Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 27 (2015) 112 – 115

Procedia Environmental Sciences

Spatial Statistics 2015: Emerging Patterns

Using Bootstrap Methods to Investigate Coefficient Non-stationarity in Regression Models: An Empirical Case Study

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

[a]*Rothamsted Research, North Wyke, UK*
[b]*National Centre for Geocomputation, Maynooth University, Ireland*
[c]*University of Bristol, UK*
[d]*Ritsumeikan University, Kyoto, Japan*

Abstract

In this study, parametric bootstrap methods are used to test for spatial non-stationarity in the coefficients of regression models (i.e. test for relationship non-stationarity). Such a test can be rather simply conducted by comparing a model such as geographically weighted regression (GWR) as an alternative to a standard regression, the null hypothesis. However here, 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. In this short presentation, the bootstrap approach is applied empirically to an educational attainment data set for Georgia, USA. Results suggest value in the bootstrap approach, providing a more informative test than any related test that is commonly applied.

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns committee

Keywords: GWR; Georgia Data; Hypothesis Testing; Spatial Regression; Spatial Nonstationary

* Paul Harris. Tel.: +44-1837-883535.
E-mail address: paul.harris@rothamsted.ac.uk

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns committee
doi:10.1016/j.proenv.2015.07.106

1. Introduction

The method of GWR [1] provides a means of exploration of a multiple linear regression (MLR) model in which the coefficients show a tendency to vary over space. GWR is essentially spatial, in the sense that the value of a predicted response variable or a regression coefficient depends on the location in space. For the case where there are several predictor variables $y_1, y_2, \ldots y_p$ and $i = 1, \ldots, n$, MLR has this form for response variable z:

$$z_i = \beta_0 + \sum_{j=1}^{p} \beta_j \, y_{ij} + \varepsilon_i \qquad (1)$$

where the coefficients β, are commonly estimated by ordinary least squares. 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:

$$z_i = \beta_0(u_i, v_i) + \sum_{j=1}^{p} \beta_j(u_i, v_i) y_{ij} + \varepsilon_i \qquad (2)$$

where (u_i, v_i) is the spatial location of the i^{th} observation and $\beta_j(u_i, v_i)$ is a realisation of the continuous function $\beta_j(u, v)$ at point i. As with (ordinary) MLR, the ε_i's in GWR are random error terms which are independently normally distributed with zero mean and common variance σ^2. Therefore 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 function. Various methods have been proposed to assess the validity of GWR in comparison to MLR [1,2]. However GWR is only one of many spatial models. In particular, there are a number of models in which the *z*-variable or the error term exhibits spatial autocorrelation, although the regression coefficients remain fixed over space [3]. Among these models is the spatial simultaneous autoregressive error (ERR) model:

$$\left.\begin{aligned} z_i &= \beta_0 + \sum_{j=1}^{p} \beta_j \, y_{ij} + \gamma_i \\ \text{where} \qquad \gamma_i &= \lambda \sum_{j=1}^{n} c_{ij} \gamma_j + \varepsilon_i \end{aligned}\right\} \qquad (3)$$

where c_{ij} is the ij^{th} 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 spatial moving average (SMA) model:

$$\left.\begin{aligned} z_i &= \beta_0 + \sum_{j=1}^{p} \beta_j \, y_{ij} + \gamma_i \\ \text{where} \qquad \gamma_i &= \lambda \sum_{j=1}^{n} c_{ij} \varepsilon_j + \varepsilon_i \end{aligned}\right\} \qquad (4)$$

As before, λ governs the degree of spatial association. A further alternative is the spatial simultaneous autoregressive lag (LAG) model:

$$z_i = \beta_0 + \sum_{j=1}^{p} \beta_j \, y_{ij} + \lambda \sum_{j=1}^{n} c_{ij} \, z_j + \varepsilon_i \qquad (5)$$

In this case, each z_i depends on the neighbouring *z*-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.

