Living with Collinearity in Local Regression Models

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

[1]People Space and Place, Roxby Building, University of Liverpool,L69 7ZT , UK
Tel. +44 151 794 2837
Christopher.Brunsdon@liverpool.ac.uk
[2]National Centre for Geocomputation, National University of Ireland,
Maynooth,Co. Kildare, IRELAND

Summary: We investigate the issue of collinearity in data when using Geographically Weighted Regression to explore spatial variation in data sets – and show how the ideas of condition numbers and variance inflation factors may be `localised' to detect and respond to problems caused by this phenomenon.

KEYWORDS: Geographically Weighted Regression, Collinearity, Variance Inflation Factor, Condition Number, Model Diagnostics

1. Introduction

The problem of collinearity in regression models has long been acknowledged. In general if a multivariate linear regression model has a *response* variable **y** and a matrix of column predictor variables **X**, with a regression model of the form $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ where $\boldsymbol{\beta}$ is a vector of coefficients and $\boldsymbol{\varepsilon}$ is a vector of independent Gaussian error terms with variance $\sigma^2\mathbf{I}$ and zero mean, then there are often problems encountered when attempting to estimate $\boldsymbol{\beta}$ if any of the variables of **X** have a high degree of correlation, or are close to exhibiting a deterministic linear relationship. Collinearity has a number of adverse effects on the estimation of the regression coefficients include loss of precision and power. In designed laboratory experiments collinearity can be often avoided by design – the columns of **X** frequently correspond to quantities such as concentration of a some chemical, or drug, and so levels can be controlled, and therefore chosen in advance. In this situation, values are selected to avoid such linear dependencies – indeed **X** may be chosen so that each column has zero correlation to the others. However, researchers studying spatial data do not generally have this luxury – both social and physical geography often require observations to be made *in situ* without any way of directly influencing the values of **X**. Thus, the issues of collinearity outlined above may be unavoidable and therefore they are particularly pertinent in this situation.

This issue becomes even more relevant when considering the use of Geographically Weighted Regression (GWR) (Brunsdon *et al,* 1996). This technique essentially operates by calibrating regression models using a moving spatially weighted window – so that localised estimates of $\boldsymbol{\beta}$ can be obtained. This is a useful tool for exploring whether the relationship between the predictor variables in **X** and the response variable **y** alters across space. Collinearity can be an important issue because

- The localised data samples may be fairly small if the size of the geographical window is also small. The effects of collinearity can be more pronounced with smaller samples.
- 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 even if none are apparent when fitting a global regression model.

Thus, the aim here is to gain understanding of the way that collinearity influences the outcome of GWR, and to suggest steps that can be taken to identify any undesirable influences that might be occurring, and if so how they may be remedied. In the next sections we will outline some of the

approaches to this – and give a practical example of how these may be applied to a real-world data set used to investigate voter turnout in the 2004 Irish General Election, in the Dublin area. A key point is that existing methods to calibrate GWR choose parameters in terms of predictive performance – collinearity tends not to affect this but it does affect the parameter estimates. The approaches outlined here are intended to address performance in the latter issue.

2. Identifying Collinearity

One key aspect of the collinearity issue is measuring the *degree* of collinearity that exists in a given data set. Fortunately, much work has already been done in this area. The key modification for GWR is to adapt these ideas to work on the same localised moving window approach as GWR itself. Key measurements of collinearity are considered below.

2.1 The Condition Number

Typically, global collinearity is measured using the *condition number* of the matrix $\mathbf{X^TX}$, defined to be the ratio of the largest to the smallest eigenvalue of that matrix. If any fully collinear relationship existed within the columns of $\mathbf{X}$ then the smallest eigenvalue would be zero, and if the relationship is very nearly collinear (i.e. a linear relationship holds between some columns in $\mathbf{X}$ with only minor residuals) then this eigenvalue is very *close* to zero, and so the condition number is very large. This can be adapted to assess *local* collinearity in the GWR context by replacing $\mathbf{X^TX}$ with $\mathbf{X^TWX}$ in the definition of the condition number – where $\mathbf{W}$ is a diagonal matrix whose values w_{ii} are the weights applied to the observations to create the locally weighted window – so that $\mathbf{W}$ varies with location. In doing this, there is a condition number associated with every point in the study area at which GWR coefficients are estimated.

