GEOGRAPHICALLY WEIGHTED REGRESSION

Martin Charlton[1]

Stewart Fotheringham[1]

Chris Brunsdon[2]

1. National Centre for Geocomputation
National University of Ireland Maynooth
Maynooth
Co. Kildare
IRELAND

2. Department of Geography
University of Leicester
Leicester
UK

The first two authors acknowledge generous funding from Science Foundation Ireland which helped create the National Centre for Geocomputation

ESRC National Centre for Research Methods
NCRM Methods Review Papers
NCRM/006

1

Introduction

1.1 Overview

This text is written as a follow-up to a two-day workshop on Geographically Weighted Regression (GWR) held at the University of Leeds, June 2005. The aim of this text is both to introduce the reader to the basic concept of GWR through several empirical examples and also to demonstrate how to run GWR with software specifically written for this purpose. The software, GWR 3.0, is available from the authors and details can be found at: http://ncg.nuim.ie/GWR It is highly recommended that this software be used in conjunction with this text. It should be noted that GWR 3.0 produces a set of localised statistics that can be imported into other software for mapping.

1.2 Introduction

By far the most common statistical modelling technique used in the social sciences is that of regression. In standard applications of regression, a dependent variable is linked to a set of independent variables with one of the main outputs of regression being the estimation of a parameter that links each independent variable to the dependent variable. A major problem with this technique when applied to spatial data is that the processes being examined are assumed to be constant over space – that is, one model fits all. Geographically Weighted Regression (GWR) is a statistical technique developed by the authors that allows the modelling of processes that vary over space. GWR results in a set of local parameter estimates for each relationship which can be mapped to produce a parameter surface across the study region. In this way, GWR provides valuable information on the nature of the processes being investigated and supersedes traditional global types of regression modelling.

There are at least three reasons to suspect that relationships will vary over space. The *first* and simplest is that there will inevitably be spatial variations in observed relationships caused by random sampling variations. The contribution of this source of spatial non-stationarity is not usually of great interest in itself but it does need to be recognised and accounted for if we are to identify other, more interesting, sources of spatial non-stationarity. That is, we are only interested in relatively large variations in parameter estimates that are unlikely to be due to sampling variation alone.

The *second* reason is that the relationships might be intrinsically different across space. Perhaps, for example, there are spatial variations in people's attitudes or preferences or there are different administrative, political or other contextual issues that produce different responses to the same stimuli over space. It is difficult to conjecture an example of this cause of spatial non-stationarity in physical geography where the relationships being measured tend to be governed by laws of nature. The idea that human behaviour can vary intrinsically over space is consistent with post-modernist beliefs on the importance of place and locality as frames for understanding such behaviour. Those

who hold such a view sometimes criticise quantitative analysis in geography as having little relevance to 'real-world' situations where relationships are very complex and possibly highly contextual. Local statistical indicators address this criticism by recognising such complexity and attempting to describe it (see Fotheringham, 2006, for further development of this argument).

The *third* reason why relationships might exhibit spatial non-stationarity is that the model from which the relationships are measured is a gross misspecification of reality and that one or more relevant variables are either omitted from the model or are represented by an incorrect functional form. This view, more in line with the positivist school of thought, assumes a global statement of behaviour can be made (and hence is applicable to relationships in physical as well as human geography) but that the structure of our model is not sufficiently well-formed to allow us to make it. In a nutshell, can all contextual effects be removed by a better specification of individual level effects? (Hauser, 1970). If model misspecification is the cause of parametric instability, the calculation and subsequent mapping of local statistics is useful in order to understand the nature of the misspecification more clearly.

In the remainder of this document we describe the basic elements of GWR and the use of various spatial weighting schemes before describing in detail the use of software for GWR. We then describe the use of Arcmap for visualising the results of running the GWR software and conclude with a worked example based on an examination of the determinants of educational attainment in the state of Georgia.

1.3 GWR Basics

Consider a global regression model given by:

$$y_i = a_0 + \Sigma_k \, a_k x_{ik} + \varepsilon_i$$

In the calibration of this model, one parameter is estimated for the relationship between each independent variable and the dependent variable and this relationship is assumed to be constant across the study region. The estimator for the parameters in this model is:

$$\mathbf{a} = (\mathbf{X}^t \, \mathbf{X})^{-1} \mathbf{X}^t \, \mathbf{y}$$

where **a** represents the vector of global parameters to be estimated, **X** is a matrix of independent variables with the elements of the first column set to 1, and **y** represents a vector of observations on the dependent variable. GWR is a relatively simple technique that extends the traditional regression framework by allowing local rather than global parameters to be estimated so that the model is rewritten as:

$$y_i = a_0(u_i,v_i) + \Sigma_k \, a_k(u_i,v_i) \, x_{ik} + \varepsilon_i$$

where (u_i,v_i) denotes the co-ordinates of the ith point in space and $a_k(u_i,v_i)$ is a realisation of the continuous function $a_k(u,v)$ at point i (Brunsdon *et al.* 1996, 1998a, 1998b; Fotheringham *et al.* 1996, 1997a, 1997b, 1998, 1999). That is, we allow there to be a continuous surface of parameter values and measurements of this surface are taken at certain points to denote the spatial variability of the surface. Note that the global model is a special case of the GWR model in which the parameter surface is assumed to be constant over space.

In the calibration of the GWR model it is assumed that observed data near to point i have more of an influence in the estimation of the $a_k(u_i,v_i)$'s than do data located farther from i.

In essence, the equation measures the relationships inherent in the model *around each point i*. Hence weighted least squares provides a basis for understanding how GWR operates. In GWR an observation is weighted in accordance with its proximity to point i so that the weighting of an observation is no longer constant in the calibration but varies with i. Data from observations close to i are weighted more than data from observations farther away. Algebraically, the GWR estimator is:

$$\mathbf{a}(u_i,v_i) = (\mathbf{X}^t \mathbf{W}(u_i,v_i)\, \mathbf{X})^{-1} \mathbf{X}^t \mathbf{W}(u_i,v_i)\, \mathbf{y}$$

where $\mathbf{W}(u_i,v_i)$ is an n by n matrix whose off-diagonal elements are zero and whose diagonal elements denote the geographical weighting of observed data *for point i.* That is,

$$\mathbf{W}(u_i,v_i) = \begin{matrix} w_{i1} & 0 & 0 & \dots & 0 \\ 0 & w_{i2} & 0 & \dots & 0 \\ 0 & 0 & w_{i3} & \dots & 0 \\ . & . & . & \dots & . \\ 0 & 0 & 0 & \dots & w_{in} \end{matrix}$$

where w_{in} denotes the weight of the data at point n on the calibration of the model around point i. Clearly, these weights will vary with i which distinguishes GWR from traditional Weighted Least Squares where the weighting matrix is constant. Below we describe how these weights can be defined.

It should be noted that as well as producing localised parameter estimates, the GWR technique described above will produce localised versions of all standard regression diagnostics including goodness-of-fit measures such as r-squared. The latter can be particularly informative in understanding the application of the model being calibrated and for exploring the possibility of adding additional explanatory variables to the model. It is also useful to note that the points for which parameters are locally estimated in GWR need not be the points at which data are collected: estimates of parameters can be obtained for any location. Hence, in systems with very large numbers of data points, GWR estimation of local parameters can take place at pre-defined intervals such as at the intersections of a grid placed over the study region. Not only does this reduce computing time but it can also be beneficial for mapping the results.

1.4 Weighting Schemes

Until this point, it has merely been stated in GWR that $\mathbf{W}(u_i,v_i)$ is a weighting scheme based on the proximity of i to the sampling locations around i without an explicit relationship being stated. The choice of such a relationship is now considered. Firstly, consider the implicit weighting scheme of the traditional global OLS model which is:

$$\begin{matrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \end{matrix}$$

$$\mathbf{W}(u_i,v_i) = \begin{matrix} 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \ldots & 1 \end{matrix} .$$

That is, the global model is equivalent to a local model in which each observation has a weight of unity so that there is no spatial variation in the estimated parameters. An initial step towards weighting based on locality might be to exclude from the calibration of the model at location i observations that are further than some distance d from i. Suppose i represents a calibration point and j represents a data point. Then, the weights could be defined as:

$$w_{ij} = 1 \text{ if } d_{ij} \leq d$$

$$w_{ij} = 0 \text{ otherwise}$$

so that the diagonal elements would be 0 or 1 depending on whether or not the above criterion is met. Examples of the use of a discrete weighting function in GWR are provided in Fotheringham *et al.* (1996) and Charlton *et al.* (1997).

However, this discrete spatial weighting function does not reflect actual geographical processes very well because it suffers from the problem of discontinuity. As i varies around the study area, the regression coefficients could change drastically as one sample point moves into or out of the circular buffer around i which defines the data to be included in the calibration for location i. Although sudden changes in the parameters over space might genuinely occur, in this case changes in their estimates would be artefacts of the arrangement of sample points, rather than any underlying process in the phenomena under investigation. One way to combat this is to specify w_{ij} as a continuous function of d_{ij}, the distance between i and j. One obvious choice is to define the diagonal elements of the weighting function by:

$$w_{ij} = \exp(-d_{ij}^2 / h^2)$$

where h is referred to as the bandwidth. If i and j coincide (that is, i also happens to be a point in space at which data are observed), the weighting of data at that point will be unity. The weighting of other data will decrease according to a Gaussian curve as the distance between i and j increases. In the latter case the inclusion of data in the calibration procedure becomes 'fractional'. For example, in the calibration of a model for point i, if $w_{ij} = 0.5$ then data at point j contribute only half the weight in the calibration procedure as data at point i itself. For data a long way from i the weighting will fall to virtually zero, effectively excluding these observations from the estimation of parameters for location i.

