GEOGRAPHICALLY WEIGHTED REGRESSION

WHITE PAPER

MARTIN CHARLTON

A STEWART FOTHERINGHAM

National Centre for Geocomputation

National University of Ireland Maynooth

Maynooth, Co Kildare, IRELAND

March 3 2009

The authors gratefully acknowledge support from a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan.

sfi
science foundation ireland
fondúireacht eolaíochta éireann

Contents

Regression

Regression encompasses a wide range of methods for modelling the relationship between a dependent variable and a set of one or more independent variables. The dependent variable is sometimes known as the y-variable, the response variable or the regressand. The independent variables are sometimes known as x-variables, predictor variables, or regressors. A regression model is expressed as an equation.

In its simplest form a linear regression model can take the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{for } i=1 \ldots n$$

In this equation y_i is the response variable, here measured at some location *i*, x_i is the independent variable, ε_i is the error term, and β_0 and β_1 are *parameters*[1] which are to be estimated such that the value of $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is minimised over the *n* observations in the dataset. The $\hat{y}_i$ is the predicted or *fitted value* for the *i*th observation, given the *i*th value of x. The term $(y_i - \hat{y}_i)$ is known as the *residual* for the *i*th observation, and the residuals should be both independent and drawn identically from a Normal Distribution with a mean of zero. Such a model is usually fitted using a procedure known as Ordinary Least Squares (OLS).

More generally, a multiple linear regression model may be written:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_m x_{mi} + \varepsilon_i \qquad \text{for } i=1 \ldots n$$

where the predictions of the dependent variable are obtained through a linear combination of the independent variables. The OLS estimator takes the form:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

where $\hat{\beta}$ is the vector of estimated parameters, X is the design matrix which contains the values of the independent variables and a column of 1s, *y* is the vector of observed values, and $(X^T X)^{-1}$ is the inverse of the variance-covariance matrix.

[1] The term *coefficient* is sometimes used instead of parameter.

Sometimes it is desirable to weight the observations in a regression, for example, different levels of data uncertainty. The weights are placed in the leading diagonal of a square matrix W and the estimator is altered to include the weighting:

$$\hat{\beta} = (X^T W X)^{-1} X^T W y$$

The ability of the model to replicate the observed y values is measured by the goodness of fit. This is conveniently expressed by the r^2 value which runs from 0 to 1 and measures the proportion of variation in the observed y which is accounted for (sometimes "explained by") by variation in the model. The r^2 can often be increased merely by adding variables, so the adjusted r^2 is often reported – the adjustment takes into account the number of independent variables in the model and reflects model parsimony.

In a regression model we may want to determine whether the value of a parameter is sufficiently different from zero so that the changes in the variable to which it is attached will influence changes in the predictions. To determine whether variables contribute significantly to the model in this way, we divide the parameter estimate for each variable by its standard error. The resulting statistics have a t-distribution and may be compared with critical values from a t distribution, given the number of degrees of freedom in the model.

Regression with Spatial Data

There are a number of assumptions underlying the basic regression model described here, one of which is that the observations should be independent of one another. This is not always the case with data for spatial units and Tobler's observation that “"Everything is related to everything else, but near things are more related than distant things." (Tobler, 1970) can be recalled. Not only might the variables in the model exhibit spatial dependence (that is, nearby locations will have similar values) but also the model's residuals might exhibit spatial dependence. The latter characteristic can be observed if the residuals from the basic regression are plotted on a map where commonly the residuals in neighboring spatial units will have a similar magnitude and sign.

These characteristics of spatial data have implications for the estimates of the parameters in the basic model. If there is spatial structure in the residuals from the model, this will lead to *inefficient* estimates of the parameters, which in turn means that the standard errors of the parameters will be too large. This has implications for inference where potentially significant parameter estimates may appear not to be so. Spatial structure in the data means that the value of the dependent variable in one spatial unit is affected by the independent variables in nearby units. This leads to parameter estimates which are both *biased* and inefficient. A biased estimates is one that is either too high or too low as an estimate of the unknown true value.