An important linkage here is between the condition number and the *bandwidth* of a GWR model. The latter is essentially the radius of the moving window used in the GWR. For example, a typical weighting scheme might be as below:

$$w_{ii} = \begin{cases} \left(1-\frac{d_i^2}{h^2}\right)^2 \text{ if } d_i^2 < h^2 \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

where d_i is the distance from observation i to the location at which the GWR is calibrated, and h is the bandwidth. At a given location, there is a deterministic relationship between the bandwidth and the condition number – for example Figure 1 shows the relationship between the two quantities for a model used to explore voting patterns in Dublin in 2004 in relation to a number of Irish Census derived variables. The variables are listed in Table 1, and apply to 323 Dublin enumeration districts (EDs).

Table 1. Variables used in GWR model

Variable	Units
Voter Turnout (***y*** variable)	% Voting population
Different Address 1 Year Ago	% Population
Local Authority Renting	% Households
Head of Household Social Class 1	% Households
Unemployed	% Population
Low Education Level	% Population
Age 18-24	% Population
Age 25-44	% Population
Age 45-64	% Population

The relationship is monotone, with the condition number increasing as the bandwidth gets smaller. In short, if the bandwidth is too small, a high degree of collinearity may result. Both Myers (1986) and Belsey *et al*. (2004) suggest that condition numbers above around 30 indicate regression calibration problems – in Figure 1 it can be seen that this happens when the bandwidth is less than around 3km in this particular example. One remedy is to work with adaptive bandwidths as set out in Fotheringham *et al*. (2002), where the bandwidth is chosen to match the *n*th nearest point to the regression point, but toapply a further rule, where if the bandwidth selected in this way leads to a condition number below a threshold (here, we choose 20 to ensure values are well below the problematic value of 30), then the bandwidth is increased until the threshold is reached. This is relatively easy to achieve computationallyrequiring the numerical solution for *h* in the equation

$$\kappa(h) = 20 \tag{2}$$

where $\kappa(.)$ denotes the function mapping bandwidth *h* to the condition number.