To this stage, it is assumed that the spatial weighting function is applied equally at each calibration point. In effect, this is a global statement of the weight-distance relationship and as such it suffers from the potential problem that in some parts of the region, where data are sparse, the local regressions might be based on relatively few data points. To offset this potential problem, spatially adaptive weighting functions can be incorporated into GWR. These would have relatively small bandwidths in areas where the data points are densely distributed and relatively large bandwidths where the data points are sparsely distributed. The following weighting function produces spatially adaptive kernels.

$$w_{ij} = [1 - (d_{ij} / h_i)^2]^2 \quad \text{if } d_{ij} < h_i$$

$= 0$ otherwise

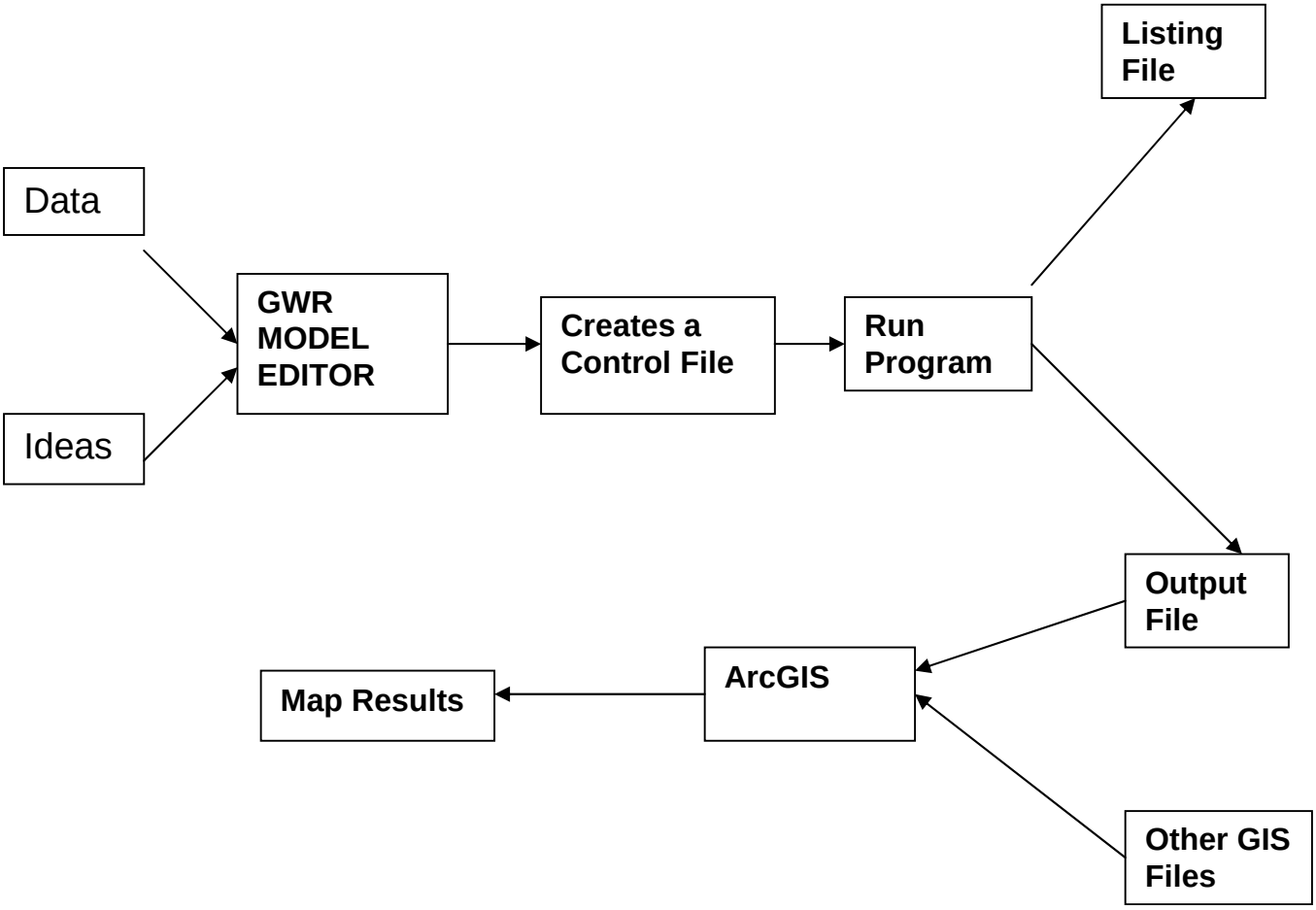
where h_i is the Nth nearest neighbour distance from i.

Whatever the specific weighting function employed, the essential idea of GWR is that for each point i sampled observations near to i have more influence in the estimation of the parameters for point i than do sampled observations farther away. Obviously, whichever weighting function is selected, the estimated parameter surfaces will be, in part, functions of the definition of that weighting function. For example, as h tends to infinity (no distance-decay), the weights tend to one for all pairs of points so that the estimated parameters become uniform and GWR becomes equivalent to OLS. Conversely, as the bandwidth becomes smaller, the parameter estimates will increasingly depend on observations in close proximity to i and hence will have increased variance. The problem is therefore how to select the optimal bandwidth and this is described below.

Once a weighting function has been selected and calibrated, the output from GWR will be a set of local parameter estimates for each relationship in the model. Because these local estimates are all associated with specific locations, each set can be mapped to show the spatial variation in the measured relationship. Similarly, local measures of standard errors and goodness-of-fit statistics are obtained. Given that, as we identified earlier, there might be different causes of spatial non-stationarity, one of which is random sampling variation, it is useful to ask the question: "Does the set of local parameter estimates exhibit significant spatial variation?" Below we describe techniques to answer this question.

1.5 Software for GWR and its Operation

Specialised software to undertake GWR, GWR 3.0, has been written by the authors. The following diagram summarises the basic operation of GWR 3.0 and how its outputs are linked to a GIS.



The user supplies a data file plus ideas on what form of model to calibrate into the user-friendly GWR Model Editor which is completed in a series of 'Windows-style' menus and tick boxes. Unseen to the user, this creates a control file for a large FORTRAN program which produces two types of output. A Listing File is written to the screen and an Output File is saved in the user's workspace. This latter file contains location-specific parameter estimates and other diagnostics which can be read into a GIS (along with other spatially referenced data) for mapping.

2

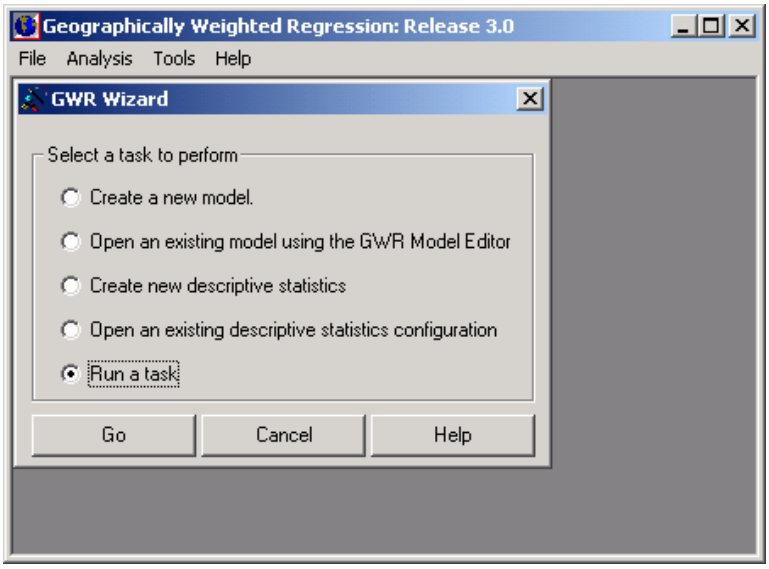
A Primer on Running GWR3

2.1 Introduction

This section shows how to set up and run a GWR model using the Visual Basic GWR Model Editor. There are several different varieties of regression model that can be run – here we will assume that you wish to run a geographically weighted regression with a Gaussian error term. This is the geographically weighted equivalent of an ordinary least squares regression, such as you might find in SPSS and is perhaps the most frequently encountered application of GWR.

The GWR software will be located in a folder on your machine or on a network. The folder is usually named GWR3, although it may be installed in a different folder on your system. There are two binaries in this folder. The name of the GWR Model Editor is GWR30.exe. If your system is set up to allow this you can create a link to the file GWR30.exe from the desktop or in a toolbar which will create a GWR icon. There is also a subfolder named SampleData which contains some test data for the software. The appropriate icon is selected to run the program. We assume that you will place your own data and the results from any analyses on those data in a different folder.

The main GWR program window is shown on the right; it has four items in the menu bar, '**File**', '**Analysis**', **Tools**' and '**Help**'. The program assumes that the user will wish to proceed with one of five initial options, and provides a 'Wizard' for guidance through the processes.



2.2 Model Specification

The general outline of specifying a GWR model is shown below. The actual program that computes the GWR is a FORTRAN program, and the software you are using is a front end to guide you through the following steps:

1. Select a task
2. Select a data file
3. Decide where to estimate the parameters

4. Specify the name of the parameter estimate file
5. Use the Model Editor to:
    5.1 Title the run
    5.2 Specify the dependent variable
    5.3 Specify the independent variable(s)
    5.4 Specify the data point location variables
    5.5 Specify the weighting scheme
    5.6 Specify the calibration method
    5.7 Specify the type of parameter estimate file
    5.8 Save the model control file
    5.9 Run the model
6. Examine the diagnostics

Following this you import the parameter estimate file into a mapping package so that you can examine any spatial variation in parameter estimates.