Anselin (1988) describes model forms to deal with these cases. A *spatial error* model is appropriate when there appears to be structure in the residual term, and a *spatial lag* model is appropriate when spatial structure is present in the variables in the model. Unbiased parameter estimates can be found from both model types when maximum likelihood is used as the fitting method.

Spatial heterogeneity is another phenomenon in spatial modelling. It is assumed when fitting all of the regression models described above that the relationships being modelled are the same everywhere within the study area from which the data are drawn. This assumption is referred to as one of *homogeneity*. However, there is often good reason to question whether this assumption when dealing with spatial data as the processes generating them might vary across space. This condition is referred to as *spatial heterogeneity* An early example of a regression model which attempts to deal with spatial heterogeneity is the spatial expansion method (Casetti 1972). Parameters in such models are themselves functions of location where the user determines the nature of the function (usually some linear polynomial). For example, if we take a simple model with two independent variables, x_1 and x_2, with three parameters a, b, and c, that are to be estimated, i.e.:

$$y_i = a + bx_{1i} + cx_{2i}$$

We then expand the parameters so that they are some linear function of the locations (u_i,v_i) of the observations. For example, if a 1st order linear polynomial ised specified, then:

$$a_i = \alpha_0 + \alpha_1 u_i + \alpha_2 v_i$$
$$b_i = \beta_0 + \beta_1 u_i + \beta_2 v_i$$
$$c_i = \gamma_0 + \gamma_1 u_i + \gamma_2 v_i$$

The full expansion model itself can then be re-written as:

$$y_i = \alpha_0 + \alpha_1 u_i + \alpha_2 v_i + \beta_0 x_{1i} + \beta_1 u_i x_{1i} + \beta_2 v_i x_{1i} + \gamma_0 x_{2i} + \gamma_1 u_i x_{2i} + \gamma_2 v_i x_{2i}$$

This is easily fitted using OLS, as the analyst only needs to supply the extra terms $u_i x_{1i}$, $v_i x_{1i}$, $u_i x_{2i}$, and $v_i x_{2i}$ in the model. As the locations of the observations are also known, the spatially varying parameter estimates are easily computed and mapped. A disadvantage of this model is that the analyst must determine the nature of the parameter expansion prior to the modelling exercise. It not may be immediately clear what order of polynomial should be used in advance, and specifying the wrong expansion may hide important local variation in model form.

Alternative approaches that account for spatial heterogeneity also exist, examples being spatially adaptive filtering (Foster and Gorr, 1986) in which parameter estimates are allowed to ‘drift’ across the study area, and multi-level modelling (Goldstein, 1987) in which models for individual and spatially aggregate characteristics are combined within the same overall model. A fourth approach extends the random coefficients model (Rao, 1965) to the spatial case (Swamy, 1971) where coefficientsvary randomly across the study area. Again in both spatially adaptive filtering and random coefficients models, the parameter estimates may be mapped to examine local variations in model form. Further insights into these models can be found in Fotheringham (1997)

Geographically Weighted Regression (GWR)

Geographically Weighted Regression (GWR) is a fairly recent contribution to modelling spatially heterogeneous processes (Brunsdon et al, 1996; Fotheringham et al 1996; 1997; 2002). The underlying idea of GWR is that parameters may be estimated anywhere in the study area given a dependent variable and a set of one or more independent variables which have been measured at places whose location is known. Taking Tobler's observation about nearness and similarity into account we might expect that if we wish to estimates parameters for a model at some location **u**[2] then observations which are nearer that location should have a greater weight in the estimation than observations which are further away.

We shall assume that the analyst has a dataset consisting of a dependent variable y and a set of *m* independent variable(s) X_k, *k=1…m*, and that for each of the *n* observations in the dataset a measurement of its position is available in a suitable coordinate system.