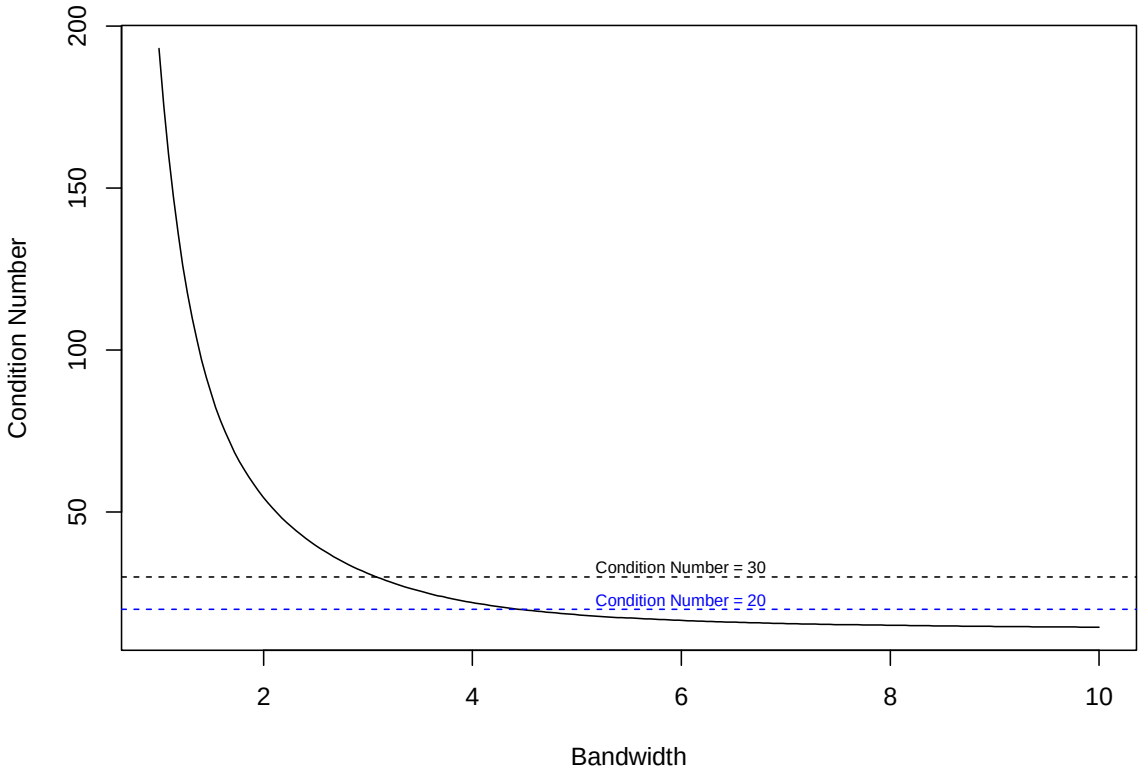

Figure 1. Relationship between bandwidth (km) and Condition Number

2.2 Variance Inflation Factors (VIFs)

An alternative measure of the effects of collinearity is the variance inflation factor (VIF) – unlike the condition number which assesses the whole model, VIFs consider each variable in turn. Essentially they estimate the degree to which the sampling variance of an individual parameter estimate is amplified by the collinearity in **X**, in comparison to an ideal situation in which all columns of **X** are uncorrelated. See, for example, Hair *et al* (2006) – as a general rule VIFs that exceed 10 are

regarded as potentially problematic. As before, it is possible to compute VIFs in a weighted, localised framework – in this case we obtain a map for each variable in a regression model.

Whereas the condition numbers may be used to modify the local bandwidth , VIFs can be used as a diagnostic. Observing that some variables are prone to very high VIFs in particular locations warns that the local estimates of the corresponding GWR coefficients may be suspect. One possible course of action then is to remove these variables from the model. However, as they may actually be associated with the dependent variable it is very difficult to calibrate their regression coefficients, and their presence also may be detrimental to the estimation of the coefficients for other variables. An alternative approach, is to attempt to increase the bandwidth until further problematic VIFs are eliminated.

For the Dublin data, the number of VIFs exceeding 10 are tabulated against the number of nearest neighbours in a nearest-neighbour based bandwidth selection model in Table 2.

Table 2. Variance Inflation Factors for the Dublin data described in the text. The numbers are counts of areas with a VIF exceeding 10 for each variable, as the number of nearest neighbours in the adaptive GWR increases.

No. Near Neigh-bours	Different Address 1 Year Ago	LA Renting	Social Class 1	Unem-ployed	Low Education Level	Age 18-24	Age 25-44	Age 45-64
25	74	62	52	108	23	32	22	9
50	20	19	7	35	5	6	3	1
75	10	12	2	17	1	1	0	0
100	1	10	0	10	1	1	0	0
125	0	8	0	8	1	1	0	0
150	0	5	0	5	0	0	0	0
175	0	4	0	5	0	0	0	0
200	0	3	0	4	0	0	0	0
225	0	1	0	1	0	0	0	0
250	0	0	0	1	0	0	0	0
275	0	0	0	0	0	0	0	0
300	0	0	0	0	0	0	0	0

From this it can be seen that when the number of nearest neighbours exceeds around 150, only the local authority renting and unemployment variables have problematic VIFs, and only in the same locations. Mapping these for the number of nearest neighbours being 150 yields the maps in Figure 2. These indicate that the problem areas lie on the southern edge of the study area. This leaves a number of potential remedies: either omit the variables, work with a larger number of nearest neighbours, or simply treat any ‘interesting’ patterns in these areas with caution.

3. The Talk

In the presentation, we will outline the above ideas, and in addition show by means of simulation the

degree to which GWR is affected by collinearity. We will also consider some further remedies to the problem, including those of Wheeler (2007), in combination with our own approaches.

