Geographically Weighted Regression

A Tutorial on using GWR in ArcGIS 9.3

Martin Charlton
A Stewart Fotheringham

Introduction

Geographically Weighted Regression (GWR) is a powerful tool for exploring spatial heterogeneity. Spatial heterogeneity exists when the structure of the process being modelled varies across the study area. We term a simple linear model such as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

a *global* model – the relationship between y and x is assumed to be constant across the study area – at every possible location in the study area the values of β_0 and β_1 are the same. The residuals from this model ε_i are assumed to be independent and normally distributed with a mean of zero (sometimes this is termed *iid* – independent and identically distributed).

This short tutorial is designed to introduce you to the operation of the Geographically Weighed Regression Tool in ArcGIS 9.3. It assumes that you understand both regression and Geographically Weighted Regression (GWR) techniques. A separate ESRI White Paper is available which outlines the theory underlying GWR.

Modelling the Determinants of Educational Attainment in Georgia

We use a simple example: modelling the determinants of educational attainment in the counties of of the State of Georgia. The dependent variable in this example is the proportion of residents with a Bachelor's degree or higher in each county (**`PctBach`**). The four independent variables that we shall use are:

Proportion of elderly residents in each county:	**`PctEld`**
Proportion of residents who are foreign born:	**`PctFB`**
Proportion of residents who are living below the poverty line:	**`PctPov`**
Proportion of residents who are ethnic black:	**`PctBlack`**

The spatial variation in each of the variables should be mapped by way of initial data exploration. There are some clear patterns in the educational attainment variable – high values around Atlanta and Athens. This is perhaps not surprising since the campuses of Georgia Institute of Technology, Georgia State University, Kennesaw State University, and Georgia Perimeter College are around Atlanta, and the University of Georgia (which has the largest enrolment of all the universities in Georgia) is located in Athens.

Mapping the individual independent variables suggests that there might be some relationships with the variation in educational attainment, and some initial analysis also suggests that these variables are reasonable as predictors. The proportion of elderly is included because concentrations of educational attainment are usually associated with concentrations of the young rather than the old – we would expect there to be increased proportions of the elderly to have a negative influence on educational attainment. It is suspected that there might be a higher value given to further education amongst recent migrants who are anxious for their children to succeed. Educational attainment is generally associated with affluence, so we would expect those parts of the State with higher proportions of those living below the poverty line to have lower proportions of those educated to degree level. Higher proportions of ethnic black residents in the population are sometimes associated with poorer access to grad schools and lower interest in higher education.

Before any analysis with regression takes place, we will have undertaken some initial statistical analysis to determine the characteristics of each of the variables which are proposed for the model. Some summary statistics for the variables in the exercise are presented in the Table 1.

Table 1: Summary Statistics

Variable	Mean	Std Deviation	Median	Minimum	Maximum
PctBach	10.95	5.70	9.40	4.20	37.50
PctEld	11.74	3.08	12.07	1.46	22.96
PctFB	1.13	1.23	0.72	0.04	6.74
PctPov	19.34	7.25	18.60	2.60	35.90
PctBlack	27.39	17.38	27.64	0.00	79.64

The correlation analysis shown in Table 2 reveals some initial associations.

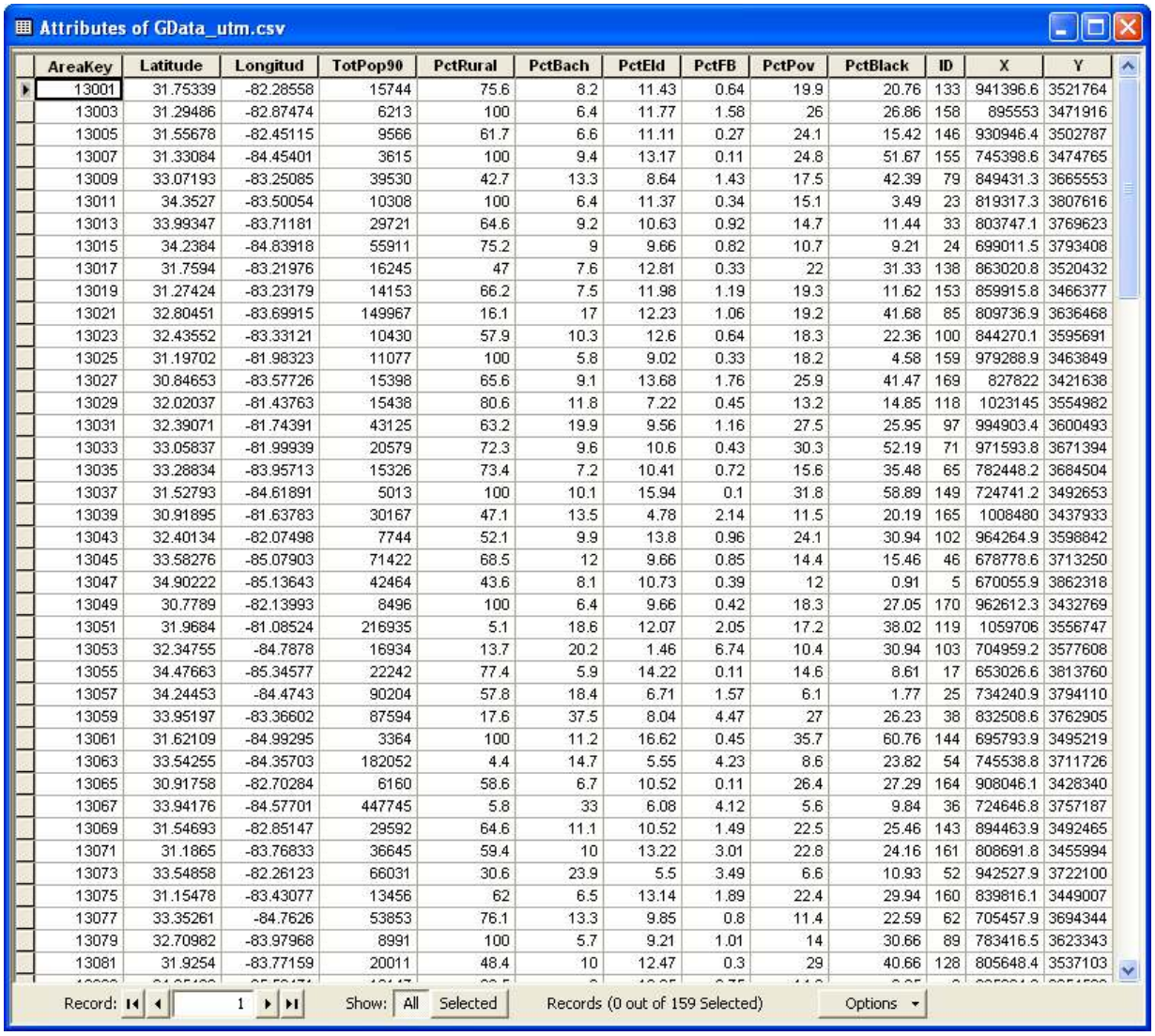
Table 2: Correlation Coefficients

	PctBach	PctEld	PctFB	PctPov
PctEld	-0.46			
PctFB	0.67	-0.48		
PctPov	-0.40	0.57	-0.33	
PctBlack	-0.11	0.30	-0.11	0.74

Most of the associations with PctBach are in the expected direction. One interesting correlation is that between PctBlack and PctPov - there is some colinearity here (r=0.74), but probably not enough for us to worry about at this stage.

