229–45.

Basile, R., and B. Gress. (2005). "Semi-Parametric Spatial Auto-Covariance Models of Regional Growth Behaviour in Europe." *Region et Development* 21, 93–118.

Beale, C. M., J. J. Lennon, J. M. Yearsley, M. J. Brewer, and D. A. Elston. (2010). "Regression Analysis of Spatial Data." *Ecology Letters* 13, 246–64.

Besag, J., J. York, and A. Mollié. (1991). "Bayesian Image Restoration, with Applications in Spatial Statistics." *Annals of the Institute of Statistical Mathematics* 43, 1–20.

Bivand, R., and G. Piras. (2015). "Comparing Implementations of Estimation Methods for Spatial Econometrics." *Journal of Statistical Software* 63, 1–36.

Bivand, R., Z. Sha, L. Osland, and I. Sandvig Thorsen. (2017). "A Comparison of Estimation Methods for Multilevel Models of Spatially Structured Data." *Spatial Statistics* 21(B), 440–59.

Bivand, R., and D. Yu. (2014). spgwr: Geographically Weighted Regression. R package version 0.6-29. https://CRAN.R-project.org/package=spgwr

Brunsdon, C., A. S. Fotheringham, and M. Charlton. (1996). "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity." *Geographical Analysis* 28, 281–98.

Brunsdon, C., A. S. Fotheringham, and M. Charlton. (1998). "Spatial Nonstationarity and Autoregressive Models." *Environment and Planning A* 30(6), 957–93.

Brunsdon, C., A. S. Fotheringham, and M. Charlton. (1999). "Some Notes on Parametric Significance Tests for Geographically Weighted Regression." *Journal of Regional Science* 39, 497–524.

Casetti, E. (1972). "Generating Models by the Expansion Method: Applications to Geographical Research." *Geographical Analysis* 4(1), 81–91.

Christensen, O. F., P. J. Diggle, and P. J. Ribeiro. (2001). "Analyzing Positive-Valued Spatial Data: The Transformed Gaussian Model. In *geoENV III – Geostatistics for Environmental Applications*, 287–98, edited by P. Monestiez, D. Allard, and R. Froidevaux. Berlin: Germany: Springer.

Cliff, A. D., and J. K. Ord. (1981). *Spatial Processes: Models and Applications*. London, UK: London Pion Ltd.

Cressie, N. (1993). *Statistics for Spatial Data*. New York, NY: Wiley.

Cressie, N., and N. H. Chan. (1989). "Spatial Modeling of Regional Variables." *Journal of the American Statistical Association* 84, 393–401.

da Silva, A. R., and A. S. Fotheringham. (2016). "The Multiple Hypothesis Testing Issue in Geographically Weighted Regression." *Geographical Analysis* 48(3), 233–47.

Dong, G., and R. Harris. (2014). "Spatial Autoregressive Models for Geographically Hierarchical Data Structures." *Geographical Analysis* 47, 173–91.

Dong, G., R. Harris, K. Jones, and J. Yu. (2015). "Multilevel Modelling with Spatial Interaction Effects with Application to an Emerging Land Market in Beijing, China." *PLoS One* 10(6), 1–18.

Dormann, C. (2007). "Effects of Incorporating Spatial Autocorrelation into the Analysis of Species Distribution Data." *Global Ecology and Biogeography* 16, 129–38.

Dormann, C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. García Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, T. Münkemüller, C. McClean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell, S. Lautenbach. (2013). "Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance." *Ecography* 36, 27–46.

Finley, A. O. (2011). "Comparing Spatially-Varying Coefficient Models for Analysis of Ecological Data with Non-Stationary and Anisotropic Residual Dependence." *Ecology and Evolution* 2, 143–54.

Fischer, M., and P. Stumpner. (2010). "Income Distribution Dynamics and Cross-Region Convergence in Europe." In *Handbook of Applied Spatial Analysis*, Vol. 4, 599–628, edited by M. Fischer and A. Getis. Springer-Verlag Berlin Heidelberg.