Thus it would be useful to compare GWR not only to MLR, but also with ERR, SMA and LAG. In this respect, a bootstrapping method [4] is proposed that assesses the variability of locally weighted estimates of the regression

coefficients from GWR under the assumptions of each of equations (1, 3, 4 and 5). The observed values of variability are then compared against these as reference distributions. The bootstrapping method provides a statistical significance test; the general methodology of which is outlined in [5].

2. The bootstrap approach

The bootstrap is a technique for estimating characteristics of the sampling distribution of a wide range of test statistics when a theoretical distribution is not readily obtainable. The underlying distribution of the data need not be known, although if it is, this information can be used by the technique to improve the estimates of the test statistic distribution. The parametric bootstrap is when the underlying data distribution is known; which is the case here.

Thus bootstrap samples of the *z*- and *y*-variables are found for each of four null models: MLR, ERR, SMA and LAG. In each case, the predictor *y*-variables are not considered random and are thus the same as the actual data. It is the *z*-variable that is simulated - simulations based on the estimated parameters of the given null model fit to the actual study data. Thus bootstrap samples $\left(\boldsymbol{Y}, z^{*}\right)$ are generated for each of the four null models in turn.

The next stage is to create a test statistic that measures spatial variability (or non-stationarity) in regression coefficients (as characterised by the GWR model) and to use this to test against the four null hypotheses (models with stationary regression coefficients). Thus a GWR model is fitted to each bootstrap sample, where each coefficient $\boldsymbol{\beta}_j(\boldsymbol{u}_k, \boldsymbol{v}_k)$ for regression point $\boldsymbol{k}$ is calibrated for a number of different locations comprising the set $\boldsymbol{L} = \{(\boldsymbol{u}_k, \boldsymbol{v}_k)\}$ and the standard deviation of these values gives a test statistic $\boldsymbol{q}_j$ for each coefficient:

$$\boldsymbol{q}_j = \sum_{\boldsymbol{k}}^{\boldsymbol{L}} \left(\boldsymbol{\beta}_j(\boldsymbol{u}_k, \boldsymbol{v}_k) - \hat{\boldsymbol{\beta}}_j\right)^2 \Big/ \boldsymbol{L} \qquad (6)$$

where $\hat{\boldsymbol{\beta}}_j = \sum_{\boldsymbol{k}}^{\boldsymbol{L}} \boldsymbol{\beta}_j(\boldsymbol{u}_k, \boldsymbol{v}_k) \Big/ \boldsymbol{L}$. Thus, for each regression coefficient, four tests are conducted. As $\boldsymbol{q}_j$ is a measure of variability of 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 coefficient variability 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 a predictor variable 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 is entirely explained by factors other than geographical variation in regression coefficients. The aim in this case is to test whether the degree of variability in local estimates exceeds the amount expected due to situations such as that above.

The bootstrap tests are run with the number of simulations set at $\boldsymbol{R} = 999$. For each regression coefficient, the 95% points of the bootstrap samples are computed, and significance levels are found for upper single tailed hypothesis tests. For the GWR fits, an adaptive Gaussian kernel weighting is specified where the bandwidth for each bootstrap sample is automatically chosen according to an AIC minimisation approach [2]. Thus the effects of sampling variation on automatic bandwidth selection in GWR are indirectly included in the bootstrap method.

3. Case study

For a case study, the Georgia educational attainment data is used, where the 'PctBach' response *z*-variable is considered a function of the 'PctRural', 'PctEld', 'PctFB', 'PctPov' and 'PctBlack' predictor *y*-variables. This data has been shown to exhibit relationship non-stationarity via a GWR analysis [2] and such an analysis is re-run (using the same GWR specifications as that given for bootstrap samples, above). It is unnecessary to show the full set of results, but the *p*-values from a randomisation test described in [6] (with $\boldsymbol{R} = 999$), for significant spatial variation

in the localised coefficients are found as: 0.544, 0.539, 0.561, 0.005, 0.471 and 0.051 for the intercept, PctRural, PctEld, PctFB, PctPov and PctBlack, respectively. Thus there is evidence for non-stationary relationships between PctBach and PctFB; and between PctBach and PctBlack. A GW correlation analysis [2] suggests that all relationships can exhibit some degree of non-stationarity. Here the relationship between PctBach and PctBlack appears the most non-stationary, whilst the relationship between PctBach and PctRural, the least non-stationary.