The attribute table for the Georgia shapefile is shown in Figure 1. You will notice that there some other variables in the file which we will not use. The AreaKey item contains the FIPS codes for the counties in Georgia. The X and Y columns contain the coordinates in a UTM projection suitable for Georgia.

Figure 1: Georgia counties feature class attribute table



Attributes of GData_utm.csv

AreaKey	Latitude	Longitud	TotPop90	PctRural	PctBach	PctEld	PctFB	PctPov	PctBlack	ID	X	Y
13001	31.75339	-82.28558	15744	75.6	8.2	11.43	0.64	19.9	20.76	133	941396.6	3521764
13003	31.29486	-82.87474	6213	100	6.4	11.77	1.58	26	26.86	158	895553	3471916
13005	31.55678	-82.45115	9566	61.7	6.6	11.11	0.27	24.1	15.42	146	930946.4	3502787
13007	31.33084	-84.45401	3615	100	9.4	13.17	0.11	24.8	51.67	155	745398.6	3474765
13009	33.07193	-83.25085	39530	42.7	13.3	8.64	1.43	17.5	42.39	79	849431.3	3665553
13011	34.3527	-83.50054	10308	100	6.4	11.37	0.34	15.1	3.49	23	819317.3	3807616
13013	33.99347	-83.71181	29721	64.6	9.2	10.63	0.92	14.7	11.44	33	803747.1	3769623
13015	34.2384	-84.83918	55911	75.2	9	9.66	0.82	10.7	9.21	24	699011.5	3793408
13017	31.7594	-83.21976	16245	47	7.6	12.81	0.33	22	31.33	138	863020.8	3520432
13019	31.27424	-83.23179	14153	66.2	7.5	11.98	1.19	19.3	11.62	153	859915.8	3466377
13021	32.80451	-83.69915	149967	16.1	17	12.23	1.06	19.2	41.68	85	809736.9	3636468
13023	32.43552	-83.33121	10430	57.9	10.3	12.6	0.64	18.3	22.36	100	844270.1	3595691
13025	31.19702	-81.98323	11077	100	5.8	9.02	0.33	18.2	4.58	159	979288.9	3463849
13027	30.84653	-83.57726	15398	65.6	9.1	13.68	1.76	25.9	41.47	169	827822	3421638
13029	32.02037	-81.43763	15438	80.6	11.8	7.22	0.45	13.2	14.85	118	1023145	3554982
13031	32.39071	-81.74391	43125	63.2	19.9	9.56	1.16	27.5	25.95	97	994903.4	3600493
13033	33.05837	-81.99939	20579	72.3	9.6	10.6	0.43	30.3	52.19	71	971593.8	3671394
13035	33.28834	-83.95713	15326	73.4	7.2	10.41	0.72	15.6	35.48	65	782448.2	3684504
13037	31.52793	-84.61891	5013	100	10.1	15.94	0.1	31.8	58.89	149	724741.2	3492653
13039	30.91895	-81.63783	30167	47.1	13.5	4.78	2.14	11.5	20.19	165	1008480	3437933
13043	32.40134	-82.07498	7744	52.1	9.9	13.8	0.96	24.1	30.94	102	964264.9	3598842
13045	33.58276	-85.07903	71422	68.5	12	9.66	0.85	14.4	15.46	46	678778.6	3713250
13047	34.90222	-85.13643	42464	43.6	8.1	10.73	0.39	12	0.91	5	670055.9	3862318
13049	30.7789	-82.13993	8496	100	6.4	9.66	0.42	18.3	27.05	170	962612.3	3432769
13051	31.9684	-81.08524	216935	5.1	18.6	12.07	2.05	17.2	38.02	119	1059706	3556747
13053	32.34755	-84.7878	16934	13.7	20.2	1.46	6.74	10.4	30.94	103	704959.2	3577608
13055	34.47663	-85.34577	22242	77.4	5.9	14.22	0.11	14.6	8.61	17	653026.6	3813760
13057	34.24453	-84.4743	90204	57.8	18.4	6.71	1.57	6.1	1.77	25	734240.9	3794110
13059	33.95197	-83.36602	87594	17.6	37.5	8.04	4.47	27	26.23	38	832508.6	3762905
13061	31.62109	-84.99295	3364	100	11.2	16.62	0.45	35.7	60.76	144	695793.9	3495219
13063	33.54255	-84.35703	182052	4.4	14.7	5.55	4.23	8.6	23.82	54	745538.8	3711726
13065	30.91758	-82.70284	6160	58.6	6.7	10.52	0.11	26.4	27.29	164	908046.1	3428340
13067	33.94176	-84.57701	447745	5.8	33	6.08	4.12	5.6	9.84	36	724646.8	3757187
13069	31.54693	-82.85147	29592	64.6	11.1	10.52	1.49	22.5	25.46	143	894463.9	3492465
13071	31.1865	-83.76833	36645	59.4	10	13.22	3.01	22.8	24.16	161	808691.8	3455994
13073	33.54858	-82.26123	66031	30.6	23.9	5.5	3.49	6.6	10.93	52	942527.9	3722100
13075	31.15478	-83.43077	13456	62	6.5	13.14	1.89	22.4	29.94	160	839816.1	3449007
13077	33.35261	-84.7626	53853	76.1	13.3	9.85	0.8	11.4	22.59	62	705457.9	3694344
13079	32.70982	-83.97968	8991	100	5.7	9.21	1.01	14	30.66	89	783416.5	3623343
13081	31.9254	-83.77159	20011	48.4	10	12.47	0.3	29	40.66	128	805648.4	3537103