The equation for a typical GWR version of the OLS regression model would be:

$$y_i(\mathbf{u}) = \beta_{0i}(\mathbf{u}) + \beta_{1i}(\mathbf{u})x_{1i} + \beta_{2i}(\mathbf{u})x_{2i} + \ldots + \beta_{mi}(\mathbf{u})x_{mi}$$

The notation $\beta_{0i}(\mathbf{u})$ indicates that the parameter describes a relationship around location **u** and is specific to that location. A prediction may be made for the dependent variable if measurements for the independent variables are also available at the location **u**. Typically the locations at which parameter estimates are obtained are those at which data are collected, but this need not necessarily be the case. This would seem to be an unusual claim, but it will be come clear when we consider the nature of the geographical weighting.

The estimator for this model is similar to the WLS (weighted least squares) *global model* above except that the weights are conditioned on the location **u** relative to the other observations in the dataset and hence change for each location. The estimator takes the form:

$$\hat{\beta}(\mathbf{u}) = (X^T W(\mathbf{u}) X)^{-1} X^T W(\mathbf{u}) y$$

[2] We shall use **u** to indicate some general location in the study area. Typically **u** will be a vector of coordinates measured in either a projected coordinate system (such as Universal Transverse Mercator) or a geodetic system such as WGS84. A particular location can be indexed $\mathbf{u}_i$, with Cartesian coordinates (u_ι, v_i) or geodetic coordinates (λ_i, ϕ_i).

W(**u**) is a square matrix of weights relative to the position of **u** in the study area; $X^T W(\mathbf{u}) X$ is the geographically weighted variance-covariance matrix (the estimation requires its inverse to be obtained), and *y* is the vector of the values of the dependent variable.

The W(**u**) matrix contains the geographical weights in its leading diagonal and 0 in its off-diagonal elements.

$$\begin{bmatrix} w_1(\mathbf{u}) & 0 & 0 & 0 \\ 0 & w_2(\mathbf{u}) & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & w_n(\mathbf{u}) \end{bmatrix}$$

The weights themselves are computed from a weighting scheme that is also known as a *kernel*. A number of kernels are possible: a typical one has a Gaussian shape:

$$w_i(\mathbf{u}) = e^{-0.5\left(d_i(\mathbf{u})/h\right)^2}$$

where $w_i(\mathbf{u})$ is the geographical weight of the *i*th observation in the dataset relative to the location **u**, $d_i(\mathbf{u})$ is some measure of the distance between the *i*th observation and the location **u**, and *h* is a quantity known as the bandwidth. The distances are generally Euclidean distances when Cartesian coordinates are used and Great Circle distances when spherical coordinates are used. However, there is no reason why non-Euclidean distances might be used (for example, distances along a road network).

The bandwidth in the kernel is expressed in the same units as the coordinates used in the dataset. As the bandwidth gets larger the weights approach unity and the *local* GWR model approaches the *global* OLS model.

As we have already stated, the locations at which parameters are estimated are generally the locations at which the observations in the dataset have been collected. This allows predictions to be made for the dependent variable and residuals to be computed. These are necessary in determining the goodness of fit of the model and we shall discuss this below. The locations at which parameters are estimated can also be *non-sample* points in the study area – perhaps the mesh points of a regular grid, or the locations of observations in a *validation* dataset which has the same dependent and independent variables as the *calibration* dataset.

It is convenient to refer to the locations at which the calibration data have been collected as the *sample points* and the locations at which parameters are estimated as the *regression* *points*. The combination of geographically weighted estimator, kernel and bandwidth can be referred to as a *local model*.

