

# Developing Grid enabled spatial regression models

Richard Harris<sup>1</sup>, Min hua Jen<sup>2</sup>, David Kilham<sup>3</sup>, Edward Thomas<sup>4</sup>, Chris Brunsdon<sup>5</sup>, Claire Jarvis<sup>6</sup>

<sup>1-4</sup>School of Geographical Sciences, University of Bristol, UK

<sup>5-6</sup>Department of Geography, University of Leicester, UK

Email address of corresponding author: M.Jen@bristol.ac.uk

**Abstract.** Various methods of spatial analysis have been developed to detect and to explain geographical ‘hot spots’ and clustering (in data), undertaking localized and non-parametric statistical testing to assess the significance of a spatial pattern. They do so by sequential repeat testing, creating computational demands as the algorithm cycles through spatial subsets of the data and then again, repeatedly, as the data randomly are redistributed across geographical locations within the study region. In this paper we introduce one such method known as Geographically Weighted Regression (GWR) and give the rationale for adapting it to the computational ‘grid’ infrastructure of e-social science. The aim of the project is to run GWR’s sequences of calibration, analysis and significance testing in parallel, applying the processes to 2001 Census data to produce a geographically calibrated index of deprivation. The research is funded by the ESRC’s small grant scheme for e-social science.

## Introduction

The use of linear regression modelling and its various variants is widespread in the geographical and other social sciences. Yet, the method is also known to present something of a geographical paradox. Generally, it makes little sense to look for interesting geographies of the relationships between a response variable,  $Y$ , and a matrix of predictor variables,  $X$ , using a method that explicitly assumes spatial independence of the regression residuals to do so. Whilst standard practice is to map the residuals and check the assumption of spatial independence has not been violated, as an approach for geographical enquiry it is unsatisfactory because if patterns of spatial autocorrelation are found, then, not only will the (global) model be biased, it will also offer little statistical explanation as to why the local variations have occurred. If the model subsequently was used to predict the attribute  $Y$  at other (non-sampled) locations it would risk giving erroneous values, especially if important local variations in relationships between the regression variables had been ‘averaged away’ by attempting a global fit.

Fortunately, the proliferation of georeferenced datasets and greatly enhanced computer infrastructures has freed spatial analysis from the constraints of traditional linear regression (and its variants) to model the effects of spatial autocorrelation and local contingencies in understanding spatial patterns and processes. One method, geographically weighted regression (GWR), assumes relationships between regression variables may change over space. Whereas global analyses treat local variations as ‘noise’, GWR specifically allows important local relationships to be measured and mapped.

The aim of the research outlined in this paper is to produce a prototype but functional, grid-enabled version of GWR that is open source and available to other researchers to run on the National Grid Service, and which is used to produce a geographically sensitive, GWR-enabled index of the correlates of deprivation across the UK. For the rest of this paper we: outline the GWR method; give an example of its application to modelling income; describe why it – as an example of spatial analysis – can benefit from a grid approach; and outline some suggestions as to how a parallelised implantation might be achieved. The research is, at this stage, embryonic and entirely theoretical: at the time of writing the project is less than two weeks old!

## About GWR

GWR builds on the intuitive nature of traditional, linear regression modelling and is based on the premise that relationships between variables measured at different locations might not be constant across geographical space. GWR offers an explicitly geographical, relatively simple but theoretically informed incorporation of local spatial relationships within the rubric of regression analysis. It is easily run on publicly available software (see <http://ncg.nuim.ie/ncg/GWR/>) and also using the *spgwr* package for the open source statistical programming language ‘R’ (Venables *et al.*, 2001; Bivand, 2006).

At its simplest, GWR can be understood as treating the global regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon \quad (1)$$

as a special case of the model

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \quad (2)$$

where the difference between (1) and (2) is that the former is spatially invariant whereas, in (2),  $\beta_k(u_i, v_i)$  is a realisation of the continuous function  $\beta_k(u, v)$  at point  $i$  (Fotheringham *et al.*, 2002).