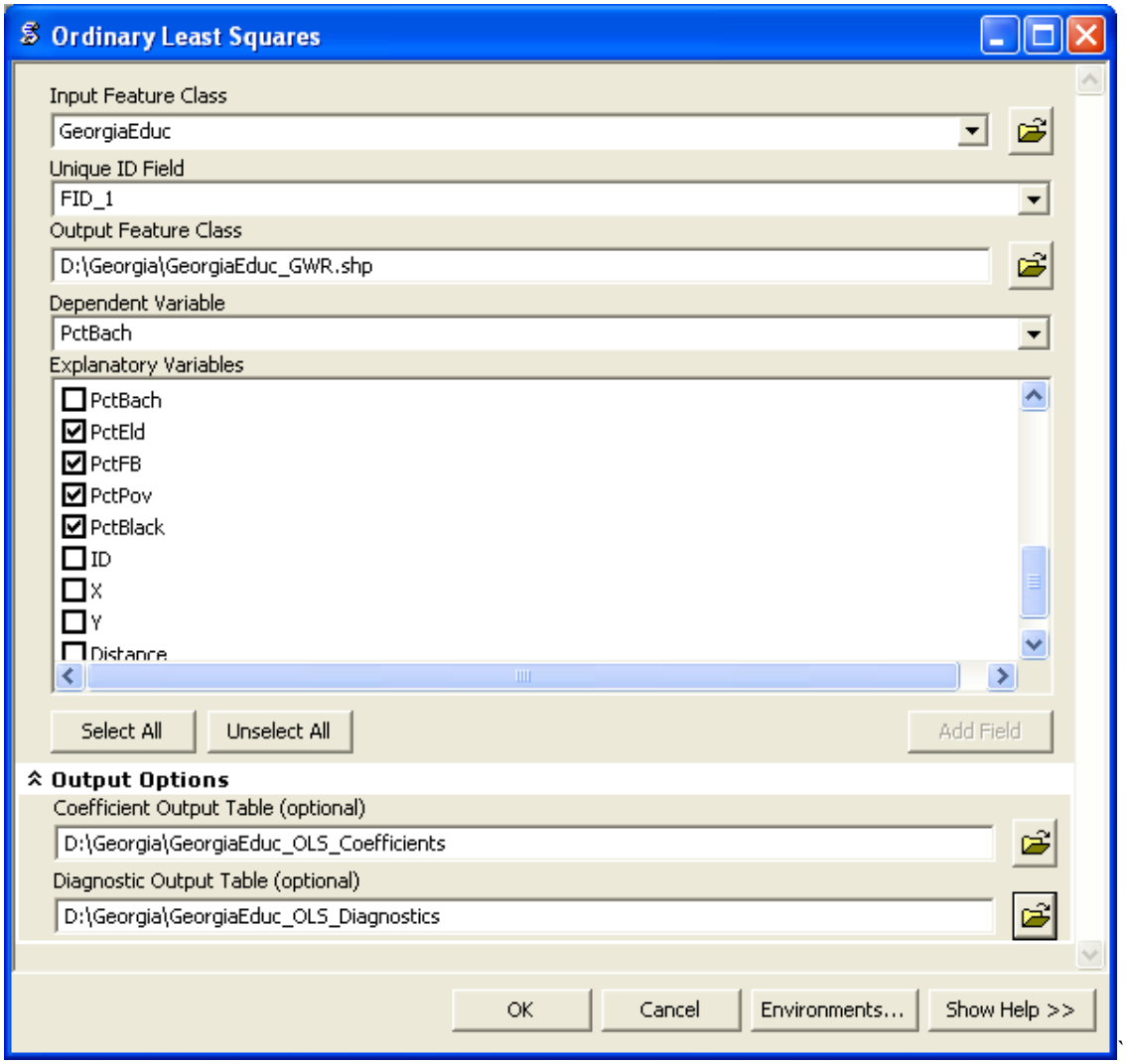
Record: 1 Show: All Selected Records (0 out of 159 Selected) Options

Getting started: OLS Regression

GWR is not a panacea for all regression ills and it should not be the automatic first choice in any regression modelling exercise. We will begin by fitting an 'ordinary' linear regression model – this is 'ordinary' in the sense that it's the default regression model in packages such as SPSS or R and the estimation of the coefficients is by Ordinary Least Squares. The residuals are assumed to be independently and identically normally distributed around a mean of zero. The residuals are also assumed to be homoscedastic – that is, any samples taken at random from the residuals will have the same mean and variance.

There is an OLS regression modelling tool in the Spatial Statistics Tools in Arc Toolbox. You may need to uncheck the Hide Locked Tools option for Arc Toolbox before you can see the tool listed. The form to specify the model structure for this example is shown in Figure 2. You should save both the coefficients and diagnostics to separate DBF tables for later scrutiny.

Figure 2: Ordinary Least Squares Tool



Clicking on the [OK] button will run the Tool. The results of the OLS analysis are shown in Figure 3.

Attributes of GeorgiaEduc_OLS

FID	Shape	FID_1	PCTBACH	PCTELD	PCTFB	PCTPOV	PCTBLACK	Estimated	Residual	StdResid
0	Polygon	130	8.2	11.43	0.64	19.9	20.76	9.17538	-0.975376	-0.25191
1	Polygon	155	6.4	11.77	1.58	26	26.86	10.3048	-3.90484	-1.0085
2	Polygon	146	6.6	11.11	0.27	24.1	15.42	6.7348	-0.134799	-0.034814
3	Polygon	156	9.4	13.17	0.11	24.8	51.67	8.558621	0.841378	0.217302
4	Polygon	74	13.3	8.64	1.43	17.5	42.39	13.7487	-0.448717	-0.11589
5	Polygon	22	6.4	11.37	0.34	15.1	3.49	8.46332	-2.06332	-0.532892
6	Polygon	32	9.2	10.63	0.92	14.7	11.44	10.7147	-1.51472	-0.391205
7	Polygon	24	9	9.66	0.82	10.7	9.21	11.502	-2.50199	-0.646187
8	Polygon	137	7.6	12.81	0.33	22	31.33	8.43034	-0.830338	-0.214451
9	Polygon	154	7.5	11.98	1.19	19.3	11.62	10.0005	-2.50053	-0.64581
10	Polygon	84	17	12.23	1.06	19.2	41.68	11.8752	5.12476	1.32357
11	Polygon	99	10.3	12.6	0.64	18.3	22.36	9.59338	0.706624	0.182499
12	Polygon	158	5.8	9.02	0.33	18.2	4.58	7.9453	-2.1453	-0.554066

Record: 1 Show: All Selected Records (0 out of 174 Selected) Options

Figure 1: Output feature class attribute table

A useful place to start is with the model diagnostics. There are a number of different *goodness-of-fit* measures: the r^2 is 0.53 and the adjusted r^2 is 0.51. The r^2 measures the proportion of the variation in the dependent variable which is accounted for by the variation in the model, and the possible values range from 0 to 1. Values closer to 1 indicate that the model has a better predictive performance. However, its values can be influenced by the number of the variables which are in the model – increasing the number of variables will never decrease the r^2. The adjusted r^2 is a preferable measure since it contains some adjustment for the number of variables in the model. In the model we have just fitted, the value of 0.51 indicates that it accounts for about half the variation in the dependent variable. This suggests that perhaps some variables have been omitted from the model, or the form of the model is not quite right: we are failing to account for 49% of the variation in educational attainment with our model.

