Living with Collinearity in Local Regression Models

Chris Brunsdon[1]*, Martin Charlton*[2] *and Paul Harris*[2]

[1] People, Space and Place, Roxby Building, University of Liverpool, L69 7ZT, UK
Christopher.Brunsdon@liverpool.ac.uk
[2] National Centre for Geocomputation, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland
Martin.Charlton@nuim.ie; Paul.Harris@nuim.ie

Abstract

In this study, we investigate the issue of local collinearity in the predictor data when using geographically weighted regression (GWR) to explore spatial relationships between response and predictor variables. Here we show how the ideas of condition numbers and variance inflation factors may be `localised' to detect and respond to problems caused by this phenomenon. Furthermore, we introduce two adapted forms of GWR where localised regressions that are resistant to collinearity effects are specified only at locations where collinearity is considered detrimental to the standard local fit. We present initial findings via the use of a simulation study designed to assess the sensitivity of GWR outputs to various levels of collinearity. This study aims to build upon, and respond to, recent research in this area.

Keywords: Geographically Weighted Regression, Variance Inflation Factor, Condition Number, Ridge Regression

1. Introduction

The problem of collinearity amongst the predictor variables of a regression model has long been acknowledged which can lead to a loss of precision and power in the coefficient estimates. This issue is heightened in the geographically weighted regression (GWR) model (Brunsdon *et al.,* 1996), which calibrates regression models using a spatially weighted moving window to obtain localised coefficient estimates. GWR is used to investigate whether the relationship between the predictor variables and the response variable alters across space. Collinearity is an issue in GWR since: (i) its effects can be more pronounced with the smaller samples that are used to calibrate each local regression; and (ii) if the data is spatially heterogeneous in terms of its correlation structure, some localities may exhibit collinearity when others do not. In both cases, collinearity may cause problems in GWR when none are found globally.

Our aim is to understand how collinearity influences the outcome of GWR, and to suggest steps to identify undesirable influences that might occur, and if they do occur, suggest remedies. These remedies build on the geographically weighted models proposed by Wheeler (2007; 2009), following the initial highlighting of this issue presented in Wheeler and Tiefelsdorf (2005). Measurements of collinearity that we consider are: (a) matrix condition numbers; and (b) variance inflation factors (VIFs), both of which can be calculated at the same scale of each local regression fit of the GWR model. Broadly, condition numbers assess collinearity for all predictor variables together, whereas VIFs consider each predictor variable in turn. Condition numbers above 30 and VIFs above 10 are usually taken to indicate collinearity problems (see Belsey *et al.,* 1980; O'Brien, 2007).

An important linkage is between the condition number and the window size (or bandwidth) of a GWR model. Here local condition numbers will tend to increase as the bandwidth gets smaller. One remedy is to calibrate a basic GWR model with some optimally found bandwidth, but to increase this bandwidth only at locations where the local condition number is above a threshold of 30. Here the bandwidth (at such locations only) is increased until this threshold is reached. An alternative remedy is to leave the bandwidth as is, and replace the local standard regression with a local ridge regression at the same locations. In this study, we adopt both remedial routes, where we investigate simple specifications for *locally-compensated bandwidth* GWR (LCB-GWR) and *locally-compensated ridge* GWR (LCR-GWR) models, respectively. LCR-GWR directly builds upon the ridge GWR models introduced by Wheeler (2007). Ridge regression itself is a standard method to address collinearity issues in regression modelling, along with principal components regression and partial least squares regression (see Frank and Friedman, 1993, for a comparison).

Whereas condition numbers may be used to modify the local regression fit, local VIFs can be used as a diagnostic. Observing that some variables are prone to high VIFs in particular locations warns that the corresponding GWR coefficient estimates may also be suspect. Possible courses of action include: (1) omitting the variables, (2) working with a larger bandwidth, or (3) simply treating any 'interesting' patterns in these areas with caution. A further diagnostic would be to map the local correlation coefficients for all predictor data pairs, which is manageable provided the number of predictor variables is small. All such diagnostics (see Figure 1) are considered an integral part of an initial analytical toolkit that should always be employed prior to the GWR analysis itself.