Fotheringham, A. S., C. Brunsdon, and M. Charlton. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships.* Chichester: Wiley.

Fotheringham, A. S., and T. M. Oshan. (2016). "Geographically Weighted Regression and Multicollinearity: Dispelling the Myth." *Journal of Geographical Systems* 18, 303–29.

Fotheringham, A. S., W. Yang, and W. Kang. (2017). "Multiscale Geographically Weighted Regression (MGWR)." *Annals of the American Association of Geographers* 107(6), 1247–1265.

Fouedjio, F. (2016). "Second-Order Non-Stationary Modelling Approaches for Univariate Geostatistical Data." *Stochastic Environmental Research and Risk Assessment* 31(8), 1887–1906.

Fuglstad, G.-A., D. Simpson, F. Lindgren, and H. Rue. (2015). "Does Non-Stationary Spatial Data Always Require Non-Stationary Random Fields?" *Spatial Statistics* 14, 505–31.

Gelfand, A. E., H.-J. Kim, C. F. Sirmans, and S. Banerjee. (2003). "Spatial Modeling with Spatially Varying Coefficient Processes." *Journal of the American Statistical Association* 98, 387–96.

Geniaux, G., and D. Martinetti. (forthcoming). "A New Method for Dealing Simultaneously with Spatial Autocorrelation and Spatial Heterogeneity in Regression Models" *Regional Science and Urban Economics*, Available at: https://doi.org/10.1016/j.regsciurbeco.2017.04.001

Gollini, I., B. Lu, M. Charlton, C. Brunsdon, and P. Harris. (2015). "GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models." *Journal of Statistical Software* 63, 1–50.

Good, P. I. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses.* Berlin: Springer International Publishing AG.

Goodchild, M. F. (2004). "The Validity and Usefulness of Laws in Geographic Information Science and Geography." *Annals of the Association of American Geographers* 94(2), 300–3.

Goovaerts, P. (2001). "Geostatistical Modelling of Uncertainty in Soil Science." *Geoderma* 103, 3–26.

Gotway, C. A., and L. J. Young. (2002). "Combining Incompatible Spatial Data." *Journal of the American Statistical Association* 97(458), 632–48.

Griffith, D. A. (2003). *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding through Theory and Scientific Visualization.* Berlin: Springer.

Griffith, D. A. (2008). "Spatial-Filtering-Based Contributions to a Critique of Geographically Weighted Regression (GWR)." *Environment and Planning A* 40, 2751–69.

Haas, T. C. (1996). "Multivariate Spatial Prediction in the Presence of Non-Linear Trend and Covariance Nonstationarity." *Environmetrics* 7, 145–65.

Harris, P., C. Brunsdon, and M. Charlton. (2013b). "The Comap as a Diagnostic Tool for Nonstationary Kriging Models." *International Journal of Geographical Information Science* 27(3), 511–41.

Harris, P., C. Brunsdon, B. Lu, T. Nakaya, and M. Charlton. (2017). "Introducing Bootstrap Methods to Investigate Coefficient Non-Stationarity in Spatial Regression Models." *Spatial Statistics* 21, 241–61.

Harris, P., M. Charlton, and A. S. Fotheringham. (2010a). "Moving Window Kriging with Geographically Weighted Variograms." *Stochastic Environmental Research and Risk Assessment* 24, 1193–209.

Harris, R., G. Dong, and W. Zhang. (2013a). "Using Contextualized Geographically Weighted Regression to Model the Spatial Heterogeneity of Land Prices in Beijing, China." *Transactions in GIS* 17(6), 901–19.

Harris, P., A. S. Fotheringham, R. Crespo, and M. Charlton. (2010b). "The Use of Geographically Weighted Regression for Spatial Prediction: An Evaluation of Models Using Simulated Data Sets." *Mathematical Geosciences* 42, 657–80.