A slightly different measure of goodness-of-fit is provided by the Akaike Information Criterion (AIC). Unlike the r^2 the AIC is not an absolute measure – it is a relative measure and can be used to compare different models which have the same independent variable. It is a measure of the 'relative distance' between the model that has been fitted and the unknown 'true' model. Models with smaller values of the AIC are preferable to models with higher values (where 5 is less than 10 and -10 is less than -5); however, if the difference in the AIC between two models is less than about 3 or 4, they are held to be equivalent in their explanatory power. The AIC formula contains log terms and sometimes the values can be unexpectedly large or negative – this is not important – it is the *difference* between the AICs that we are interested in. The AIC in this case is 969.82.

We have fitted an OLS model to *spatial* data. It is likely that there will be some structure in the residuals. We have not taken this into account in the model, which may be one contributory factor towards its rather indifferent performance. The value of the Jarque-Bera statistic indicates that the residuals appear not to be normally distributed. The OLS tool prints a warning that we should test to determine whether the residuals appear to be spatially autocorrelated.

We now examine the model coefficient estimates which are shown in Table 3 along with the t-statistics for each estimated coefficient. The signs on the coefficient estimates are as expected, with the exception of PctBlack (we have already noted a raised correlation between it and PctPov).

Table 3: OLS Model Parameter Estimates

Variable	Coefficient	t-Statistic
Intercept	12.789636	8.410249
PctEld	-0.116422	-0.896639
PctFB	2.538762	8.928844
PctPov	-0.272978	-3.723328
PctBlack	0.073405	2.800241

The t-statistics test the hypothesis that the value of an individual coefficient estimate is not significantly different from zero. With the exception of PctEld, the coefficient estimates are all statistically significant (this is, their values are sufficiently large for us to assume that they are not zero in the population from which our sample data have been drawn). The Variance Inflation Factors are all reasonably small, so there is no strong evidence of variable redundancy.

In completing the OLS model form we specified DBF output tables for the coefficient estimates and the regression diagnostics. These may be examined – the coefficient estimates from the OLS model are shown in Figure 4 and the diagnostics table is shown in Figure 5.

Figure 4: Report from Ordinary Least Squares Tool

```
Executing: OrdinaryLeastSquares GeorgiaEduc FID_1 D:\Georgia\GeorgiaEduc_OLS.shp PctBach
PctEld;PctFB;PctPov;PctBlack D:\Georgia\GeorgiaEduc_OLS_Coefficients.dbf
D:\Georgia\GeorgiaEduc_OLS_Diagnostics.dbf
Start Time: Mon Jan 19 14:59:13 2009
Running script OrdinaryLeastSquares...
                              Summary of OLS Results
Variable  Coefficient StdError t-Statistic Probability Robust_SE Robust_t  Robust_Pr VIF [1]
Intercept 12.789636   1.520720 8.410249    0.000000*   2.012447  6.355267  0.000000* --------
PCTELD    -0.116422   0.129842 -0.896639   0.371177    0.143566  -0.810931 0.418536  1.828902
PCTFB     2.538762    0.284333 8.928844    0.000000*   0.585029  4.339550  0.000028* 1.353249
PCTPOV    -0.272978   0.073316 -3.723328   0.000276*   0.122553  -2.227421 0.027229* 3.332809
PCTBLACK  0.073405    0.026214 2.800241    0.005701*   0.033751  2.174877  0.031020* 2.418869

                                   OLS Diagnostics
Number of Observations:      174          Number of Variables:                          5
Degrees of Freedom:          169          Akaike's Information Criterion (AIC) [2]:     969.823038
Multiple R-Squared [2]:      0.525104     Adjusted R-Squared [2]:                       0.513864
Joint F-Statistic [3]:       46.716921    Prob(>F), (4,169) degrees of freedom:         0.000000*
Joint Wald Statistic [4]:    89.061691    Prob(>chi-squared), (4) degrees of freedom:   0.000000*
Koenker (BP) Statistic [5]:  43.772814    Prob(>chi-squared), (4) degrees of freedom:   0.000000*
Jarque-Bera Statistic [6]:   275.825399   Prob(>chi-squared), (2) degrees of freedom:   0.000000*


                          Notes on Interpretation
 *  Statistically significant at the 0.05 level.
[1] Large VIF (> 7.5, for example) indicates explanatory variable redundancy.
[2] Measure of model fit/performance.
[3] Significant p-value indicates overall model significance.
[4] Significant p-value indicates robust overall model significance.
[5] Significant p-value indicates biased standard errors; use robust estimates.
[6] Significant p-value indicates residuals deviate from a normal distribution.

WARNING 000851: Use the Spatial Autocorrelation (Moran's I) Tool to ensure residuals are not spatially
autocorrelated.
Writing Coefficient Output Table....
D:\Georgia\GeorgiaEduc_OLS_Coefficientx.dbf
Writing Diagnostic Output Table....
D:\Georgia\GeorgiaEduc_OLS_Diagnostic.dbf
Completed script OrdinaryLeastSquares...
Executed (OrdinaryLeastSquares) successfully.
```

In the diagnostics DBF table shown in the Figure 5 those statistics which have been discussed are highlighted.

Attributes of OLS_Coefs

OID	Field1	Variable	Coef	StdError	t_Stat	Prob	Robust_SE	Robust_t	Robust_Pr
0	0	Intercept	12.789636	1.52072	8.410249	0	2.012447	6.355267	0
1	0	PCTELD	-0.116422	0.129842	-0.896639	0.371177	0.143566	-0.810931	0.418536
2	0	PCTFB	2.538762	0.284333	8.928844	0	0.585029	4.33955	0.000028
3	0	PCTPOV	-0.272978	0.073316	-3.723328	0.000276	0.122553	-2.227421	0.027229
4	0	PCTBLACK	0.073405	0.026214	2.800241	0.005701	0.033751	2.174877	0.03102

Record: 1 Show: All Selected Records (0 out of 5 Selected) Options

Figure 5: OLS Model Coefficient Estimate DBF Table

The output feature class attribute table shown in Figure 6 contains three extra columns in addition to the original observed data.

Attributes of OLS_Diags

OID	Field1	Diag_Name	Diag_Value	Definition
0	0	AIC	969.823038	Akaike's Information Criterion: A relative measure of performance used to compare models; the smaller AIC indicates the superior model.
1	0	R2	0.525104	R-Squared, Coefficient of Determination: The proportion of variation in the dependent variable that is explained by the model.
2	0	AdjR2	0.513864	Adjusted R-Squared: R-Squared adjusted for model complexity (number of variables) as it relates to the data.
3	0	F-Stat	46.716921	Joint F-Statistic Value: Used to assess overall model significance.
4	0	F-Prob	0	Joint F-Statistic Probability (p-value): The probability that none of the explanatory variables have an effect on the dependent variable.
5	0	Wald	89.061691	Wald Statistic: Used to assess overall robust model significance.
6	0	Wald-Prob	0	Wald Statistic Probability (p-value): The computed probability, using robust standard errors, that none of the explanatory variables have an effect on the dependent variable.
7	0	K(BP)	43.772814	Koenker's studentized Breusch-Pagan Statistic: Used to test the reliability of standard error values when heteroskedasticity (non-constant variance) is present.
8	0	K(BP)-Prob	0	Koenker (BP) Statistic Probability (p-value): The probability that heteroskedasticity (non-constant variance) has not made standard errors unreliable.
9	0	JB	275.825399	Jarque-Bera Statistic: Used to determine whether the residuals deviate from a normal distribution.
10	0	JB-Prob	0	Jarque-Bera Probability (p-value): The probability that the residuals are normally distributed.
11	0	Sigma2	14.991807	Sigma-Squared: OLS estimate of the variance of the error term.