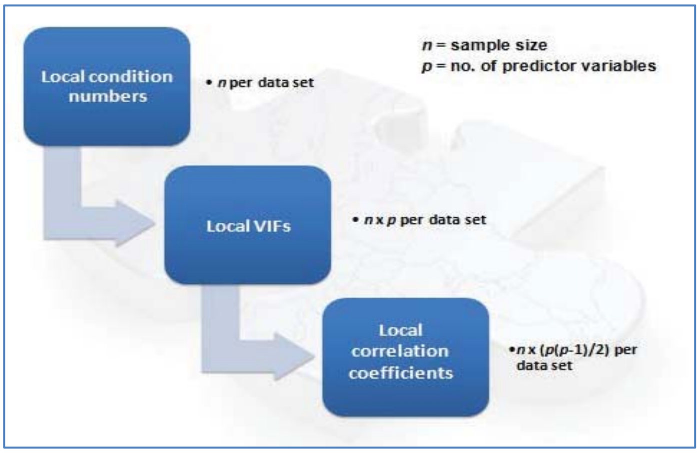
Local condition numbers
• n per data set
n = sample size
p = no. of predictor variables
Local VIFs
• n x p per data set
Local correlation coefficients
•n x (p(p-1)/2) per data set

Figure 1: Levels of complexity for different localised collinearity diagnostics.

We investigate our GWR models by means of a simulation study designed to assess the sensitivity of GWR outputs to various levels of collinearity. Issues of sample size, spatial co-dependence/dependence in the coefficients and spatial dependence in the sample data are also investigated. A key point worth emphasising is that GWR is (commonly) calibrated in terms of its predictive performance. In this respect, issues of collinearity are by design over-looked as collinearity tends not to affect prediction accuracy, but does affect the coefficient estimates.

2. Methodology

2.1. Models

We calibrate four regression models in total: (i) multiple linear regression (MLR); (ii) GWR; (iii) LCB-GWR and (iv) LCR-GWR. For predictor variables $y_1, y_2, ... y_p$ and $i = 1, \ldots, n$, MLR has this form for response variable z :

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

To estimate the parameters β , we use ordinary least squares.

The GWR model is defined as:

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

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 MLR, the ε_i 's are random error terms. Here a local regression is calibrated at i with data near to i given more influence than data further away by weighting observations according to a distance-decay function. Coefficients of the GWR model are estimated from:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = \left(\mathbf{Y}^{\mathrm{T}} \mathbf{W}(u_i, v_i) \mathbf{Y}\right)^{-\mathbf{1}} \mathbf{Y}^{\mathrm{T}} \mathbf{W}(u_i, v_i) \mathbf{z} \tag{3}$$

where $\mathbf{W}(u_i, v_i)$ is a $(n \times n)$ spatial weighting diagonal matrix; $\mathbf{Y}$ is a $(n \times (p+1))$ predictor data matrix; and $\mathbf{z}$ is a $(n \times 1)$ response data vector. To find the local coefficient standard errors, first let $\mathbf{C}_i = \left(\mathbf{Y}^{\mathrm{T}} \mathbf{W}(u_i, v_i) \mathbf{Y}\right)^{-\mathbf{1}} \mathbf{Y}^{\mathrm{T}} \mathbf{W}(u_i, v_i)$, and then evaluate the square root of $\mathrm{VAR}\left[\hat{\boldsymbol{\beta}}(u_i, v_i)\right] = \mathbf{C}_i \mathbf{C}_i^{\mathrm{T}} \sigma^2$, where the variance of the errors σ^2 is estimated from the residuals of the GWR model by:

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (z_i - \hat{z}_i)^2 \Big/ (n - (2v_1 - v_2)) \tag{4}$$

Here $v_1 = \mathrm{tr}(\mathbf{S})$ and $v_2 = \mathrm{tr}\left(\mathbf{S}^{\mathrm{T}} \mathbf{S}\right)$, where $\mathbf{S}$ is the hat matrix, which maps $\hat{\mathbf{z}}$ on to $\mathbf{z}$ by $\hat{\mathbf{z}} = \mathbf{S} \mathbf{z}$. The i^{th} row of $\mathbf{S}$, r_i is given by $r_i = \mathbf{Y}_i \mathbf{C}_i$ (where $\mathbf{Y}_i$ is a $(1 \times (p+1))$ predictor data vector at i). For this study, the weighting matrix is specified using a Gaussian kernel with a fixed bandwidth. *Leave-one-out* cross-validation is used to automatically select an optimal bandwidth.