Harris, P., A. S. Fotheringham, and S. Juggins. (2010c). "Robust Geographically Weighted Regression: A Technique for Quantifying Spatial Relationships between Freshwater Acidification Critical Loads and Catchment Attributes." *Annals of the Association of American Geographers* 100, 286–306.

Jetz, W., C. Rahbek, and J. W. Lichstein. (2005). "Local and Global Approaches to Spatial Data Analysis in Ecology." *Global Ecology and Biogeography* 14, 97–8.

Kammann, E. E., and M. P. Wand. (2003). "Geoadditive Models." *Journal of Royal Statistical Society C, Applied Statistics* 52(1), 1–18.

Kissling, W. D., and G. Carl. (2007). "Spatial Autocorrelation and the Selection of Simultaneous Autoregressive Models." *Global Ecology and Biogeography* 17, 59–71.

Lambert, D. M., W. Xu, and R. J. Florax. (2014). "Partial Adjustment Analysis of Income and Jobs, and Growth Regimes in the Appalachian Region with Smooth Transition Spatial Process Models." *International Regional Science Review* 37(3), 328–64.

Lark, R. M., B. R. Cullis, and S. J. Welham. (2006). "On Spatial Prediction of Soil Properties in the Presence of a Spatial Trend: The Empirical Best Linear Unbiased Predictor (E-BLUP) with REML." *European Journal of Soil Science* 57, 787–99.

Leong, Y. Y., and J. C. Yue. (2017). "A Modification to Geographically Weighted Regression." *International Journal of Health Geographics* 16(1), 11.

LeSage, J., and K. Pace. (2009). *Introduction to Spatial Econometrics*. Boca Raton: CRC Press.

LeSage, J., and K. Pace. (2014). "The Biggest Myth in Spatial Econometrics." *Econometrics* 2, 217–49.

Lloyd, C. D. (2010). *Local Models for Spatial Analysis*, 2nd ed. Boca Raton: CRS Press.

Loader, C. (2004). *Smoothing: Local Regression Techniques*. Berlin: Center for Applied Statistics and Economics (CASE), Humboldt-Universit.

Lu, B., C. Brunsdon, M. Charlton, and P. Harris. (2017). "Geographically Weighted Regression with Parameter-Specific Distance Metrics." *International Journal of Geographical Information Science* 31(5), 982–998.

Lu, B., P. Harris, M. Charlton, and C. Brunsdon. (2014). "The GWmodel R Package: Further Topics for Exploring Spatial Heterogeneity Using Geographically Weighted Models." *Geo-Spatial Information Science* 17, 85–101.

Matérn, B. (1986). "Spatial Variation." In *Lecture Notes in Statistics*, edited by B. Matérn. Berlin: Springer.

McMillen, D. P. (1996). "One Hundred Fifty Years of Land Values in Chicago: A Nonparametric Approach." *Journal of Urban Econimics* 40, 100–24.

McMillen, D. P. (2001). "Nonparametric Employment Subcentre Identification." *Journal of Urban Economics* 50, 448–73.

McMillen, D. P. (2003). "Spatial Autocorrelation or Model Misspecification?" *International Regional Science Review* 26, 208–17.

Mellin, C., K. Mengersen, C. J. A. Bradshaw, and M. J. Caley. (2014). "Generalizing the Use of Geographical Weights in Biodiversity Modelling." *Global Ecology and Biogeography* 23, 1314–23.

Montero, J., R. Mínguez, and M. Durbán. (2012). "SAR Models with Nonparametric Spatial Trends. A P-Spline Approach." *Estadística Española* 54, 89–111.

Murakami, D., and M. Tsutsumi. (2015). "Area-to-Point Parameter Estimation with Geographically Weighted Regression." *Journal of Geographical Systems* 17(3), 207–25.