Record: 0 Show: All Selected Records (5 out of 12 Selected) Options

Figure 6: OLS Model Diagnostics DBF Table

The column headed **PCTBACH** contains the observed dependent variable values and the columns headed **PCTELD**, **PCTFB**, **PCTPOV** and **PCTBLACK** contain the values for the independent variables in the model. The column headed **Estimated** contains the predicted y values given the model coefficients and the data for each observation. The predicted y values are sometimes known as the fitted values. The residual is the difference between the observed values of the dependent variable (in this case in the column headed PCTBACH) and the fitted values – these are found in the column headed **Residual**. Finally, the column headed **StdResid** contains standardised values of the residuals: these have a mean of zero and a standard deviation of 1. Observations of interest are those which have positive standardised residuals greater than 2 (model underprediction) or negative standardised residuals less than -2 (model overprediction).

The report from the OLS advised that we should carry out a test to determine whether there is spatial autocorrelation in the residuals. If the residuals are sufficiently autocorrelated then the results of the OLS regression analysis are unreliable – autocorrelated residuals are not iid, so one of the underlying assumptions of OLS regression has been violated. An appropriate test statistic is Moran's I: this is a measure of the level of spatial autocorrelation in the residuals. This tool is available under Spatial Statistics Tools / Analyzing Patterns / Spatial Autocorrelation and is shown in Figure 7

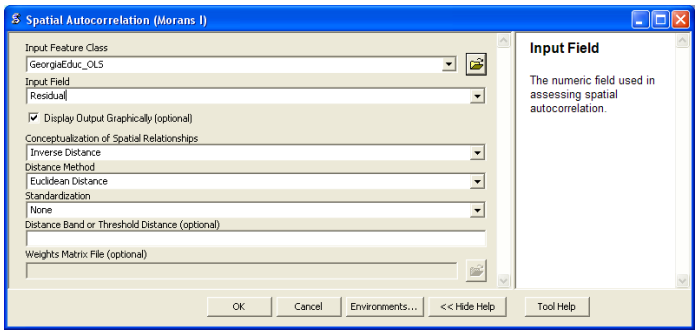


Figure 7: Spatial Autocorrelation Tool

The Input Feature Class should be the Output Feature Class specified in the OLS Regression tool. The Input Field should be Residual (the results are the same if you use StdResid instead). The other choices should be left as their defaults.

The report from the tool is shown in Figure 8

Figure 8: Report from Spatial Autocorrelation Tool

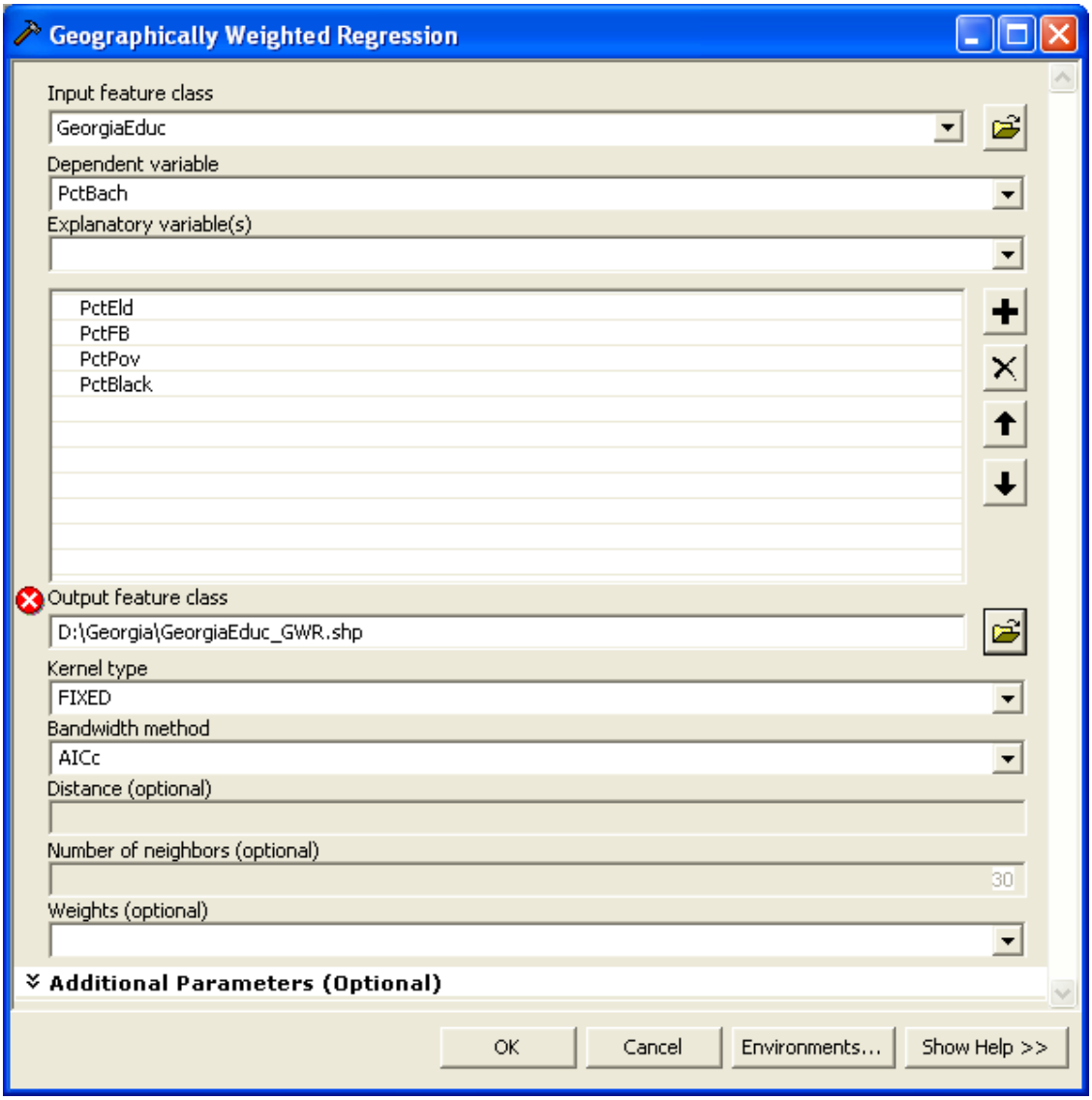
```
Executing: SpatialAutocorrelation GeorgiaEduc_OLS Residual true "Inverse Distance" "Euclidean
Distance" None # # 0 0 0
Start Time: Tue Jan 06 16:16:06 2009
Running script SpatialAutocorrelation...
WARNING 000853: The default neighborhood search threshold was 40696.962105194.

 Global Moran's I Summary
Moran's Index:   0.144841
Expected Index:  -0.005780
Variance:        0.017554
Z Score:         1.136833
p-value:         0.255608

Completed script SpatialAutocorrelation...
Executed (SpatialAutocorrelation) successfully.
```