To estimate  $y_i$ , a kernel of radius  $r_i$  is positioned around location  $(u_i, v_i)$  and a regression model fitted to the  $n_i$  number of observations found within distance  $r_i$  from  $i$ . This process is then repeated at other locations within the study region. The name of geographically weighted regression arises because each observation’s contribution to the local regression model is weighted ( $w_j$ ) as a function of the inverse of the distance from it ( $j$ ) to  $i$ . It is localized, because beyond distance  $r_i$  each observation is weighted as contributing zero. For example, where  $d_{ij}$  is the distance between  $i$  and  $j$ :

$$w_j = \begin{cases} \exp\left[-\frac{1}{2}\left(d_{ij}/r_i\right)^2\right] & d_{ij} < r_i \\ 0 & d_{ij} \geq r_i \end{cases} \quad (3)$$

which creates a Gaussian shaped kernel

Other weighting functions can be used by the actual choice has rather less effect on the estimation of  $y_i$  than the choice of bandwidth,  $r$ . The bandwidth can be defined in various ways, including the distance from  $i$  to its  $N$ th neighbour. Since this distance will vary by location so determining the bandwidth by  $N$  leads to a kernel that adapts to the local density

of observations around  $i$ . This is useful in geographic applications where, for example, population density rarely is uniform across space.

As Fotheringham *et al.* (*op. Cit.*) note, GWR will not lead to an unbiased estimate of the local regression coefficients because the outcome of the non-stationary process at location  $i$  is inferred from data collected at proximate locations other than  $i$ . There is a trade-off between bias and standard error because if, as the method presumes, the regression coefficients vary continuously over space then that variation must also exist within the spatial extent of the kernel placed over and around  $i$ . On this basis, the solution would seem to be to limit the size of  $N$  (or  $r_i$ ). However, this needs to be balanced by the fact that as  $N$  – the sample size – is decreased then so the standard errors of the estimates increase. Calibration of GWR is therefore required, for example by iterating through values of  $N$  until the (squared) sum of differences between the observed and predicted values of the data points is minimized.

Calibrating in this way also permits the premise of spatial autocorrelation to be tested. This premise is not unique to GWR but underpins a whole swathe of spatial and geostatistical techniques that build on Tobler's 'First Law of Geography': everything is related to everything else, but near things are more related than distant things (Tobler, 1970). These include various types of point pattern analysis (Diggle, 2003) and geostatistical interpolation based upon kriging (Isaaks & Srivastava, 1990). Where GWR differs from these techniques is by couching the spatial component in a more traditional and widely understood regression framework for statistical analysis and explanation.

## A GWR model of income

GWR has been used to show that even within a single city (Bristol England) the determinants of poverty vary spatially (Longley & Tobón, 2004). Further variations can be expected at more aggregate scales (e.g. between cities, between regions or between rural and urban settlements). However, standard measures of deprivation tend not to be sensitive to these local or regional differences. Consider the simple regression model

$$y = \beta_0 + \sum_{k=1}^4 \beta_k x_k + \varepsilon \quad (4)$$

where the parameters and model diagnostics are as in Table 1, having been obtained based on a stepwise procedure applied to a collection of 84 census variables describing socio-economic and demographic conditions, and used to predict net household weekly income in the 107 census wards of the three contiguous unitary authorities of Bristol, Bath and North Somerset, and South Gloucestershire (total population: 795256 persons; 332243 households), each located in the South West of England. An intriguing property of this model is the selection of the 'born in Scotland' variable as a significant predictor of net household income in the South Western wards. Given the distance between the study region and Scotland (about 300 miles), it may, in fact, be a confounded relationship (i.e. the 'born in Scotland' variable is correlated with one or more other variables that have a more obvious relationship with income but are not included in the model). In any case, it seems sensible to probe further and ask 'is the relationship between the proportion of Scots and weekly income a constant across the region or does it display geographical variation?'

To answer the question, a GWR model is fitted, centring the localized regression kernel on the geographical centres of each of the 107 census wards, in turn. The analysis reveals there to be significant spatial variability in the Scottish variable (see the last column of Table I): the

proportion of all people born in Scotland is a significant predictor of net weekly income across the study region but it has greater and more significant effect towards the centre-west of the region (the City of Bristol) – Figure 1. The point is not the (somewhat curious) relationship between Scots and household income within the study region but that predictors of income (and, by extension, deprivation) can have spatial variability that need to be considered.