2.3 Data Organisation

The data file for GWR is an ASCII file which will normally have the filetype of .dat or .csv. The assumptions in the software about the organisation of the data are as follows:

1. The first line of the data file is a comma separated list of the names of the variables in the remainder of the file
2. The variable names should not contain any spaces
3. The variable names should be no more than 8 characters in length
4. The variable names should be formed from upper and lower case alphabetic characters and the numbers 0 … 9 inclusive
5. The only other character which is allowed is the underscore (_)
6. The remaining lines in the file contain the data
7. There are as many lines as there are observations (“data points”)
8. Each line contains the same number of attributes as there are variables
9. Attributes values are separated by commas
10. All attributes are numeric, and may be signed. Unsigned data are treated as positive
11. At least one of the attributes will be a dependent variable
12. There are two variables which specify the location of each data point

As an example, here are the first 11 lines of the data file for the Georgia educational attainment data to be used in the following labs:

```
ID,Latitude,Longitud,TotPop90,PctRural,PctBach,PctEld,PctFB,PctPov,PctBlack
13001,31.753389,-82.285580,15744,75.6,8.2,11.43,0.635,19.9,20.76
13003,31.294857,-82.874736,6213,100.0,6.4,11.77,1.577,26.0,26.86
13005,31.556775,-82.451152,9566,61.7,6.6,11.11,0.272,24.1,15.42
13007,31.330837,-84.454013,3615,100.0,9.4,13.17,0.111,24.8,51.67
13009,33.071932,-83.250851,39530,42.7,13.3,8.64,1.432,17.5,42.39
13011,34.352696,-83.500539,10308,100.0,6.4,11.37,0.340,15.1,3.49
13013,33.993471,-83.711811,29721,64.6,9.2,10.63,0.922,14.7,11.44
13015,34.238402,-84.839182,55911,75.2,9.0,9.66,0.816,10.7,9.21
13017,31.759395,-83.219755,16245,47.0,7.6,12.81,0.332,22.0,31.33
```

If you have been using ArcMap to integrate your data for an analysis, you can export a .dbf file as a .txt file. This can be renamed in the Explorer. When ArcGIS does this it places quotes around the variable names. These are not however stripped off by the FORTRAN program so the files will need further editing. You can also create .csv files in Excel (save your data in comma-separated variable form), Notepad, and other

applications capable of writing ASCII files. Do not name the first variable ID – Excel will assume your file is a wrongly formatted SYLK file and will refuse to open it!

2.4 Parameter Estimate Files

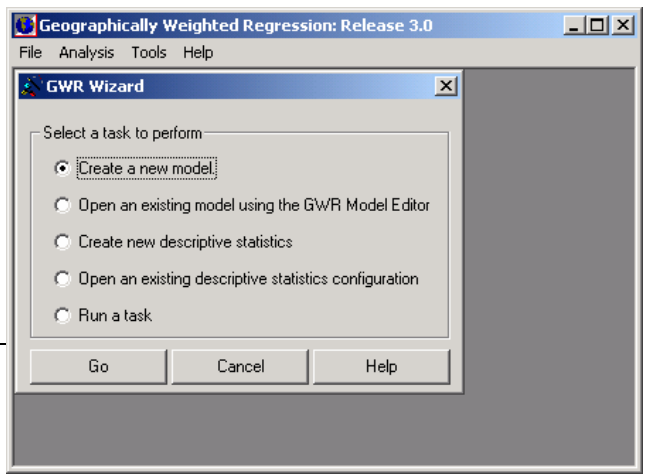
The output from GWR can be voluminous. At every regression point there will be a set of parameter estimates, a set of associated standard errors, and some diagnostic statistics. For this reason we have decided to make these outputs available as a file which can then be post-processed.

The outputs are

PARM_1 … PARM_n	Values of the estimates of the parameters at each regression point. n is one more than the number of independent variables with PARM_1 containing the values of the intercept term.
SVAL_1 … SVAL_n	Values of the estimates of the standard errors of the parameters at each regression point. The numbering of these variables is as for the parameter estimate variables.
TVAL_1 … TVAL_n	Pseudo-t values
OBS	Observed y variable value
PRED	Predicted y variable value
RESID	Unstandardised residual
HAT	Leverage value
STDRES	Standardised residual
COOKSD	Cook's Distance
LOCRSQ	Pseudo-R^2 values

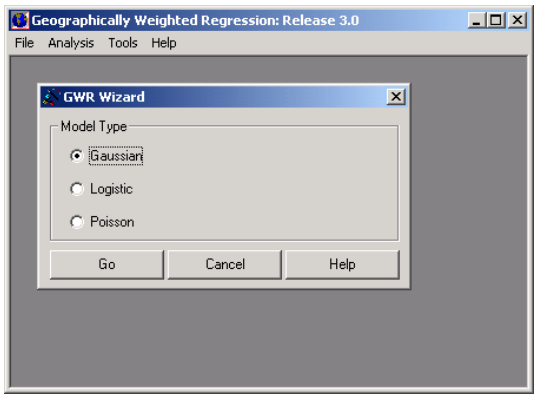
Three types of output format are available

1. ArcInfo uncompressed export format. This may be imported into ArcInfo to create a point coverage (where the coordinates of each point are those of the regression points). On a PC the coverage can be created using ArcToolBox (or Import71 if you are using ArcView 3.x). The filetype is .e00.
2. Comma-separated-variable format. This may be imported into Excel or SPSS for further processing. The names of the variables are included at the head of the file. Small numbers are not dealt with very elegantly and may be converted to scientific notation – ArcToolBox has trouble with these conversions. You should note that some numbers may be printed using scientific notation – the abscissa may be written as D+04 to represent 10^4. You will need to change these to E+04 otherwise Excel will treat them as text.
3. MapInfo Interchange Format. A .mif/.mid pair of files is created. These can be imported into MapINFO. The files are ASCII files and can be hand edited to remove any anomalies. This is somewhat experimental at the moment.



2.5 The Model Editor

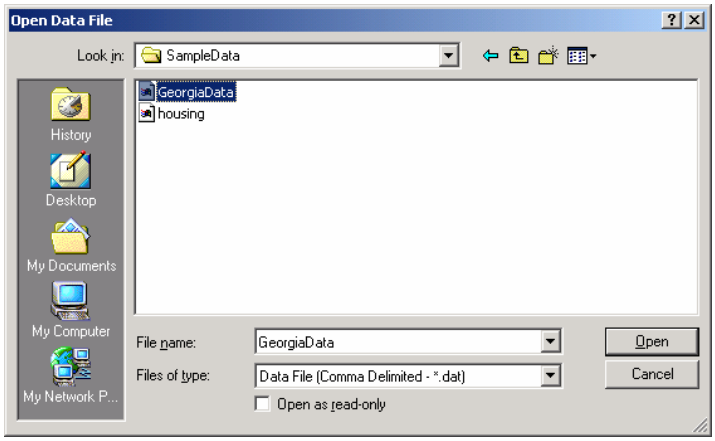
The first step is to create a new model to use with your data. If there is an existing model control file, then this can be run or the model editor can be invoked to change the variables or some other control parameters. Geographically weighted descriptive statistics may also be requested. At this point the user has the option of clicking on 'Go' to proceed with the new model, 'Cancel' to close the Wizard, or 'Help' to obtain some assistance on what to do next.



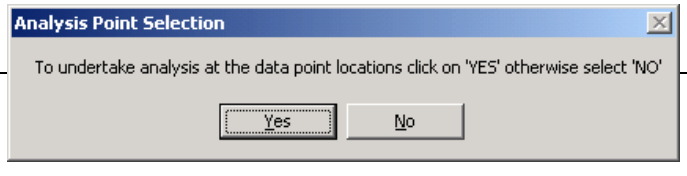
In this example, we have checked 'Create a new model' and clicked on 'Go'. We next need to determine what type of GWR model we wish to fit. In many cases this will be a Gaussian model. Select Gaussian and then 'Go'

Before a new model can be created, a data file must be selected from the data folder (see section 2.2 for details of the data file structure). The model editor will extract the names of the variables from the first line of the data file that is selected. We will base this description of the use of the Model Editor around the data concerning educational attainment in the counties of the state of Georgia, USA. These data have been described briefly in the previous section and further information is also given in section 4 (lab 2).

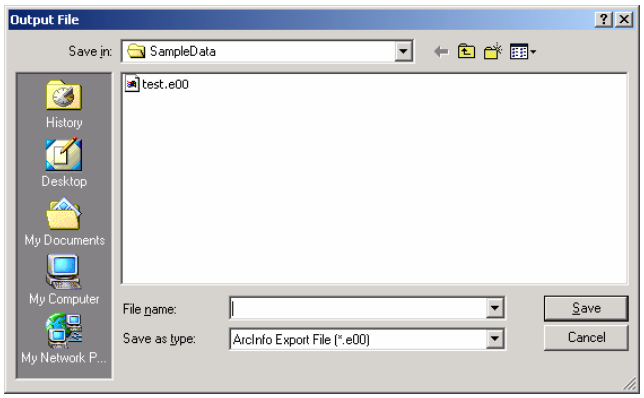
The form shown on the right will now appear – it is a standard Windows type 'File Open' form. There are only two data files in the data folder, one for the example we are using and one which is supplied with the software. Click on the relevant data file name to highlight it and click on 'Open' to proceed.



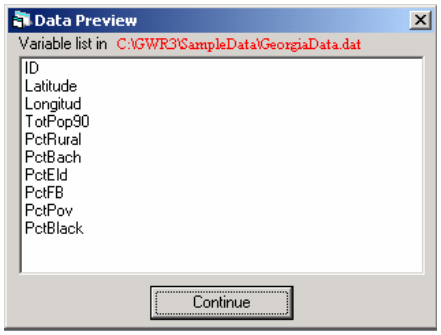
GWR estimates may be produced at locations other than those at which data are sampled. Locations where observations are recorded are referred to as *data points* (or *sample points*) and the locations at which the estimates are produced as *regression points*. In most instances, the regression points and the data points will be the same. However, there is an option in GWR3.x to produce estimates of local parameters at locations other than those at which data are recorded, for example at the mesh points of a regular grid. The prompt shown above allows the user to make this decision. In this instance, we click on '**Yes'**. Clicking 'No' brings up another form to allow the user to select a separate file of regression point locations. Note that using this second option means the automatic bandwidth selection and a range of diagnostic statistics will not be



available.