Kernels other than the Gaussian can be used in GWR although in practice it generally matters very little as long as the kernel is 'Gaussian-like'. In terms of influencing the fit of the model, the choice of a bandwidth is more important than the shape of the kernel. If the sample points are reasonably regularly spaced in the study area, then a kernel with a *fixed* bandwidth is a suitable choice for modelling. If the sample points are not regularly spaced but are clustered in the study area, it is generally desirable to allow the kernel to accommodate this irregularity by increasing its size when the sample points are sparser and decreasing its size when the sample points are denser. A convenient way of implementing this *adaptive* bandwidth specification is to choose a kernel which allows the same number of sample points for each estimation. This is usually accomplished by sorting the distances of the sample points from the desired regression point **u** and setting the bandwidth so that it includes only the first *p* observations, where the optimal value of *p* is found from the data. The weight can be computed by using the specified kernel and setting the value for any observation whose distance is greater than the bandwidth to zero, thereby excluding them from the local calibration. One such kernel is the bisquare:

$$w_i(\mathbf{u}) = (1 - ({}^{d_i(\mathbf{u})}\!/_{h})^2)^2$$

where $w_i(\mathbf{u})$ is zero when $d_i(\mathbf{u}) > h$. This is a near-Gaussian function with the useful property that the weight is zero at a finite distance. In the ArcGIS implementation a fixed radius kernel is Gaussian and the adaptive kernel is based on the bisquare.

When the sample and regression points coincide residuals and predictions of the dependent variables are available. These values can be used to measure the goodness of fit of the model. For the conventional global model, the usual goodness of fit measure is the r^2 or adjusted r^2 value. The adjusted value is preferable if several models are compared as it compensates for the number of variables or parameters in the model. In general a model with more variables or parameters is likely to have a higher r^2 than one with fewer. The situation is a little more complex with GWR and we need to consider the *effective number of parameters* in the model when computing a goodness-of-fit measure.

An interesting matrix in regression modelling is known as the *hat* matrix, **S**. When the observed *y* values are premultiplied by **S** we obtain the predicted (fitted) values thus:

$$\hat{y} = \mathbf{S}y$$

The trace of the matrix **S** in a global model yields the number of parameters in that model – the trace is the sum of the values in the leading diagonal of this matrix, usually expressed as tr(**S**). The trace of S for an OLS regression is the number of parameters in the model. In a GWR model the *effective number of parameters* is obtained from the expression 2tr(**S**)-tr($\mathbf{S}^T\mathbf{S}$). The effective number of parameters in the model depends on the number of independent variables and the bandwidth and can often be large and is usually not an integer. However, it is useful in evaluating the fit of the model.

The measure of goodness of fit which we use extensively in GWR is the *corrected* Akaike Information Criterion (Hurvich et al, 1998). This takes the following form:

$$AIC_c = 2n\log_e(\hat{\sigma}) + n\log_e(2\pi) + n\left(\frac{n + tr(\mathbf{S})}{n - 2 - tr(\mathbf{S})}\right)$$

where n is the number of observations in the dataset, $\hat{\sigma}$ is the estimate of the standard deviation of the residuals, and tr(**S**) is the trace of the hat matrix. The AIC_c can be used to compare models of the same y variable which have very different right hand sides and it contains a penalty for the complexity (degrees of freedom) of the model.

The AIC_c provides a measure of the *information distance* between the model which has actually been fitted and the unknown 'true' model. This distance is not an absolute measure but a relative measure known as the Kullback-Leibler information distance. Two separate models which are being compared are held to be equivalent if the difference between the two AIC_c values is less than 3. This is a widely accepted rule of thumb, however the more cautious analyst might use 4 instead. As the AIC_c is a relative measure, the actual values which are reported in the GWR output might be counter-intuitively large or small. This does not matter, as it is the differences in the AIC_c values which are important. The AIC_c formula contains log terms, and with a little manipulation it can be shown that the difference between AIC_c values for two models with identical degrees of freedom corresponds to the ratio of the likelihoods of the models, although it should be stressed that the AIC_c is not a likelihood ratio test.

