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 $y_1, y_2, ...y_p$ 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