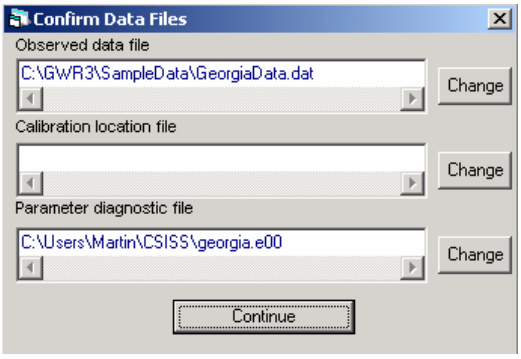


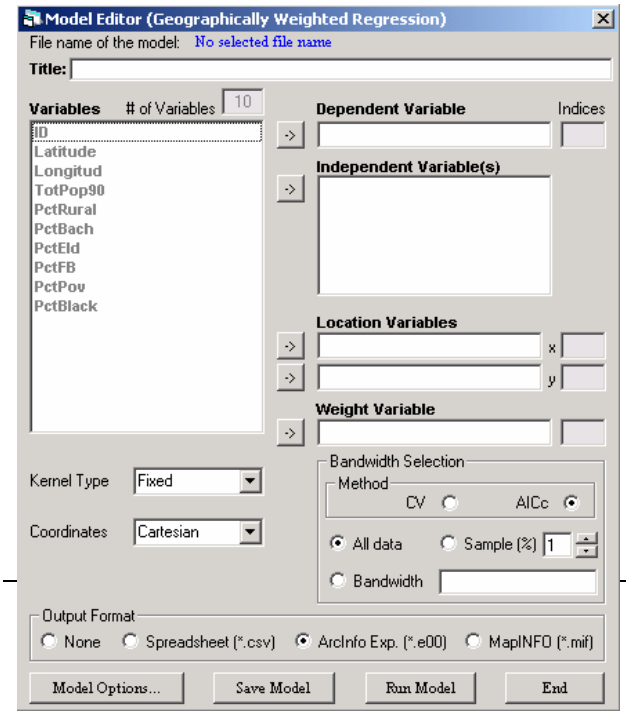
Next, the name of the file into which the results will be written must be specified. This file can be in one of several formats (comma-separated variable, ArcInfo uncompressed export, and MapInfo Interchange). The user also needs to specify the appropriate filetype - **.e00** for an ArcInfo export file, **.csv** for a comma-separated variable file, and **.mif** for a MapInfo Interchange File. You will need to navigate to the appropriate folder for the output file. Note that you cannot proceed without specifying a filename here.



The Data Preview window allows you to check that you have loaded the correct file – it lists the variable names which it has found in the first line of the file and gives you the location of the file.

As well as a check on the names of the variables, GWR also prints the names of the files which you selected thus far. If you have made a mistake, you have the option of correcting this before you continue. (Note: the various folder names we use here may be different from the ones you will use!). As we have decided to fit the model at the data points, the calibration location filename is blank.





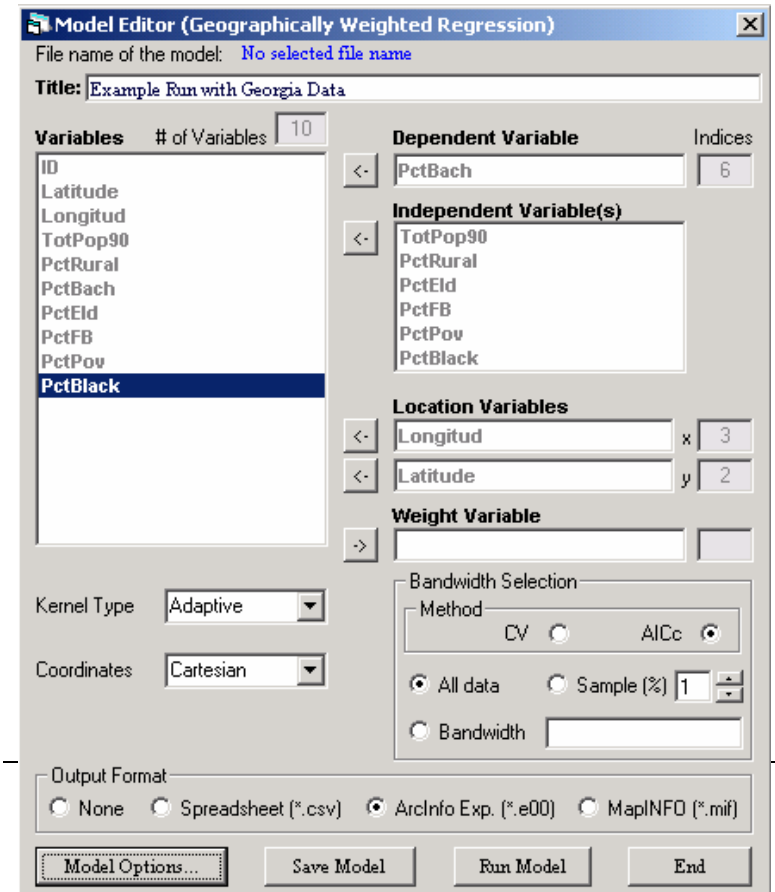
The Model Editor Window appears next and is shown on the left. It allows a GWR model to be created, saved and run. The **Title** box allows the user to input a title which will then appear in the output listing. The list of **Variables** is read automatically from the comma-

separated list on the first line of the data file that has been specified. From this, a **Dependent Variable** and one or more **Independent Variable(s)** are selected by highlighting the variable name and moving it with the appropriate arrow key. Next, two variables representing the coordinates of the data points, the **Location Variables** need to be assigned, and an optional **Weight Variable** can be selected. Note that this weight variable is *not* a geographical weight but simply allows data points to be weighted by some attribute reflecting different levels of uncertainty about the measurements taken across the data points. In most cases, this will be left empty. In the special case of Poisson regression, this variable will be used as an *offset variable*.

Once the variables have been selected, which essentially defines the model, the **Kernel Type** is chosen for the GWR. The choices are either 'Fixed' (Gaussian) or 'Adaptive' (bi-square). The kernel bandwidth is determined by either crossvalidation (**CV**) or AIC (**AICc**) minimisation (see Fotheringham *et al.* 2002 for more details of this). Alternatively, an *a priori* value for the bandwidth can be entered by clicking on the **Bandwidth** option and entering the bandwidth in the window. If you are using a Fixed kernel, the bandwidth needs to be specified in terms of the distance units used in your model. If you are using an Adaptive kernel, the bandwidth is specified as the number of data points in the local sample used to estimate the parameters. If you specify too small a bandwidth, you may get unpredictable results, or the program may be unable to estimate the model. With a very large data set (perhaps in excess of 10,000 observations), bandwidth selection can be made using a sample of data points in order to save time. This is achieved by clicking on **Sample (%)** and entering the desired percentage of the data used for the bandwidth selection procedure. The default is that the procedure will use **All data**.

If your coordinates are in some projected coordinate system (UTM, for example) then the Coordinate Type should be specified as Cartesian. If your measurements are in degrees of Latitude and Longitude, then select Spherical unless your study area is in a relatively low latitude or is relatively small when you can use Cartesian as the type. With Spherical coordinates, the distance computations in the geographical weighting use Great Circle distances. For more details on this, see Fotheringham *et al.* (2002).

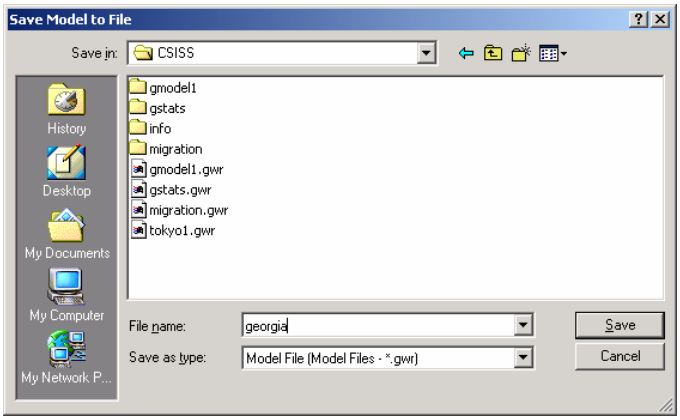
The Model Options include specifying the type of output required and the type of significance test to be employed on the local parameter estimates. Apart from the default output listings (described later), the user has the option of outputting **List Bandwidth Selection, List Predictions** and **List Pointwise Diagnostics**. Examples of these are shown below. The significance testing options are: **Monte Carlo,** or **None** (see above). Finally, the format of the output file needs to be specified: this should be compatible with the previous selection of an output filetype (see above).



A completed example of the GWR Editor is shown on the left. The dependent variable

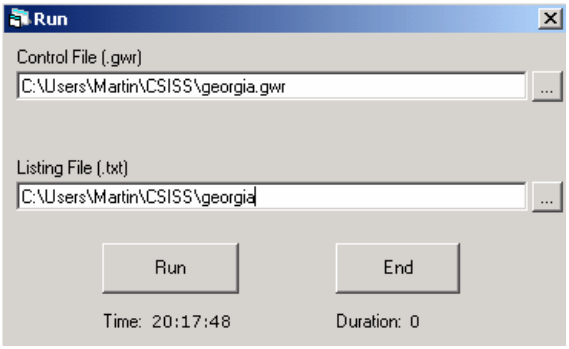
is the proportion of the county population with education to degree level. Suppose we are interested to see how this is related to total population within each county, the percentage rural, the percentage elderly, the percentage foreign born, the percentage below the poverty line and the percentage black. We would also like to see if there are any geographical variations in the relationships between educational attainment and these variables.