Murakami, D., T. Yoshida, H. Seya, D. A. Griffith, and Y. Yamagata. (2017). "A Moran Coefficient-Based Mixed Effects Approach to Investigate Spatially Varying Relationships." *Spatial Statistics* 19, 68–89.

Mur, J., F. López, and A. Angelo. (2008). "Symptoms of Instability in Models of Spatial Dependence." *Geographical Analysis* 40, 289–11.

Mur, J., F. López, and A. Angelo. (2010). "Instability in Spatial Error Models: An Application to the Hypothesis of Convergence in the European Case." *Journal of Geographical Systems* 12(3), 259–80.

Myers, D. E. (1989). "To Be or Not to Be. . . .Stationary? That is the Question." *Mathematical Geology* 21(3), 347–62.

Opsomer, J., Y. Wang, and Y. Yang. (2001). "Nonparametric Regression with Correlated Errors." *Statistical Science* 16(2), 134–53.

Osland, L., I. S. Thorsen, and I. Thorsen. (2016). "Accounting for Local Spatial Heterogeneities in Housing Market Studies." *Journal of Regional Science* 56, 895–920.

Pace, K., and J. Lesage. (2004). "Spatial Autoregressive Local Estimation." In *Spatial Econometrics and Spatial Statistics*, 31–51, edited by A. Getis, J. Mur, and H. Zoller. Londres, UK: Palgrave.

Páez, A., S. Farber, and D. Wheeler. (2011). "A Simulation-Based Study of Geographically Weighted Regression as a Method for Investigating Spatially Varying Relationships." *Environment and Planning A* 43, 2992–3010.

Pebesma, E. J. (2004). "Multivariable Geostatistics in S: The Gstat Package." *Computers & Gesciences* 30, 683–91.

Sampson, P. D., D. Damian, and P. Guttorp. (2001). Advances in modeling and inference for environmental processes with nonstationary spatial covariance. NRCSE-TRS No 61, National Research Centre for Statistics and the Environment, University of Washington, Technical Report Series.

Shen, S.-L., C.-L. Mei, and Y.-J. Zhang. (2011). "Spatially Varying Coefficient Models: Testing for Spatial Heteroscedasticity and Reweighting Estimation of the Coefficients." *Environment and Planning A* 43, 1723–45.

Stefanova, K., A. Smith, and B. Cullis. (2009). "Enhanced Diagnostics for the Spatial Analysis of Field Trials." *Journal of Agricultural, Biological and Environmental Statistics* 14, 1–19.

Su, L., and S. Jin. (2010). "Profile Quasi-Maximum Likelihood Estimation of Partially Linear Spatial Autoregressive Models." *Journal of Econometrics* 157, 18–33.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications.* Berlin Heidelberg: Springer-Verlag.

Wall, M. M. (2004). "A Close Look at the Spatial Structure Implied by the CAR and SAR Models." *Journal of Statistical Planning and Inference* 121, 311–24.

Wang, N., C. L. Mei, and X. D. Yan. (2008). "Local Linear Estimation of Spatially Varying Coefficient Models: An Improvement on the Geographically Weighted Regression Technique." *Environment and Planning A* 40, 986–1005.

Wheeler, D. C. (2007). "Diagnostic Tools and a Remedial Method for Collinearity in Geographically Weighted Regression." *Environment and Planning A* 39, 2464–81.

Wheeler, D., and M. Tiefelsdorf. (2005). "Multicollinearity and Correlation Among Local Regression Coefficients in Geographically Weighted Regression." *Journal of Geographical Systems* 7, 161–87.

Yu, D., N. A. Peterson, and R. J. Reid. (2009). "Exploring the Impact of Non-Normality on Spatial Non-Stationarity in Geographically Weighted Regression Analyses: Tobacco Outlet Density in New Jersey." *GIScience & Remote Sensing* 46(3), 329–46.

Zhang, J., P. Atkinson, and M. Goodchild. (2014). *Scale in Spatial Information and Analysis.* Boca Raton: CRC Press.