The value of Moran's I for the OLS model is 0.14, and the p-value for the hypothesis that this value is not significantly different from zero is 0.26 (Z = 1.14). We would normally accept the hypothesis that autocorrelation is not present in the residuals given this value of p, but the graphical output warns that although the pattern is "somewhat clustered" it may also be due to "random chance".However, we have made the assumption here that the model structure is spatially stationary – in other words we assume that the process we are modelling is homogenous. Although we have a model that performs moderately well with reasonably random residuals, we nevertheless would be justified in attempting to improve the reliability of the predictions from the models by using GWR. We also will be able to map the values of the county specific coefficient estimates to examine whether the process appears to be spatially heterogenous.

Geographically Weighted Regression

The ArcGIS 9.3 GWR tool is an exploratory tool. It can be found in Spatial Statistics Tools / Modeling Spatial Relationships / Geographically Weighted Regression. The model choices are specified in a form. The choices we use in this example are shown in Figure 9.

Figure 9: Geographically Weighted Regression Tool



The *Input feature class* will be the same as that which was specified in the OLS model. The *Output feature class* will contain the coefficient estimates and their associated standard errors, as well as a range of observation specific diagnostics. The *Dependent variable* and the *Explanatory variable(s)* will be those which were specified for the OLS model. There are a number of options which may be specified which need some initial thought from the user.

There are two possible choices for the *Kernel type*: FIXED or ADAPTIVE. A spatial kernel is used to provide the geographic weighting in the model. A key coefficient in the kernel is the bandwidth – this controls the size of the kernel. Which kernel is chosen largely depends on the spatial configuration of the feature in the Input feature class. If the observations are either reasonably regularly positioned in the study area (perhaps they are the mesh points of a regular grid) then a FIXED kernel is appropriate; if the observations are clustered so that the density of observations varies around the study area, then an ADAPTIVE kernel is appropriate. If you are not sure which to use, ADAPTIVE will cover most applications.

There are three choices for the *Bandwidth method*: AICc, CV and BANDWIDTH COEFFICIENT. The first two choices allow you to use an automatic method for finding the bandwidth which gives the best predictions, the third allows you to specify a bandwidth. The AICc method finds the bandwidth which minimises the AICc value – the AICc is the corrected Akaike Information Criterion (it has a correction for small sample sizes). The CV finds the bandwidth which minimises a CrossValidation score. In practice there isn't much to choose between the two methods, although the AICc is our preferred method. The AICc is computed from (a) a measure of the divergence between the observed and fitted values and (b) a measure of the complexity of the model. The complexity[1] of a GWR model depends not just on the number of variables in the model, but also on the bandwidth. This interaction between the bandwidth and the complexity of the model is the reason for our preference for the AICc over the CV score.

There may be some modelling contexts where you wish to supply your own bandwidth. In this case, the *Bandwidth method* is BANDWIDTH COEFFICIENT. If you have chosen a FIXED kernel, the coefficient will be a **distance** which is in the **same units** as the coordinate system you are using for the feature class. Thus if your coordinates are in metres, this will be a distance in metres; if they are in miles, the distance will be in miles. If you are using geographic coordinates in decimal degrees, this value will be in degrees – large values (90 for example) will create very large kernels which will cover considerable parts of the earth's surface and the geographical weights will be close to 1 for every observation! If you have chosen an ADAPTIVE kernel the bandwidth is a **count** of the number of **nearest observations** to include under the kernel – the spatial extent of the kernel will change to keep the number of observations in the kernel constant. In general you should have good reasons for specifying an *a priori* bandwidth, and for most applications allowing the GWR tool to chose an 'optimal' bandwidth is good practice.

In the example described here, we have chosen an ADAPTIVE kernel whose bandwidth will be found by minimising the AICc value.

There are a number of optional *Additional coefficients* which are for more advanced users of GWR. One of the features of GWR is that while a model can be fitted to data collected at one set of locations, coefficients may also be estimated as locations at which no data have been collected (for example, the mesh points of a raster) or at other locations for which the ***y***s and ***x***s are known (for example a model can be fitted to a calibration set of data and then used to estimate coefficients and predictions for a validation set).

---

[1] We use the term complexity here as a shorthand for the *number of parameters* in the model. In an OLS regression model, the number of parameters one more than the number of independent variables (the intercept is also a parameter). In a GWR model the equivalent measure is known as the *effective number of parameters* and is usually much larger than that for an OLS model with the same variables and need not be an integer.

Figure 10: Report from the Geographically Weighted Regression Tool

```
Executing: GeographicallyWeightedRegression GeorgiaEduc PctBach PctEld;PctFB;PctPov;PctBlack
D:\Georgia\Georgia_GWR.shp ADAPTIVE AICc # 30 # # 1819.529 # # # D:\Georgia\Georgia_GWR_supp.dbf #
Start Time: Sat Oct 18 11:44:48 2008
Neighbours          : 121
ResidualSquares     : 1815.1926630181806
EffectiveNumber     : 19.691366786638696
Sigma               : 3.4297799048143567
AICc                : 937.9369828885825
R2                  : 0.6597640641724919
R2Adjusted          : 0.6185513689517778
Executed (GeographicallyWeightedRegression) successfully.
```

The GWR tool will create a report and a DBF table which contains the diagnostic statistics which are also listed in the report shown in Figure 10. The report is the first place to start when interpreting the results from a GWR exercise as it provides not only a list of the coefficients which have used by the tool, but also a set of important diagnostic statistics. Recall that the bandwidth of the model has been estimated for an adaptive kernel, unsing AICc minimisation. The *Neighbours* value is the number of nearest neighbours that have been used in the estimation of each set of coefficients. In this case it's 121: this is large in comparison with the number of observations in the dataset (175), and means that under each kernel there are about 70% of the data. There may be some evidence of spatial variation in the coefficient estimates. The *ResidualSquares* value is the sum of the squared residuals – this is used in several subsequent calculations. The *EffectiveNumber* is a measure of the complexity of the model – it is equivalent to the number of parameters in the OLS model and is usually larger than the OLS value and is usually not an integer. It is also used in the calculation of several diagnostics. *Sigma* is the square root of the normalised residual sum of squares. The *AICc* is the corrected Akaike Information Criterion, and with *R2* (r^2) and *R2Adjusted* (the adjusted r^2) provide some indication of the goodness of fit of the model. These diagnostics are also saved in a DBF table whose name is that of the output feature class with the suffix _supp.