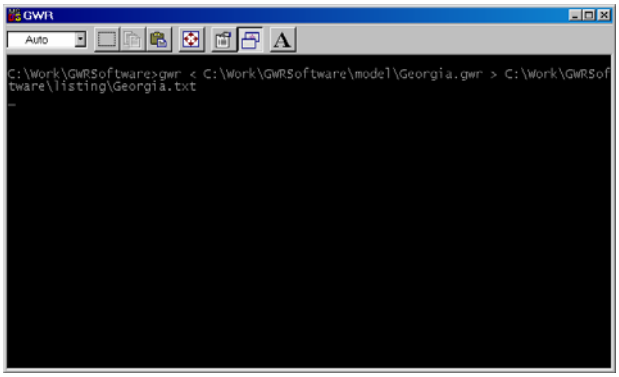
The sample point location variables are **Longitud** (x) and **Latitude** (y). There is no aspatial weight variable. We have chosen an adaptive kernel and the bandwidth will be chosen by AICc minimisation using all the data. A Monte Carlo significance testing procedure has also been selected for the local parameter estimates. Printing of a range of diagnostics has been requested and the output will be written to an ArcInfo export file. Some of the output will, by default, also be written to the screen in a listing file.



Before the model can be run, it must be saved. Clicking on **Save Model** will open the standard window shown on the left which depicts the contents of the model folder where the model control files are stored. Type the name of the file in the **Filename** box or click on an existing filename and then click on **Save**.

Once the model has been saved, it can be run. Simply click on the **Run** button in the Model Editor window and this brings up the form shown on the right. A name must be specified for the **Model Listing File (.txt)**. This file will be placed in the listing folder. To specify a filename click on the [...] button to the right of the filename box. Once this is done, click on the **Run** button. The model control file is now passed to the GWR program and the program is invoked and run in a DOS window as shown below. [1]

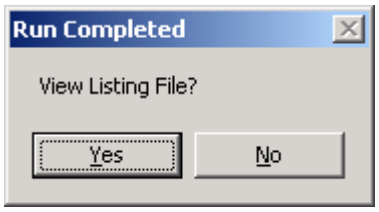




With small data sets and simple models, the program runs very quickly. For instance, calibrating a bivariate GWR model using the 159 counties of Georgia on a Pentium III PC took less time than it has taken to type this sentence. However, the time requirements increase rapidly as both model complexity and the number of data points increases. One solution to very slow run times is to use the option in the Model Editor which allows the

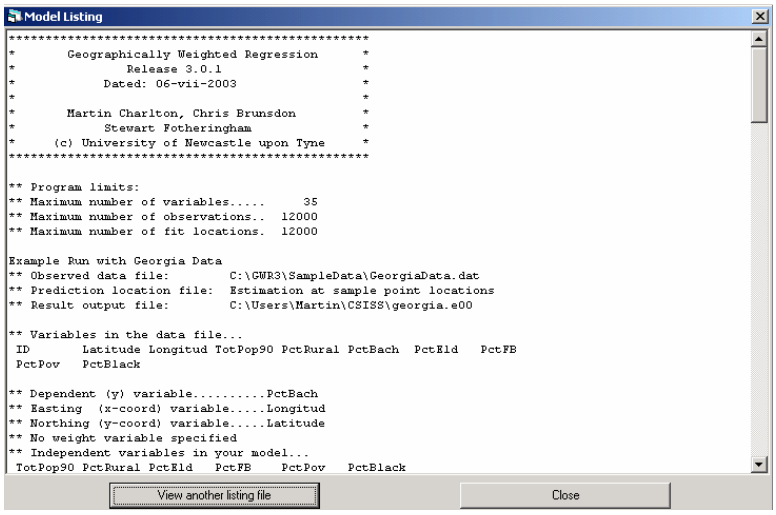
[1] You may need to make a small alteration in your Windows setup so that the DOS box closes on program termination.

user to supply a percentage of the data points on which to base the bandwidth selection procedure.



When the run has completed, the DOS window closes, and you are asked whether you wish to examine the listing file:

2.6 Printed Outputs



Once the program has run, the user is asked if the output listing is to be viewed. This listing appears in a separate window; an example of this for the Georgia educational attainment model is shown on the left. The user can scroll down the file to view other sections. The listing file is a text file with the filetype of **.txt** so that it can also be opened in MS Word or Notepad for viewing or printing.

Following a description of the model that has been calibrated, the first section of the output from GWR3 contains the parameter estimates and their standard errors from a global model fitted to the data. This is shown below.

```
*********************************************************
*           GLOBAL REGRESSION PARAMETERS                *
*********************************************************
Diagnostic information...
Residual sum of squares.........         1816.210715
Effective number of parameters..            7.000000
Sigma...........................            3.456697
Akaike Information Criterion....          855.443391
Coefficient of Determination....            0.645830

Parameter               Estimate               Std Err                     T
---------            ------------          ------------          ------------
Intercept        14.779297592328        1.705507562188         8.665630340576
 TotPop90         0.000023567534        0.000004746089         4.965675354004
 PctRural        -0.043878182061        0.013715372112        -3.199197292328
 PctEld          -0.061925096691        0.121460075458        -0.509839117527
 PctFB            1.255536084016        0.309690422174         4.054164886475
 PctPov          -0.155421764065        0.070388091758        -2.208069086075
 PctBlack         0.021917908085        0.025251694359         0.867977738380
```

There are two parts to the output from the global model. In the first panel, some useful diagnostic information is printed which includes the residual sum of squares, the number

of parameters in the global model, the standard error of the estimate, the Akaike Information Criterion (corrected version) and the coefficient of determination. In the second panel the matrix contains one line of information for each variable in the model. The columns are:

(a) the name of the variable whose parameter is being estimated
(b) the estimate of the parameter
(c) the standard error of the parameter estimate and
(d) the t statistic for the hypothesis that the true parameter value = 0.

These global results suggest that educational attainment is positively related to total population and percentage foreign born and is negatively related to percentage rural and percentage below the poverty line. Educational attainment does not appear to be related to the remaining two variables, percentage elderly and percentage black. The model replicates the data reasonably well (65% of the variance in educational attainment is explained by the model) but there are clearly some factors that are not captured adequately by the global model.

From this point, the output listing contains the results of the GWR. The first section is an optional calibration report which lists the calculated value of the criterion statistic at various bandwidths, as shown below. The utility of printing this section is to observe the speed of convergence and also to plot the results to see the shape of the convergence function (for more details on this, see Fotheringham *et al.* 2002). If the calibration report is not requested, the program will print only the optimal value of the bandwidth.

```
 Dependent mean=  10.9471693
 Number of observations, nobs= 159
 Number of predictors,   nvar= 6
 Observation Easting extent:   4.41947222
 Observation Northing extent:  4.20193577
*Finding bandwidth...
  ... using all regression points
 This can take some time...
 *Calibration will be based on  159 cases
*Adaptive kernel sample size limits:      10    159
 *AICc minimisation begins...
            Bandwidth                    AICc
          56.043532255000         952.763365832809
          84.500000000000         894.827422579517
         112.956467745000         872.102336481384
         130.543532046749         862.364688964195
         141.412935569545         859.863227740004
         148.130596397659         857.532739228028
         152.282339122725         856.699997311380
         154.848257244551         855.820209809022
** Convergence after     8 function calls
** Convergence: Local Sample Size=    155
```

The next section of the output presents diagnostics for the GWR estimation. There are two panels in this section. The first panel provides some general information on the model: it includes (a) a count of the number of data points or observations (b) the number of predictor variables (this is the number of columns in the design matrix) (c) the bandwidth for the type of kernel specified (here it is the number of nearest neighbours to be included in the bisquare kernel) and (d) the number of regression points. The second panel contains similar information to the corresponding panel for the global model. This includes (a) the residual sum of squares (b) the effective number of parameters, (c) the standard error of the estimate, (d) the Akaike Information Criterion (corrected) and (e) the coefficient of determination. The latter is constructed from a comparison of the predicted values from different models at each regression point and the observed values.

The coefficient has increased from 0.646 to 0.706 although an increase is to be expected given the difference in degrees of freedom. However, the reduction in the AIC from the global model suggests that the local model is better fit to the data even accounting for differences in degrees of freedom.

```
*********************************************************
*                GWR ESTIMATION                         *
*********************************************************
 Fitting Geographically Weighted Regression Model...
 Number of observations............ 159
 Number of independent variables... 7
  (Intercept is variable 1)
 Number of nearest neighbours...... 155
 Number of locations to fit model.. 159

Diagnostic information...
Residual sum of squares.........         1506.219121
Effective number of parameters..           12.814342
Sigma...........................            3.209901
Akaike Information Criterion....          839.193981
Coefficient of Determination....            0.706280
```

Casewise diagnostics can be also requested (as shown below for the first 10 observations in the Georgia data set). These include:

1. the observation sequence number
2. the observed data
3. the predicted data
4. the residual
5. the standardised residual
6. the local pseudo r-square
7. the influence and
8. Cook's D.