For LCB-GWR and LCR-GWR, GWR is adapted at locations where the local condition number is above 30. In the spirit of parsimonious model development, we only present a simple form of LCB-GWR where only two bandwidths are specified: (a) bandwidth-A from the basic GWR model and (b) bandwidth-B used at locations where bandwidth-A resulted in a high condition number. Bandwidth-B is chosen so that *all* of the effected locations now have a reduced condition number, below 30.

At the same locations where bandwidth-B is specified in LCB-GWR, LCR-GWR applies a local ridge regression. For LCR-GWR, the same bandwidth as that found optimally with GWR (bandwidth-A) is used everywhere. Alternative LCR-GWR models can be calibrated using a dual optimisation of both the ridge terms and the bandwidth. More formally, collinearity in a local design matrix of a GWR model, $\mathbf{Y}^{\mathrm{T}}\mathbf{W}(u_i, v_i)\mathbf{Y}$ entails that is difficult to invert numerically, having a denominator close to zero. In turn, this causes problems in estimating $\hat{\boldsymbol{\beta}}(u_i, v_i)$. The ridge approach to this problem modifies Equation (3) by adding a displacement to the leading diagonal of the design matrix, giving this modified formula:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = \left(\mathbf{Y}^{\mathrm{T}}\mathbf{W}(u_i, v_i)\mathbf{Y} + \lambda_i \mathbf{I}\right)^{-1} \mathbf{Y}^{\mathrm{T}}\mathbf{W}(u_i, v_i)\mathbf{z} \qquad (5)$$

for some choice of λ_i. Here it is possible to specify λ_i, such that the local condition number is below our chosen threshold. For a square symmetrical matrix $\mathbf{V}$, the condition number is defined by $\kappa(\mathbf{V}) = e_1 / e_m$, where e_1 is the largest eigenvalue of $\mathbf{V}$ and e_m is the smallest. A key drawback to ridge regression is that it does not provide usable coefficient standard errors and confidence intervals (CIs).

2.2. Simulation design: a brief guide

1. Generate sample data sets of sizes: n=144, 289, 484 and 961 on square grids of 12x12, 17x17, 22x22 and 31x31, respectively.
2. For each permutation of step 1, generate three predictor variables (y_1, y_2, y_3) using geostatistical un-conditional co-simulations (e.g. see Pebesma, 2004). Vary the global correlation (collinearity) amongst the predictor variables from: (i) weak, (ii) moderate (positive only) and (iii) strong (positive only). This procedure should ensure that the number of locations with a strong degree of local collinearity should in turn be: (a) small, (b) moderate and (c) large, respectively.
3. For each permutation of steps 1-2, generate four non-stationary regression coefficient surfaces (for $\beta_0, \beta_1, \beta_2, \beta_3$) also using geostatistical un-conditional co-simulations. Vary the global correlation amongst the coefficients from: (i) weak and (ii) strong. Specify three smoothness levels for the surfaces: (a) low, (b) moderate and (c) high; via the smoothing parameter of the Matérn variogram model that is specified in the co-simulations (see Figure 2).
4. For each permutation of steps 1-3, generate the response variable z. Vary the random error term ε from: (i) relatively small and (ii) relatively large.
5. Fit the four study regression models to the resultant simulated data sets and assess the accuracy of the *estimated* regression coefficients to the *actual* regression coefficients found in step 3 (using the model diagnostics described below). Only fit models to simulated data that has a global (MLR) condition number below 30. The rationale for this exclusion, is that a usual (global) exploratory data analysis would often discard a redundant predictor variable in such cases.
6. Repeat steps 1-5 for a designated number of simulations to provide a distribution of model diagnostics for assessment. In this study, 25 simulations were run.

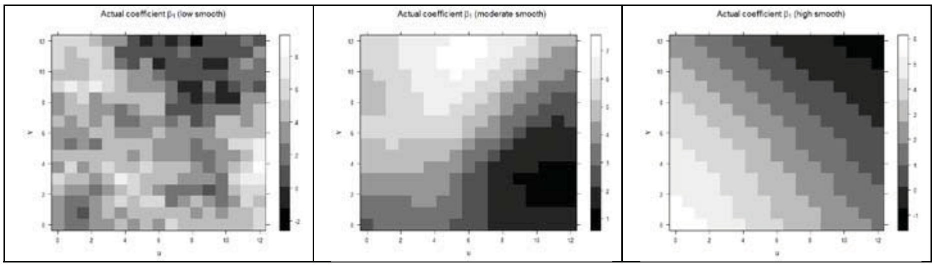

Figure 2: Example actual coeff. surfaces for β_1 with low/moderate/high smoothness levels.