Table I. Regression model of net weekly income in census wards of Bristol, Bath and North Somerset and South Gloucestershire

	Variable definition	Source	$\beta_k$	t value	p value	Significance	
						Global	Spatial <sup>o</sup>
Y	Estimate of net household weekly income (£) (2001/2)	Office for National Statistics (ONS)					
$x_{k=0}$	Intercept		343	18.82	0.000	***	
$x_{k=1}$	Proportion of households with two cars or vans	ONS: 2001 Census	447	17.25	0.000	***	
$x_{k=2}$	Proportion of people aged 16-74 in employment working in elementary occupations	ONS: 2001 Census	-539	-7.65	0.000	***	
$x_{k=3}$	Proportion of all people born in Scotland	ONS: 2001 Census	2629	5.33	0.000	***	**
$x_{k=4}$	Proportion of all people of 'other' ethnic group <sup>Δ</sup>	ONS: 2001 Census	-2779	-3.73	0.000	***	

R-sq. (adjusted) = 90.7%  
 \*\*\* significant at .1% level  
 \*\* significant at 1% level  
 \* significant at 5% level  
 Δ not White, British, Black British, Asian, Asian British, Mixed or Chinese  
 o significance of the spatial variability of parameters (estimated by GWR: see text)

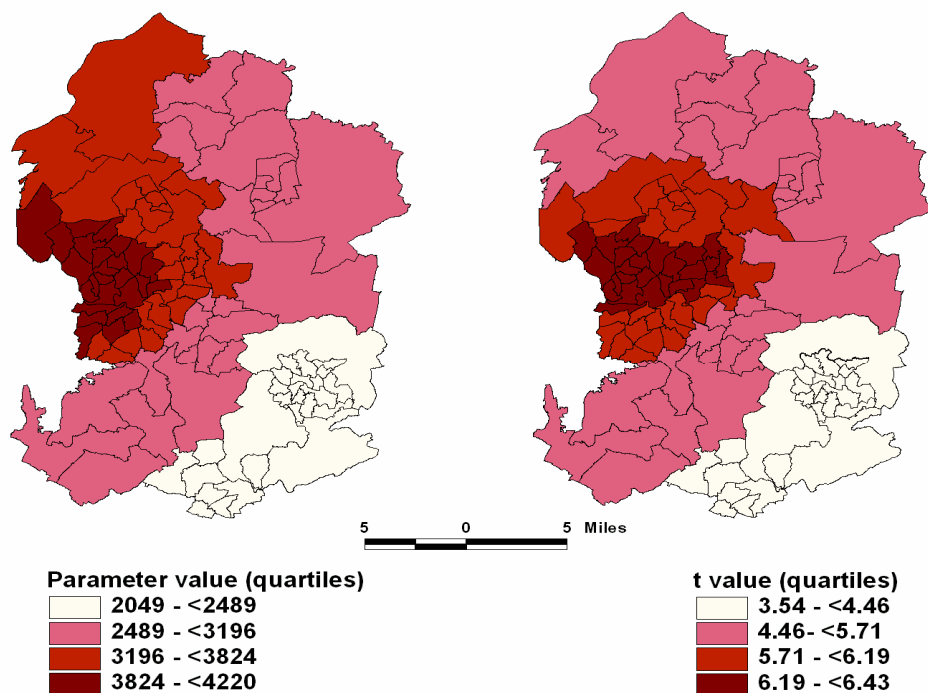


Figure 1. GWR fitted Scottish parameter and t values across the study region

## Spatial statistics and high performance computing

The application of high performance computing to spatial analysis has long been of interest to geographical scientists and has spearheaded research in geocomputation. Of particular note is the pioneering work undertaken by Stan Openshaw at the University of Newcastle and at the Centre for Computational Geography at Leeds University, of which an exemplar is the ‘Geographical Analysis Machine’ (GAM: Openshaw *et al.*, 1987). Although different in their statistical approaches (GWR is more firmly rooted in the conventions of regression modelling), what GAM, GWR and other methods of spatially localized analysis have in common is a general sequence of: (a) calibrating the size of the kernel or search window to the amount of spatial autocorrelation found in the attributes of the data being examined; (b) creating spatially overlapping subsets of the data to reflect this; (c) allowing the kernel to pass from one subset to the next, applying a statistical test in each; and (d) simulating confidence intervals for the statistical result by detaching the data attributes from the geographical coordinates at which they were captured, then repeatedly reattaching the attributes to randomly selected locations and applying the test again.