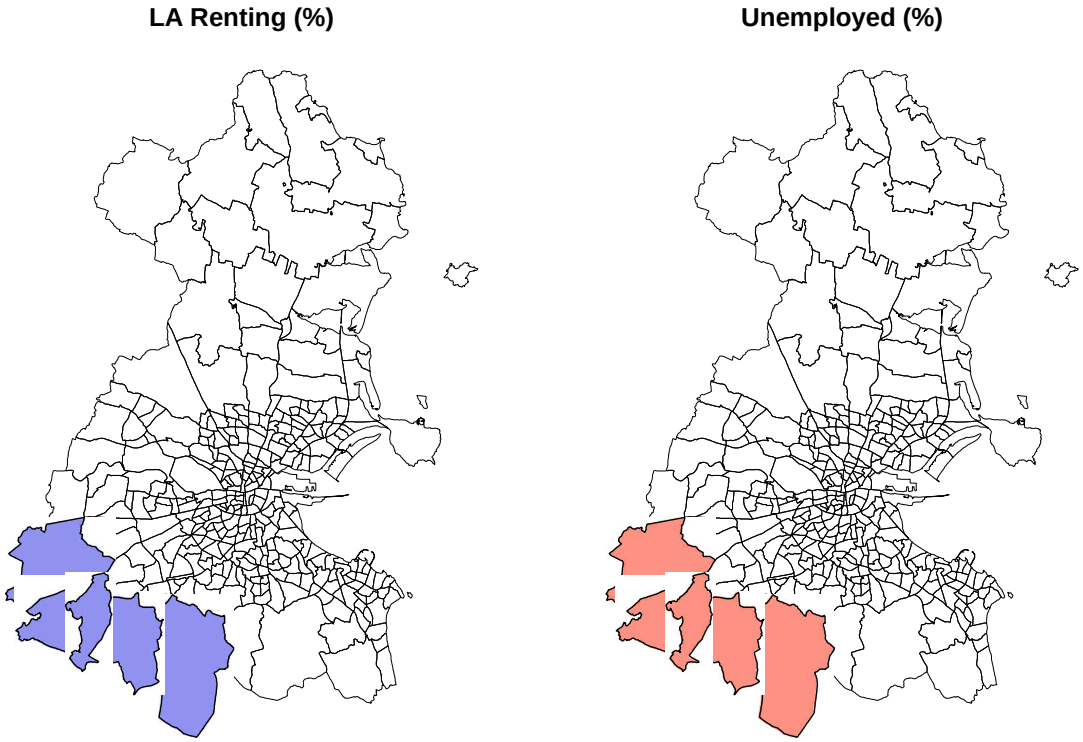

Figure 2. Maps showing areas in which the VIF exceeds 10 when using a 150th nearest neighbour-based GWR.

4. Acknowledgements

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. The authors gratefully acknowledge this support.

5. References

Belsey DA, Kuh E and Welsch RE, 2004, *Regression Diagnostics: identifying influential data and sources of collinearity*, Hoboken: Wiley

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

Fotheringham AS, Brunsdon C and Charlton ME (2002). *Geographically Weighted Regression - the analysis of spatially varying relationships*. Wiley, Chichester.

Hair JF, Anderson R, Tatham RL and Black WC (2006). *Multivariate Data Analysis*. Prentice Hall: Upper Saddle River, N.J. 2006.

Myers RH (1986). *Classical and Modern Regression with Applications*, Boston: Duxbury Press.

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

8. Biography

Chris Brunsdon is Professor of Human Geography at the University of Liverpool. One of the

developers of Geographically Weighted Regression he is a leading researcher in geographical information science.

Martin Charlton is Senior Research Associate in the National Centre for Geocomputation at the National University of Ireland Maynooth. Together with Chris Brunsdon and Stewart Fotheringam, he is one of the developers of Geographically Weighted Regression.

Paul Harris is a Postdoctoral Research Fellow at the National Centre for Geocomputation and has a PhD in Geostatistics from Newcastle University.
All three of them are leading practitioners in the ancient and noble art of pie-shifting.