We start by comparing the fit of the OLS and GWR models. We'll refer to the OLS model as the *global* model and the GWR model as the *local* model. The global adjusted r^2 is 0.51 and the local adjusted r^2 is 0.62 which suggests that there has been some improvement in model performance. Our preferred measure of model fit is the AICc, the global model's value is 969.82, and the local model's value is 937.94 – the difference of 31.88 is strong evidence of an improvement in the fit of the model to the data[2].

[2] As a general rule of thumb, if the AICc difference between the two models is less than about 4 there is little to choose between them; if the difference between them is greater than about 10 there is little evidence in support of the model with the larger AICc. For further discussion of issues in using the AICc see Burnham and Anderson (2002).

Visualising the GWR output

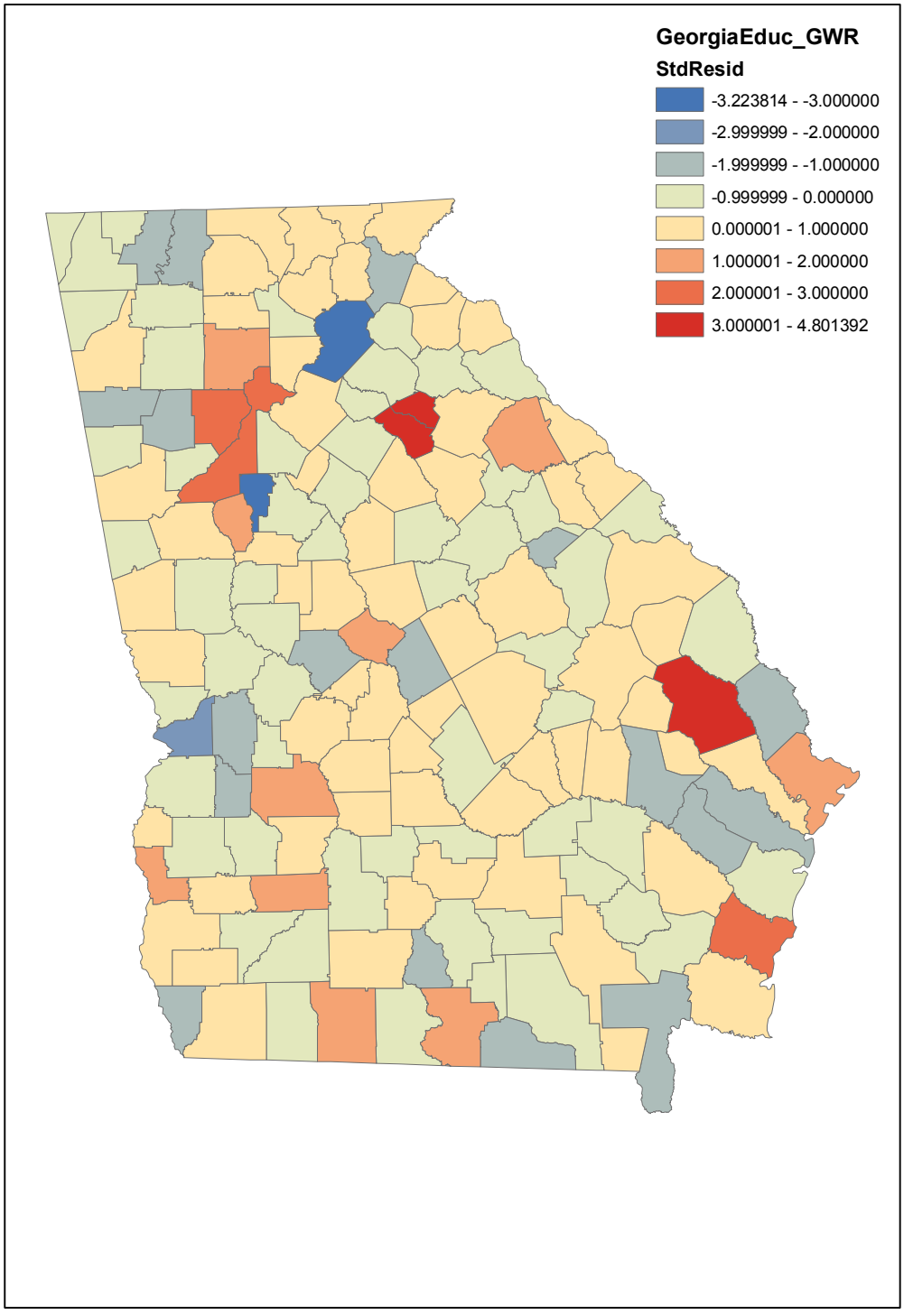
The attribute table for the output feature class contains the coefficient estimates, their standard errors, and a range of diagnostic statistics. Descriptions of the main column headings in this table are given in Table 4.

Table 4: Items in the output feature class attribute table

Observed	The observed value of the dependent (y) variable
Cond	The condition number of the data matrix – local collinearity produces unreliable coefficient estimates – the results should be treated with caution. Values around 5 to 10 suggest weak dependencies in the data, whereas values greater than 30 suggest moderate or stronger dependencies in the data. See Belsley *et al* (2004) for further discussion.
LocalR2	The locally weighed r^2 between the observed and fitted values. The statistic is a measure of how well the model replicates the local *y* values around each observation. See Fotheringham *et al* (2002, 215-216) for further discussion
Predicted	The local prediction of the y variable (fitted value)
Intercept	The local intercept
C*n_abc*	The coefficient for the *n*th independent variable in the model whose item name is *abc* (C1_PctEld, for example)
Residual	The residual – the difference between the observed and fitted value
StdError	The standard error of the residual
StdErr_Int	The locally weighed standard error of the Intercept
StdErrC*n*_P	The locally weighed value of the coefficient for the *n*th variable in the model
StdResid	The standardised residual – these have a mean of zero and a standard deviation of unity.
Source_ID	The FID of the corresponding feature in the Input feature class attribute table.

Mapping the values of *StdResid* (the standardised residual) is a good starting point – these are shown in Figure 11. There are two questions of interest (a) where are the unusually high or low residuals and (b) are the residuals spatially autocorrelated? Not surprisingly those counties with the large universities have very large positive residuals (StdResid > 3) (University of Georgia, Georgia Southern University), and there are large positive residuals for those counties in and around Atlanta which contain major university campuses. We would expect that, given the variables we have in the model, the model would underpredict the levels of educational attainment in these counties. Two counties have noticeable over-prediction and would certainly warrant closer inspection to discover possible reasons for this.