Whilst in general it might be helpful to look at a printout of these statistics, it is probably a little more useful to be able to map them: with a large data set you run the risk of being swamped in output. All of these statistics are saved automatically in the output results file so that requesting them in the listing file should be done judiciously. This panel is not available when the regression points are different from the data points.

```
*******************************************************
*                CASEWISE DIAGNOSTICS                 *
*******************************************************

   Obs       Observed      Predicted       Residual   Std Resid    R-Square   Influence    Cook's D
 ----- -------------- -------------- -------------- ----------- ----------- ----------- -----------
     1        8.20000        9.26692       -1.06692   -0.258875    0.819218    0.021879    0.000117
     2        6.40000        7.33714       -0.93714   -0.232802    0.820589    0.066868    0.000303
     3        6.60000        8.70596       -2.10596   -0.525272    0.819776    0.074367    0.001730
     4        9.40000        8.11559        1.28441    0.319607    0.840207    0.069997    0.000600
     5       13.30000       13.58140       -0.28140   -0.070091    0.839357    0.071855    0.000030
     6        6.40000        8.79625       -2.39625   -0.591102    0.844322    0.053656    0.001546
     7        9.20000       11.61571       -2.41571   -0.587443    0.846859    0.026203    0.000725
     8        9.00000       11.61646       -2.61646   -0.636924    0.852840    0.028236    0.000920
     9        7.60000       10.26846       -2.66846   -0.654270    0.826147    0.042107    0.001468
    10        7.50000        9.48755       -1.98755   -0.489605    0.822446    0.051028    0.001006
```

Another optional set of information that can be printed to the screen concerns the predicted values (as shown below for the first 10 observations in the Georgia data set). If this option is selected, the following data are printed to the screen:

1. **Obs** the sequence number of the observation
2. **Y(i)** the observed value
3. **Yhat(i)** the predicted value
4. **Res(i)** the residual
5. **X(i)** the x-coordinate of the regression point
6. **Y(i)** the y-coordinate of the regression point and
7. **F/T** an indicator of whether the matrix inverse was computed using either the Gauss-Jordan method (F) or a generalised inverse (T). The latter is only used if there is severe multicollinearity in the design matrix

This set of output is not available when the regression points are different from the sample points.

```
Predictions from this model...
   Obs        Y(i)    Yhat(i)      Res(i)        X(i)        Y(i)
    1       8.200       9.267      -1.067     -82.286      31.753  F
    2       6.400       7.337      -0.937     -82.875      31.295  F
    3       6.600       8.706      -2.106     -82.451      31.557  F
    4       9.400       8.116       1.284     -84.454      31.331  F
    5      13.300      13.581      -0.281     -83.251      33.072  F
    6       6.400       8.796      -2.396     -83.501      34.353  F
    7       9.200      11.616      -2.416     -83.712      33.993  F
    8       9.000      11.616      -2.616     -84.839      34.238  F
    9       7.600      10.268      -2.668     -83.220      31.759  F
   10       7.500       9.488      -1.988     -83.232      31.274  F
```

Next in the output listing is a panel of results of an ANOVA in which the global model is compared with the GWR model. The ANOVA tests the null hypothesis that the GWR model represents no improvement over a global model. The results are shown below where it can be seen that the F test suggests that the GWR model is a significant improvement on the global model for the Georgia data.

```
*********************************************************
*                      ANOVA                            *
*********************************************************
        Source                    SS        DF            MS             F
OLS Residuals                1816.2      7.00
GWR Improvement               310.0      5.81         53.3150
GWR Residuals                1506.2    146.19         10.3035        5.1745
```

The main output from GWR is a set of local parameter estimates for each relationship. Because of the volume of output these local parameter estimates and their local standard errors generate, they are not printed in the listing file but are automatically saved to the output file. However, as a convenient indication of the extent of the variability in the local parameter estimates, a 5-number summary of the local parameter estimates is printed. For the Georgia data, this is shown in below. The 5-number summary of a distribution presents the median, upper and lower quartiles, and the minimum and maximum values of the data. This is helpful to get a 'feel' for the degree of spatial non-stationarity in a relationship by comparing the range of the local parameter estimates with a confidence interval around the global estimate of the equivalent parameter.

Recall that 50% of the local parameter values will be between the upper and lower quartiles and that approximately 68% of values in a normal distribution will be within $\pm$ 1 standard deviations of the mean. This gives us a reasonable, although very informal, means of comparison. We can compare the range of values of the local estimates

between the lower and upper and quartiles with the range of values at ±1 standard deviations of the respective global estimate (which is simply 2 x S.E. of each global estimate). Given that 68% of the values would be expected to lie within this latter interval, compared to 50% in the inter-quartile range, if the range of local estimates between the inter-quartile range is greater than that of 2 standard errors of the global mean, this suggests the relationship might be non-stationary.

```
*********************************************************
*          PARAMETER 5-NUMBER SUMMARIES                 *
*********************************************************
   Label       Minimum  Lwr Quartile        Median  Upr Quartile       Maximum
Intrcept     12.620986     13.754251     15.823232     16.312238     16.489399
TotPop90      0.000014      0.000018      0.000022      0.000025      0.000028
PctRural     -0.060218     -0.051780     -0.039342     -0.031651     -0.025801
PctEld       -0.255508     -0.203092     -0.164197     -0.129393     -0.058400
PctFB         0.504876      0.825190      1.432738      2.003490      2.417666
PctPov       -0.204510     -0.164793     -0.110038     -0.056264     -0.004242
PctBlack     -0.036187     -0.013582      0.006294      0.031046      0.076566
```

As an example, consider the parameter estimates for the two variables PctEld (percentage elderly) and PctFB (percentage foreign born) in the Georgia study.

The global results provide the following information:

	S.E.	**2 x S.E.**
PctEld	0.121	0.242
PctFB	0.310	0.620

while the 5-number summary yields:

	Lower quartile	**Upper quartile**	**Range**
PctEld	-0.203	-0.129	0.074
PctFB	0.825	2.003	1.178

For PctEld the interquartile range of the local estimates is much less than 2 x S.E. of the global estimate indicating a stationary relationship.

For PctFB the interquartile range of the local estimates is much greater than 2 x S.E. of the global estimate indicating a non-stationary relationship.

Finally, we can examine the significance of the spatial variability in the local parameter estimates more formally by conducting a Monte Carlo test. The results of a Monte Carlo test on the local estimates indicates that there is significant spatial variation in the local parameter estimates for the variables PctFB and PctBlack. The spatial variation in the remaining variables is not significant and in each case there is a reasonably high probability that the variation occurred by chance. This is useful information because now in terms of mapping the local estimates, we can concentrate on the two variables, PctFB and PctBlack, for which the local estimates exhibit significant spatial non-stationarity. It is interesting to note that these results reinforce the conclusions reached above with the informal examination of local parameter variation for the variables PctEld and PctFB.

```
**********************************************
*                                            *
*   Test for spatial variability of parameters  *
*                                            *
**********************************************

Tests based on the Monte Carlo significance test
procedure due to Hope [1968,JRSB,30(3),582-598]

Parameter                  P-value
----------      ------------------
Intercept            0.22000
TotPop90             0.09000
PctRural             0.17000
PctEld               0.68000
PctFB                0.00000
PctPov               0.50000
PctBlack             0.00000
```

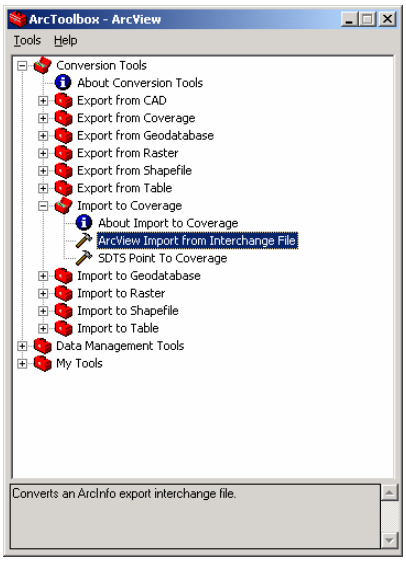
3

Visualising the Output from GWR with ArcMap

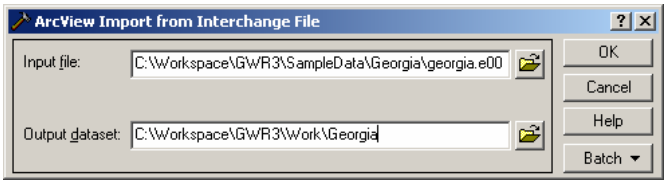
The main output from GWR is a set of localised parameter estimates and associated diagnostics. Unlike the single global values traditionally obtained in modelling, these local values lend themselves to being mapped. Indeed, with large data sets, mapping, or some other form of visualisation, is the only way to make sense of the large volume of output that will be generated. We now describe ways of visualising the output from GWR. Although we concentrate only on displays of the local parameter estimates, in many instances it might be instructive to plot other local statistics such as the influence and Cook's D statistics. Similarly, it might be useful to plot the local r-square statistic or the local standard deviation. No matter which local statistic is mapped, however, there is a choice of map types that can be employed. We now describe some of these briefly after first discussing mapping the results in a commonly used, PC-based, Geographic Information System (GIS), ArcMap.

3.1 Creating a Coverage

We assume that the user has available some software for visualising the results. Most commonly, this will be some mapping package, or preferably, a GIS in which both the results and the data can be manipulated. Saving the output file as either an uncompressed ESRI export file (.e00) or a MapInfo interchange file (.mif) means the output can be viewed relatively easily within a GIS. For instance, if we convert the .e00 file to a coverage it can be viewed in ArcMap. The conversion is carried out using the ArcToolBox program which is part of ArcMap.