If a study region is divided into  $n$  overlapping grid squares then the first and third stages of the sequence are completed by allowing the kernel to expand from a minimum to a maximum width through  $z$  increments and determining an optimal result. This requires  $n \times z$  processes. For the fourth stage (and assuming the kernel size is now fixed), the data are redistributed  $m$  times, requiring a further  $n \times m$  tests. In total, then, the method invokes approximately  $n(z + m)$  processes (plus others to subset the data, run the statistical tests and so forth). The attraction of high performance computing and, in particular, the use of parallelisation, arises from the course granularity of the overall sequence of events (granularity being the size of computation that can be performed between communication or synchronization points: Lumb, 2004 after Wilkinson & Allen, 1999). For many spatial statistical procedures, each of the stages of calibration, fitting and assessing significance can be parallelised with processes that will operate without communication to the others (since, for example, the outcome of a model fitted to one spatial subset of the data does not affect or modify the outcome of a model fitted to another). In principle, each of the  $n(z + m)$  processes can be sent to separate computational nodes; their outputs need only be pooled and assessed once the results have been established.

## Towards running GWR on the National Grid Service

The current focus of the GWR-Grid project is on feasibility testing and thinking carefully about the best path to take for enabling GWR on the National Grid Service (NGS). There are a number of options, reflecting the availability of GWR source code in R and Fortran 77, and the potential to recode (in C/C++, for example). Of these, the R route is attractive: in part because an implementation of GWR already is available as an R package (*spgwr*: see [cran.r-project.org](http://cran.r-project.org) for details); in part because of potential research synergies with the NCESS pilot project ‘SABRE in R’ that aims to develop a serial and parallel implementation of SABRE (a program for the statistical analysis of binary, ordinal and count recurrent events) and to make these freely available as R Objects (see [www.ncess.ac.uk/research/](http://www.ncess.ac.uk/research/)); and in part because there is a further R package called *snow* which can be used for simple parallel computing in R. It does so by using a master/slave design – the master R process creates a cluster of slave R processes with communication using a socket interface, PVM or MPI (Rossini *et al.*, 2003).

It would be extremely fortuitous if these various packages could be bolted together to enable GWR on the NGS. Unfortunately, life rarely is simple and so it may be preferable to recode the GWR algorithm (or use the Fortran code), writing it around the task and NGS

environment, therefore minimising reliance on third party components. Given the early stage of the project, comments and experience that might contribute to our decision making warmly are invited.

## Acknowledgments

The research is funded by ESRC grant number RES-149-25-1041. Please see [www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Minisite/gwr/](http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Minisite/gwr/) for further details and developments.

## References

- Bivand, R. S. (2006): 'Implementing spatial data analysis software tools in R', *Geographical Analysis*, vol. 38, no. 1, pp. 23–40.
- Diggle, P. J. (2003): *Statistical Analysis of Spatial Point Patterns*, Arnold, London.
- Fotheringham, A. S., Brunson, C. and Charlton, M. (2002): *Geographically Weighted Regression: the analysis of spatially varying relationships*, Wiley, Chichester and New York.
- Isaaks, E. H. and Srivastava, R. M. (1990): *Applied Geostatistics*. Oxford University Press Inc, USA.
- Longley, P. A. and Tobón, C. (2004): 'Spatial Dependence and Heterogeneity in Patterns of Hardship: an intra-urban analysis', *Annals of the Association of American Geographers*, vol. 94, pp. 503-19.
- Lumb, I. (2004): 'HPC Grids', in A. Abbas (ed.): *Grid Computing: a practical guide to technology and applications*, Charles River Media, Hingham, MA, 2004, pp. 119-33.
- Openshaw, S., Charlton, M., Wymer, C. and Craft, A. W. (1987): 'A Mark I Geographical Analysis Machine for the automated analysis of point datasets', *International Journal of Geographical Information Systems*, vol. 1, pp. 335-58.
- Rossini, A., Tierney, L. and Li N. (2003): Simple Parallel Statistics in R, University of Washington Working Paper Series, The Berkley Electronic Press, [www.bepress.com/uwbiostat/paper193](http://www.bepress.com/uwbiostat/paper193)
- Tobler, W. R. (1970): 'A computer movie simulating urban growth in the Detroit region', *Economic Geography*, vol. 46, pp. 234-40.
- Venables, W. N., Smith, D. M. and R Development Core Team (2001): *An Introduction to R*, Network Theory Ltd., Bristol.
- Wilkinson, B. and Allen, M. (1999): *Parallel Programming: techniques and applications using networked workstations and parallel computers*, Prentice Hall, Upper Saddle River, NJ.