The AIC_c can not only be used to compare models with different independent variable subsets, but can also be used the compare the global OLS model with a local GWR model. The AIC_c is also used in the software the determine the 'optimal' value for the bandwidth; the bandwidth with the lowest AIC_c is used in the estimation of the model parameters. However it is up to the analyst to choose the final best value, and there may be good a priori reasons for choosing one that is not suggested by the plot of AIC_c against bandwidth.

Outputs from GWR

As a minimum, GWR will produce parameter estimates and their associated standard errors at the regression points. If the regression points are the same as the sample points then GWR will produce predictions for the dependent variable (fitted values), residuals and standardised residuals. Some implementations will also output local r^2 values and influence statistics based on the hat matrix.

If the regression points are not the same as the sample points, and there are no independent variables available for the regression points, then little else besides parameter estimates and standard errors will be available – fitted values, residuals, and a hat matrix will not be available. If independent variables are available, then fitted values will be available. If there is also a dependent variable present as well, then the whole range of outputs can be created.

Interpreting parameter estimates

In a global model the analyst may be interested in whether the parameter estimates provide any insights into the process being modelled. This is not always the case, since the goal of the analysis may be only to get better predictions of the dependent variable, but it is always advisable, even if this is the goal, to check the parameter estimates.

Each parameter has a sign and a magnitude. If the sign is positive, then an increase in the value of the variable to which the parameter refers will induce an increase in the dependent variable. If the sign is negative, then a decrease will be induced. The size of the change depends on the magnitude of the parameter estimate – a change of 1 unit will lead to a change in the dependent variable of an amount equivalent to the magnitude of the parameter estimate. For example, a model of the form $y = 0.5 - 0.7x$ tells us that when x is zero, we can expect y to be 0.5, and for each unit increase in x, y will decrease correspondingly by 0.7.

The situation is analogous in GWR, except that we have a surface of parameter estimates – we do not estimate everywhere on the surface, only at the regression points, so our output is a sample from a much larger, effectively infinite, population. The spatial changes in the magnitude of the parameter estimates across the surface indicate the locally changing influence of a variable on the dependent variable – in some areas the influence might be much stronger than in other areas. This is the essence of spatial heterogeneity – the structure of the model changes from place to place across the study area as the parameter estimates change in relation to each other in the model. The local parameter estimates are mappable and should be mapped.

As well as mapping the parameter estimates, the analyst should also map the associated standard errors. Are the local standard errors sufficiently large for us to doubt whether the values of the parameter which has been estimated are non-zero?

In a global model it is usual to test whether the parameter estimates are significantly different from zero. This can be accomplished with a t-test – the t statistics and their associated p-values are usually provided on the computer output. A parameter whose estimated value is found to be not significantly different from zero is associated with a variable whose variation does not contribute to the mode. Variables with non-significant parameter estimates can be dropped from the model.

The situation with GWR is a little more complex and is the subject of current research. As there is one set of parameters associated with each regression point, as well as one set of standard errors, then there are potentially hundreds or thousands of tests that would be required to determine whether parameters are locally significant. The assumptions behind the tests mean that if the 0.05 significance level is used we would expect 5 tests in every hundred to be significant. With a 5 variable model estimated at 20000 regression points, we would expect 5000 of these tests to return a significant result. This is the spectre that is raised by *multiple testing*.

The situation also arises in analysis of variance. Having determined that at least one of the means is different from the rest, it is common to use some form of post hoc test to determine which are different. Such tests control the significance level to take into account the multiple testing. There are many such methods, associated with statisticians such as Bonferroni, Tukey, Sidak, and Scheffé.

Using a Bonferroni correction (which downweights the significance level by the number of tests being made) is inappropriate when the tests being carried out are highly correlated as it is highly conservative and therefore likely to miss many real differences. A potential solution for GWR exists in the Benjamini-Hochberg (1995) False Discovery Rate (FDR) procedure – this modifies the significance level for each separate test in a consistent fashion. The test is not implemented in any version of GWR at the time of writing, but as the DBF files which are part of the ESRI shapefile structure are easily read in Excel, the approach detailed by Thissen, Steinberg and Kuang (2002) is convenient, and the corresponding results can be mapped. Benjamini and Yekutieli (2001) report a development of the FDR approach with dependent tests.