First, start ArcToolBox[2] and in the Conversion Tools kit, select Import to Coverage. Then select ArcView Import from Interchange file. This brings up another dialog box which must be completed. The Input file is the ArcInfo Export File (also known as an Interchange file). Normally the file and path you specify are those which have already been specified in the GWR program. The output dataset (or coverage) is probably best located in the same folder. However, in

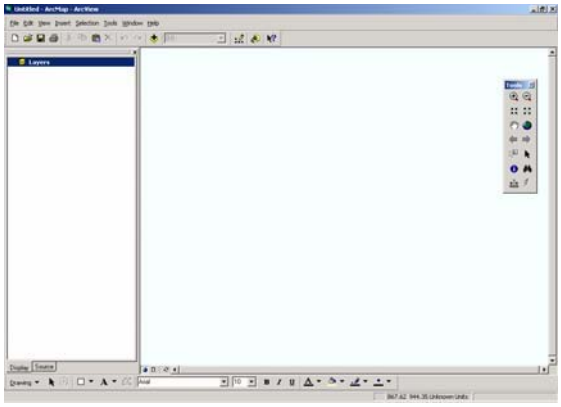


[2] The *modus operandi* for running ArcToolBox varies slightly between different versions of ArcMap.

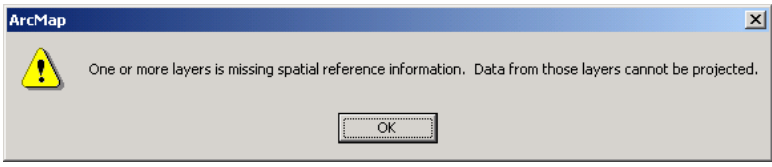
this lab the input file should be Georgia.e00 which you will find in the SampleData\Georgia folder; the output file should be located in your Work folder and named Georgia. When you have specified these, you click on [OK]. Wait a few moments until the software has finished the conversion – an hourglass will appear while conversion is taking place. Close the ArcToolBox application.

3.2 Visualizing your Coverage

Now you can start the ArcMap application and examine your data. The ArcMap window is as on the right (start with the *A new empty map* option). The converted coverage becomes a data layer. We shall add another data layer shortly. Click on the Add Data icon (it's the black cross on a yellow diamond icon in the menu bar).

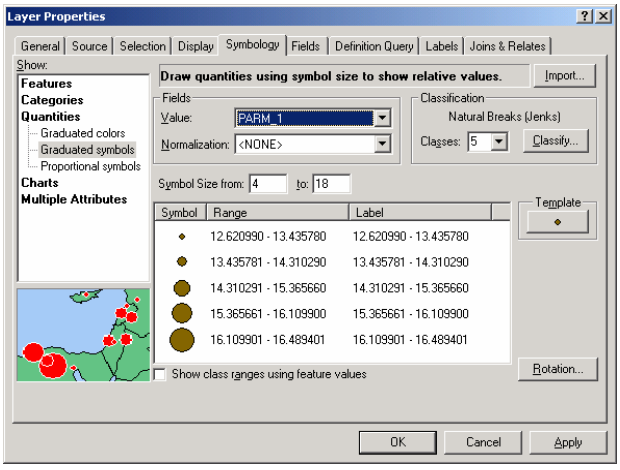


Navigate to your Work folder and click on the name of the coverage you have just created. In the case of this example, the coverage name is Georgia. You will get an error message which you can ignore for the time being.

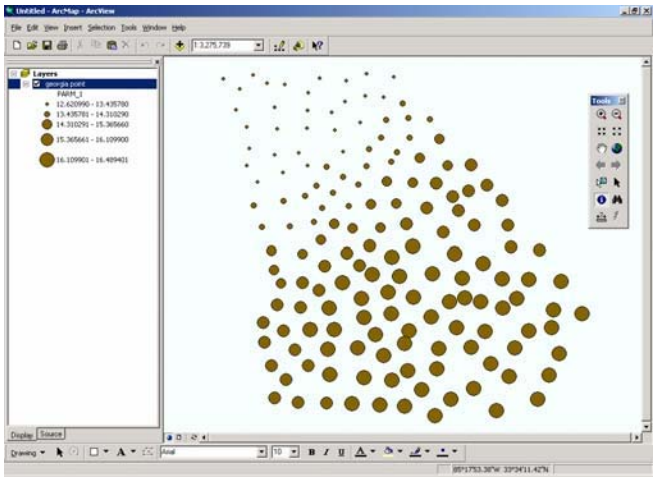


The points which you can see represent the locations of the regression points – in this case, they are the centroids of the counties of Georgia. To visualise the spatial variation in the Intercept term (this is called PARM_1 in the coverage).

1. Right click on the **georgia point** entry in the Table of Contents
2. Select Properties from the list (it's at the bottom)
3. Click on the Symbology tab
4. Select Quantities/Graduated Symbols from the Show: box
5. Select **PARM_1** from the Fields/Value dropdown list



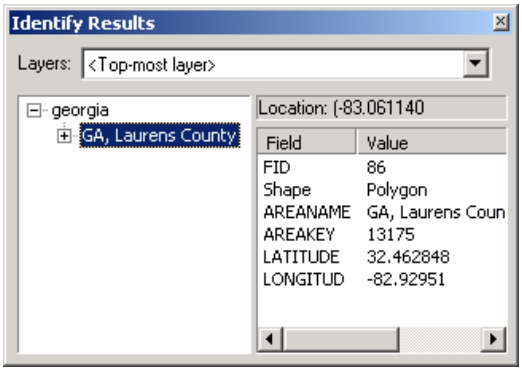
The completed dialog should be as above. If you click on [OK] you will be presented with a display of graduated circles as on the right. The circle size is related to the value of the intercept term. The resulting pattern suggests that in this case, there is a broad regional pattern with higher values of the intercept in southern Georgia, and lower values elsewhere.



You can use the Identify tool in the Toolbar list to click on one of the circles to bring up the values of all the attributes for that county.

3.3 Visualizing Variation with Shaded Polygons

This is all very well, but it might be desirable to attach the attributes to some polygons. How can this be achieved?

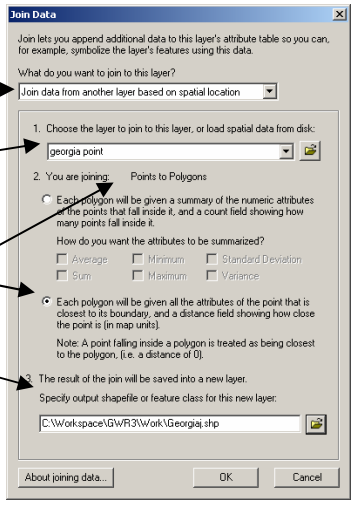


Click on the Add Data icon once more, and navigate to the SampleData\Georgia folder. You will find a data layer called **G_utm.shp**. Select this and click OK. ArcMap places this on the end of the Table of Contents, and assigns some default shading. You can use the Identify tool to find what attributes the polygons have. Here's a typical entry on the right. Unfortunately, the AREAKEY item is not present in the Georgia point coverage (the GWR output) so we cannot join the GWR output attribute table to the polygon attribute table using this key. We need to use a "spatial join" to match the attributes of the points with the attributes of the polygons.

3.4 Spatial Join

Click on **G_utm** in the Table of Contents, and then right click. Select Relates and Joins and then Joins… from the list of options. Complete the dialog that appears as below.

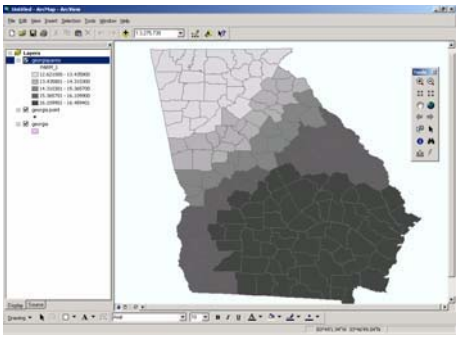
1. You are joining data from another layer based on spatial location
2. The layer from who which you wish to copy attributes to the polygon layer is the one you name in box 1
3. In 2, you are joining Points to Polygons; make sure the second option is checked
4. In box 3, navigate to your Work folder and name the output file georgiaj.shp



Click on [OK]

The join then takes place, and the shapefile (files of type .shp are called shapefiles) is added to the table of contents.

Use the following steps to create shading for the polygons. Here is the display for PARM_1 which illustrates the same spatial variation as in the previous example but in a slightly different form.



1 Right click on georgiaj
2 Click on the Symbology tab
3 Select Quantities/Graduated Colors
4 Select PARM_1 from Field/Value List
5 Click on OK

3.5 Checklist

It might be worthwhile looking back over the decisions we have had to make in carrying this out.

1. Check whether any of the data layers you need has projection information. If so, make a note of it and see the notes below on assigning a projection.
2. Convert the parameter estimate interchange file (.e00) to a coverage using ArcToolBox
3. In ArcMap add the parameter estimate coverage layer
4. Add in any other layers you might need
5. Carry out any spatial joins you need to do
6. Visualize the parameter estimate variation, either as points with graduated symbols, or as polygons with graduated colours.

Endnote: Assigning Projections

When mapping the results from *your own data sets*, depending on the source of your boundary files, you may find ArcMap automatically assigns a default projection if the coordinates look as though they might be latitude & longitude measurements. Checking the layer's properties, for example, you may find something like…

```
Data Type: Shapefile Feature Class
Shapefile: C:\GWR3\SampleData\georgia
Geometry Type: Polygon

Coordinate System:
GCS_Assumed_Geographic_1
Datum: D_North_American_1927
Prime Meridian: 0
```

This will cause a problem in that the output from the GWR program will not have this projection assigned and the two data sets are apparently not compatible. If such a problem occurs, you will need to assign the same projection to your GWR output coverage as ArcMap has assigned to your boundary data. This can be done as follows:

You will need to remove the parameter estimate coverage from the Table of Contents and assign a projection to it. Suppose the projection you wish to assign is that of geographic NAD 1927. The projection conversion is carried out in ArcToolBox. You may need to do this every time you Import an interchange file.