2.3. Model diagnostics

To assess coefficient estimation accuracy for each model, a root mean squared error is found that is relative to the properties of the generated coefficient surfaces (relRMSE). CI accuracy for the coefficient estimates (and their standard errors) is assessed using coverage probabilities. Here a coverage probability is found for a range of symmetric CIs and then summarised by the *G*-statistic, where a value of 1 is sought. For cases, where models provide similar *G*-statistics, one model can be preferred if the mean of its CI widths that contain the actual coefficient (M-CI-W) is smaller. Observe that a model's *G*-statistic and M-PCI-W values should always be viewed in conjunction, as a strong *G*-statistic is of little use if it is coupled with a large M-PCI-W value (and vice versa); see Goovaerts (2001) for details.

3. Preliminary results

Preliminary results for weak, moderate and strong levels of predictor variable collinearity are presented in Table 1 for a sample size of n=289, where the coefficients were generated with a strong spatial correlation together with high levels of smoothness. The corresponding response variables were generated via a (relatively) small random error term. Diagnostics are averages over the 25 simulations.

Table 1: Regression coefficient estimation *and* its uncertainty accuracy.

Predictor variable collinearity	MLR	GWR	LCB-GWR	LCR-GWR
	relRMSE			
Weak	1.670	**1.316**	1.509	1.376
Moderate	2.197	1.880	2.124	**1.594**
Strong	2.264	2.486	2.262	**1.676**
	G-Statistic			
Weak	0.155	0.579	0.441	N/A
Moderate	0.282	0.648	0.447	N/A
Strong	0.263	0.727	0.476	N/A
	M-CI-W			
Weak	0.609	1.759	1.421	N/A
Moderate	0.774	2.516	1.619	N/A
Strong	1.244	4.078	3.040	N/A

With respect to coefficient estimation accuracy (relRMSE), GWR marginally performs the best when predictor variable collinearity is weak. When this collinearity is moderate or strong, LCR-GWR clearly performs the best. Differences in performance between GWR and LCR-GWR increase as levels of collinearity increase. LCB-GWR performs poorly and requires a more sophisticated form. With respect to coefficient CI accuracy (*G*-statistic and M-CI-W), all models perform poorly, with MLR performing the poorest. As the strength of collinearity increases, CI accuracy is considered to weaken in all cases (any increase in the *G*-statistic is offset by a corresponding increase in the M-CI-W value).

4. Conclusion

Preliminary results has shown promise in a locally-compensated GWR model in the presence of collinearity. Further work is currently being undertaken to more fully understand the true value of these models and under what conditions they should be applied. Models with autocorrelated error are also under consideration.

Acknowledgments

For Charlton and Harris, research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by the Science Foundation Ireland under the National Development Plan.

References

Belsey, D., Kuh, E., Welsch, R., (2004), *Regression Diagnostics: identifying influential data and sources of collinearity*, Hoboken: Wiley.

Brunsdon, C., Fotheringham, A.S., Charlton, M., (1996), "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity", *Geographical Analysis*, 28(4), 281-298.

Frank, I., Friedman, J., (1993) "A statistical view of chemometrics tools", *Technometrics,* 35, 109-135.

Goovaerts, P., (2001) "Geostatistical modelling of uncertainty in soil science", *Geoderma* 103, 3-26.

O'Brien, R., (2007) "A caution regarding rules of thumb for variance inflation factors", *Quality & Quantity,* 41, 673-690.

Pebesma, E., (2004) "Multivariate geostatistics in S: the gstat package" *Computers & Geosciences,* 30:683-691.

Wheeler, D., (2007), "Diagnostic tools and a remedial method for collinearity in geographically weighted regression" *Environment and Planning A,* 39(10), 2461-2481.

Wheeler, D., (2009) "Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso", *Environment and Planning A,* 41:722-742.

Wheeler, D., Tiefelsdorf M., (2005), "Multicollinearity and correlation among local regression coefficients in geographically weighted regression" *Journal of Geographical Systems,* 7, 161-187.