Parameter estimates for a variable that are close to zero often tend to be spatially clustered indicating that in these parts of the study area, changes in this variable do not influence changes in the dependent variable. This is potentially interesting and encourages further curiosity about the processes, the data, the model, and the outcome.

Extensions to GWR

This White Paper has discussed the OLS version of the GWR model which is implemented in ArcGIS 9.3. However, this model is not appropriate for every kind of data. If the analyst is using count data as the dependent variable then, for example, a Poisson model might be appropriate as it will not predict negative quantities (the potential range of the predictions with OLS is $-\infty$ to $+\infty$). For data which is not quite Poisson distributed, a Negative Binomial model might be more useful. With case-control or similar data where the dependent variable is 0/1 valued, then a logistic (binary logit) form is appropriate. The fitted values from the model are the probabilities that the dependent variable takes the value 1.

Other extensions to GWR include the possibility of creating models where some variables are held constant across the study area, and others are allowed to vary spatially. Such models are known as mixed or semi-parametric models. Nakaya et al (2005) derive a semi-parametric Poisson model and use it in an investigation of the determinants of premature mortality in Tokyo. The development of such models raises questions of variable selection and which parameters should be allowed to vary spatially or remain constant. Work is in progress on this issue.

References

Anselin L, 1988, *Spatial econometrics: methods and models*, Dordrecht: Kluwer Academic Publishers

Benjamini Y and Hochberg Y, 1996, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B*, 57, 289-300

Benjamini Y and Yekutieli D, 2001, The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics*, 29(4), 1165-1188

Brunsdon C, Fotheringham AS and Charlton M, 1996, Geographically weighted regression: a method for exploring spatial non-stationarity, *Geographical Analysis*, 28(4), 281-298

Casetti E, 1972, Generating models by the expansion method: applications to geographic research, *Geographical Analysis*, 4, 81-91

Foster SA and Gorr WL, 1986, An adaptive filter for estimating spatially varying parameters: application to modelling police hours spent in response to calls for service, *Management Science*, 32, 878-89

Fotheringham AS, 1997, Trends in quantitative methods I: stressing the local, *Progress in Human Geography*, 21, 88-96

Fotheringham AS, Brunsdon C and Charlton M, 1996, The geography of parameter space: an investigation of spatial non-stationarity, *International Journal of Geographical Information Systems*, 10, 605-627

Fotheringham AS, Brunsdon C and Charlton M, 2002, *Geographically Weighted Regression: the analysis of spatially varying relationships*, Chichester: Wiley

Fotheringham AS, Charlton M and Brunsdon C, 1997, Two techniques for exploring non-stationarity in geographical data, *Geographical Systems*, 4, 59-82

Goldstein H, 1987, *Multilevel models in educational and social research*, London: Oxford University Press

Hurvich CM, Simonoff JS and Tsai C-L, 1998, Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of Royal Statistical Society, Series B*, 60, 271-293

Nakaya T, Fotheringham AS, Brunsdon C and Charlton M, 2005, Geographically weighted Poisson regression for disease association mapping*,* *Statistics in Medicine*, 24(17), 2695-2717

Rao CR, 1965, The theory of least squares when the parameters are stochastic and it application to the analysis of growth curves, *Biometrika*, 52, 447-458

Swamy PAV, 1971, *Statistical inference in random coefficient regression models*, Berlin: Springer

Thissen D, Steinberg L and Kuang D, 2002, Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons, *Journal of Educational and Behavioural Statistics*, 27(1), 77-83

Tobler, WR, 1970, A computer movie simulating urban growth in the Detroit region, *Economic Geography*, 46(2), 234-24