Moran's *I* tests on the *z*-variable and on the error term from a MLR fit to this data, each result in significant spatial autocorrelation (with *p*-values of 0.0000 and 0.0045, respectively). This suggests that the ERR, SMA and LAG models are also suitable for this data. AIC values for the MLR, GWR, ERR, SMA and LAG models are 876.4, 824.3, 873.5, 872.7 and 873.0, respectively. Thus GWR clearly provides the most parsimonious fit, whilst the three spatially autocorrelated regressions only provide a marginal improvement over MLR. All such results suggest some value in a GWR analysis to this data. This promise can be tested further via the bootstrap test set out above, the results of which are given in Table 1. The variables PctFB, PctPov and PctBlack give significant results for all four null hypotheses. Thus the corresponding coefficients (and relationships) can be viewed as non-stationary. The situation for the intercept, PctRural and PctEld casts strong doubt that these coefficients are non-stationary. For most variables, the 95% point of the distribution of the *q*-statistic increases in the following model order: MLR, ERR, SMA and LAG. This suggests that the degree to which one might expect local estimates of coefficients to vary when a model with fixed coefficients holds, increases in this order.

Table 1. Bootstrap test *q*-statistics for the Georgia data set (significant results are bold and under-lined).

	Interc.	PctRu.	PctEld	PctFB	PctPov	PctBl.		Interc.	PctRu.	PctEld	PctFB	PctPov	PctBl.
Actual	1.737	0.012	0.116	0.996	0.111	0.047	Actual	1.737	0.012	0.116	0.996	0.111	0.047
MLR 95%	**1.321**	**0.010**	**0.110**	**0.243**	**0.044**	**0.016**	SMA 95%	2.477	0.016	0.181	**0.421**	**0.095**	**0.036**
MLR *p*	0.029	0.033	0.044	0.000	0.002	0.002	SMA *p*	0.118	0.111	0.152	0.000	0.026	0.026
ERR 95%	2.202	0.014	0.179	**0.422**	**0.084**	**0.034**	LAG 95%	3.118	0.017	0.207	**0.412**	**0.096**	**0.039**
ERR *p*	0.093	0.087	0.122	0.000	0.027	0.011	LAG *p*	0.191	0.150	0.183	0.000	0.040	0.019

4. Concluding remarks

Results suggest value in the bootstrap approach. However, this study is only an introduction. A complementary and more objective assessment, where the bootstrap is applied to simulated data sets, with known spatial properties, will be presented elsewhere. In addition, the results for an adapted test statistic $\boldsymbol{q}_j^*$ that can account for detrimental effects on coefficient estimation in GWR due to local collinearity [7], were also not given here, and instead will be presented elsewhere. Reassuringly, the test results using $\boldsymbol{q}_j$ and those using $\boldsymbol{q}_j^*$ were little different.

References

1. Brunsdon C, Fotheringham AS, Charlton M. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis* 1996: 28: 281-298.
2. Fotheringham AS, Brunsdon C, Charlton M. *Geographically Weighted Regression: the analysis of spatially varying relationships*. 2002: Chichester: Wiley.
3. Anselin L. *Spatial Econometrics* 1998: Kluwer: Dordrecht.
4. Efron B, Tibshirani RJ. Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Statistical Science* 1986: **1**: 54-77.
5. Davsion A, Hinkley D. *Bootstrap Methods and Their Application*. 1997: Cambridge Series in Statistical and Probabilistic Mathematics: Cambridge University Press.
6. Brunsdon C, Fotheringham AS, Charlton M. Geographically Weighted Regression: Modelling Spatial Non-Stationarity. *Journal of the Royal Statistical Society* 1998: **47**: 431-443.
7. Wheeler D. Tiefelsdorf M. Multicollinearity and Correlation among Regression Coefficients in Geographically Weighted Regression. *Journal of Geographical Systems* 2005: **7**: 161-287.