1. Remove the parameter estimate coverage from the Table of Contents
2. Start the ArcToolBox application
3. Select **Data Management Tools/Projections**
4. Select **Define Projection Wizard (coverages, grids, TINs)**
5. Check **define the coordinate system interactively**: click [Next]
6. Navigate to your work folder and select **Georgia**: click [Next]
7. Select **Geographic** as the dataset projection: click [Next]
8. Select **DD** as the Units parameter: click [Next]
9. Select **NAD 1927 (US-NADCON)** as the Datum: click [Next]
10. Check the settings and click [Finish]
11. Exit the ArcToolBox application

You will have noticed that the Wizard also allows you to copy the information about the projection from another coverage, so if you create a 'master' coverage and assign the projection information, then you can use this as the source for a copy.

4

An Example of GWR: Educational Attainment Data in Georgia

4.1 Introduction

In this section, data on educational attainment in Georgia will be used to demonstrate a typical GWR application. The idea is to predict the level of education attainment from some social attributes of the counties in the State of Georgia and then to map the variation in the local parameter estimates and some diagnostics.

4.2 The Modelling Process

Often, GWR is used in some larger data exploration and modelling exercise. Typically the steps either side of GWR may include the following

1. Prepare the data – this may involve, for example, Excel, SPSS, SAS, or a GIS program.
2. Model relationships in GWR: examine printed diagnostics
3. Save the parameter estimates in a suitable format
4. Import the parameter estimates into a GIS program
5. Display the parameter variation – further analysis
6. Display the diagnostic variation – further analysis

It should be stressed that this is not the only route, but it is the one illustrated here.

4.3 Choices in Model Specification

The model specification is central to the analysis. Having specified your data file, the GWR software will read the first line and extract the names of the variables. These names appear in the Variables box. At this stage, the model variables should be specified. Also, various other model specification options may be selected. A typical set of actions is outlined here:

1. Choose something like **Georgia Educational Attainment** as the title. This does not affect the analysis, but will appear in the GWR software output.
2. Select the dependent variable: In this example **PctBach** is used.
3. Select the independent variables: for the Georgia data these are **Totpop90**, **PctRural**, **PctEld**, **PctFB**, **PctPov**, and **PctBlack**
4. Select the location variables. Here **X** is the x variable and **Y** is the y variable.
5. Select the kernel type. Here it is **Adaptive**
6. Select appropriate Model Options; In this case bandwidth selection, predicted values and pointwise diagnostics will be listed.
7. In this case monte carlo significance testing, and bandwidth selection using the AIC (Akaike Information Criterion) are also selected.
8. Finally, the output format is specified to be Arc/INFO export format

4.4 Running the Model

The last set of actions creates a work file – this contains the exact details of the model specified as in the previous section. This model is saved in the work file – here a name such as **Georgia.gwr** might be appropriate. This model may then be run. At this stage a listing file (a name such as **Georgia.txt** might be appropriate here) must be specified – this contains the output generated by the program. Once this is done the model is run, and on completion, the output will be written into **Georgia.txt.** While the program is running a DOS window appears. The window title bar indicates when the program has finished and the command Exit appears in this window. The DOS window will then disappear. At this point the Run Completed form allows the listing file to be viewed.

4.5 Examining the Outputs

At the top of the listing file, some initial values are reported: for the Georgia data these are:

```
Dependent mean=  10.9471693
Number of observations, nobs= 159
Number of predictors,   nvar= 6
Observation Easting extent:  423741.688
Observation Northing extent:  471492
```

Also, the current values of the bandwidth and the associated AIC are printed on the output - these functions can be quite messy at times.

Since the adaptive approach was selected, bandwidths are specified in terms of nearest neighbours. For the Georgia data, the program has converged at 155 nearest neighbours. As there are only 159 counties, the GWR results may be fairly close to the global results – although in the GWR model the data are weighted by geographic location, so between the centre of the kernel and the 155th nearest neighbour the weighting has gradually fallen to zero.

The global model parameters and diagnostic statistics are also listed. The AIC for the global model is 855.44 and the global coefficient of determination is 0.65. This suggests that the global model is a reasonable one, although 35% of the variation in the dependent variable is from sources other than the ones in the model. The global parameters themselves show that there is a positive association with Population, Foreign Born and Black and a negative association with Rural, Elderly and Poverty. However the coefficients for Elderly and PctBlack are small enough for us to regard them as having no effect on the model ($t < \sim 1.96$).

Next the GWR parameters and diagnostics are listed. The AIC for this model is 840.07. This is less than that for the global model and this suggests that the GWR model is "better" at modelling the data. The coefficient of determination is a little higher at 0.72. Further down, the ANOVA is listed. The computed value of 5.01 is in excess of the critical value for F with 7.0 and 146.2 degrees of freedom suggesting rejection of the null hypothesis that the GWR represents no improvement over the global model. This conclusion is in line with the AIC results above. Finally, since a significance test was requested in the model specification, it is also possible to identify which parameters exhibit significant spatial variation.

4.6.1 Mapping the Results

In this stage of a typical analysis, the GWR coefficients are usually mapped. The Arc/INFO .e00 file saved as output from running the model has to be converted into an Arc/INFO coverage. The **ArcToolBox** utility which is part of the ArcMap system carries out this task. Assume the input file is georgiaout.e00 in a folder called Work, and the output file is gparms also in the Work folder, and that **ArcToolBox** created a coverage called **gparms** also in this folder. The ArcMap program can then display this coverage, using the Add Data icon. If this icon is clicked, then navigating to the Work folder and selecting gparms will add this coverage as a theme the ArcMap window.

Right-clicking this theme's name and selecting **Open Attribute Table** from the list of options shows that the theme table contains the values of:

the pointwise parameter estimates	(PARM_1…)
the pointwise standard errors	(SVAL_1…)
the pointwise pseudo-t values	(TVAL_1…)
the observed y value	OBS
the predicted y value	PRED
the residual	RESID
the standardised residual	STDRES
the trace of the hat matrix	HAT
Cook's D	COOKSD

There are 7 sets of data for the PARM, SVAL and TVAL items numbered thus

1. Intercept
2. Totpop90
3. PctRural
4. PctEld
5. PctFB
6. PctPov
7. PctBlack

PARM_1 contains the values of the Intercept term and SVAL_1 contains the values of the corresponding standard errors.

As an example of the use of ArcMap to show geographical variation in the parameter estimates, one can show the variation in the intercept term with proportional symbols located at the centroids of the regression points with the following steps:

1. Right click on gparms point again and select Properties, then
2. Select Symbology/Quantities/Graduated symbols
3. Select Value/Field: **Parm_1**
4. Click **Apply** to apply this symbolism to the data.

4.6.2 Adding Boundaries

Whilst it is clear from the map that the value of the intercept term increases gently from North West to South East Georgia, it will enhance the map to show some county boundaries.

1. Click on the Add Data icon

2. Highlight **g_utm.shp** in the SampleData\Georgia folder and click on Add:
3. Right click on G_utm to bring up Properties/Symbology, click in the middle of the Symbol (which will be colored as the polygons on the map)
4. In the Symbol selector change Options/Fill Color to **No Color**
5. Click on the various OKs to return to the map

4.6.3 Choropleth Mapping

The data refer to the counties of Georgia rather than point locations within them. It would be desirable to attach the parameter values to the attribute table for the county boundaries. One way of doing this is to use a spatial join between the gparms point coverage and the g_utm shapefile.

1. Right click on **g_utm** in the Table of Contents and select Joins and Relates/Joins…
2. In the first box under the question "What do you want to do to this layer?" Select "Join data from another layer based on spatial location"
3. Choice 1: the layer you wish to join will be gparms point
4. Choice 2: you are joining Points to Polygon. Check the second option here "Each polygon will be given all the attributes of the points…"
5. Choice 3: name the output layer gparmsj.shp in your Work folder
6. Click OK
7. gparmsj is then added as a layer to the Table of Contents. Use Properties/Symbology to assign suitable shading to display the parameter estimates.

An interesting diagnostic produced by the software diagnostic is the STDRES – this is the standardised residual. By setting Manual class breaks, with 5 classes at –1.96, 0, 1.96, 2.58 and 3.53, one can compare these to cumulative percentage points on the Normal distribution[3]. The counties of Clarke and Oconee have unusually high positive residuals; Hall and Clayton have rather large negative residuals. Thus, following through this example, it can be seen how to specify a model, examine its output, map the results and carry out geographical diagnostics and explorations.

[3] Those listed here correspond to the two tailed 5%, 1% and 0.1% points.

5

Concluding Comments

This document has outlined the underlying ideas of GWR as well as having shown how such an analysis may be carried out in practice. Of key importance here is the emphasis on statistical testing and diagnostics. In particular, although there are many situations in which the relationships between variables vary spatially, there are others where the relationships are stationary, and it is always important to verify that a GWR approach is justified. In particular, be aware that a badly-specified global model is unlikely to be 'rescued' simply by turning it into a GWR model with the same variables.

Also be aware that the correct *kind* of GWR model is being used for a given problem. Although this document focuses attention on the basic GWR model, which assumes that the dependent variable is ratio or interval scale and has a Normal distribution, there are other GWR models (not discussed here, but available in the software an discussed in Fotheringham *et al.* 2002) which can model count data (with a Poisson distribution) and binary data (with a logistic distribution). Work is also underway to model *categorical* data (using geographically weighted discriminant analysis), so that with the correct choice of model, the geographically weighted approach can be used to model interval, ratio, categorical or count data.

Further details on the use of GWR are given in Fotheringham *et al.* 2002.

6

References

Fotheringham, A S, Brunsdon, C and M Charlton 2002 *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships* Wiley: Chichester.

Fotheringham A S 2006 “Quantification, Evidence and Positivism”, Chapter 32 in *Approaches to Human Geography* eds. G. Valentine and S Aitken. Sage: London.