Figure 11: GWR Model: Standardised Residuals



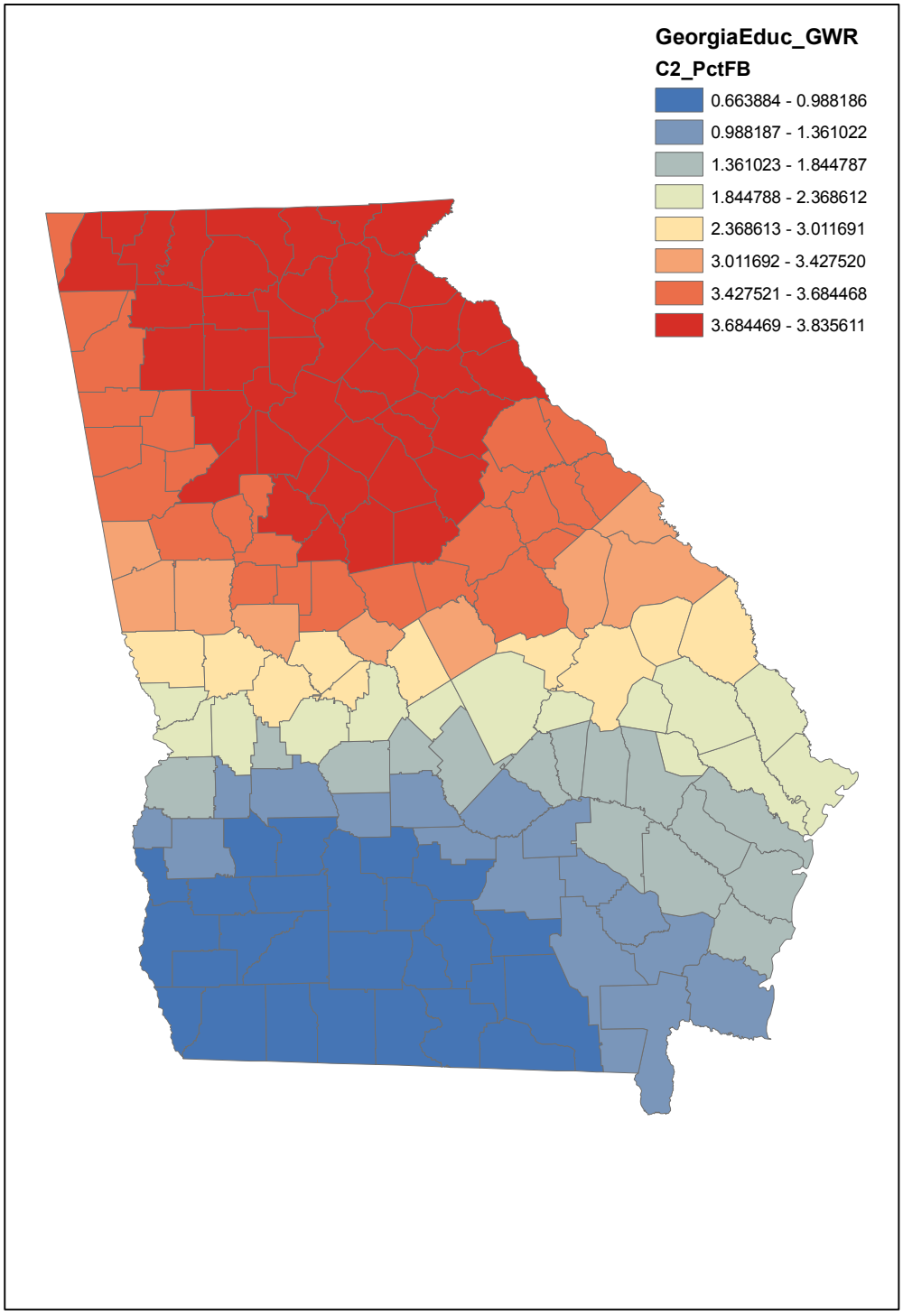
The report from the Spatial Autocorrelation Tool used on the GWR residuals is shown in Figure 12. Moran's I for the residuals is 0.04 (p=0.74) so there is little evidence of any autocorrelation in them. Any spatial dependencies which might have been present in the residuals for the global model have been removed with the geographical weighting in the local model.

Figure 12: Report from the Spatial Autocorrelation Tool on the GWR Residuals

```
Executing: SpatialAutocorrelation D:\Georgia\GeorgiaEduc_GWR.shp Residual false "Inverse Distance"
"Euclidean Distance" None # # 0 0 0
Start Time: Mon Jan 19 15:08:33 2009
Running script SpatialAutocorrelation...
WARNING 000853: The default neighborhood search threshold was 40696.962105194.

 Global Moran's I Summary
Moran's Index:   0.037049
Expected Index:  -0.005780
Variance:        0.017687
Z Score:         0.322037
p-value:         0.747424

Completed script SpatialAutocorrelation...
Executed (SpatialAutocorrelation) successfully.
```

The local coefficient estimates should also be mapped. Figure 13 shows the variation in the coefficient estimates for the PctFB variable. The estimated value for the global model was 2.54, with a standard error of 0.28. (95% CI: 2.00 - 3.09). The map for the local coefficients reveals that the influence of this variable in the model varies considerably over Georgia, with a strong north-south direction. The range of the local coefficient is from 0.67 in the southernmost counties to 3.84 in the northernmost counties – evidence which points to heterogeneity in the model structure within Georgia.

Figure 13: GWR Model: PctFB Parameter Variation



The global coefficient and all the local coefficients for this variable are positive – there is agreement between the two models on the direction of the influence of this variable. There may be some cases where most of the local coefficients have one sign, but for a few observations the sign changes. How can a variable have a positive influence in the model in some areas but a negative influence in other areas?

As the values of the coefficients change sign, they will pass through zero. The coefficients themselves are estimates and have a standard error, so for some of them they will be so close to zero that any variation in the variables concerned will not influence the local variation in the model. In an OLS model it is conventional to test whether coefficients are different from zero using a t test. Carrying out such tests in GWR is perhaps a little more contentious and raises the problem of multiple significance testing. It would be inappropriate to compute local t statistics and carry out 174 individual significance tests. Not only are the local results highly dependent, but the problem of carrying out multiple significance tests is that we would expect, with a 5% level of significance, that some 8 or 9 would be significant at random. Fotheringham et al (2002) suggest using a Bonferroni correction to the significance level; this may well be overly conservative, and a test procedure such as the Benjamini-Hochberg False Discovery Rate might be more appropriate (Thissen et al (2002)). However, answers to these problems continue to be the subject of research and future publication.

Further Reading

The definitive text on GWR is:

Fotheringham, AS, Brunsdon, C, and Charlton, ME, 2002, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Chichester: Wiley

A useful text on model selection is:

Burnham, KA and Anderson, DR, 2002, *Model Selection and Multimodel Inference: a practical information-theoretic approach*, 2nd edition, New York: Springer

An excellent text on data issues is:

Belsley, DA, Kuh, E and Welsch, R (1980), *Regression Diagnostics: identifying influential data and sources of collinearity*, Hoboken, NJ: Wiley

An implementation of the Benjamini-Hochberg False Discovery Rate procedure:

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