

GWR 3

Software for Geographically Weighted Regression

Martin Charlton

Stewart Fotheringham

Chris Brunson

Spatial Analysis Research Group
Department of Geography
University of Newcastle upon Tyne
Newcastle upon Tyne
ENGLAND NE1 7RU

Copyright Notice

© 2003. The contents of this manual are the copyright of the authors and may not be reproduced or used without their permission. This extends to the any software and data files distributed on CD.

Acknowledgements

The authors would like to acknowledge the important contributions of both Stamatis Kalogirou and Tomoki Nakaya in the development of GWR 3.0. Stamatis Kalogirou was responsible for writing the Visual Basic interface and Tomoki Nakaya provided invaluable assistance in the programming of the Poisson GWR and the Binary Logit GWR options.

Version 3.0.1 (November 2003)

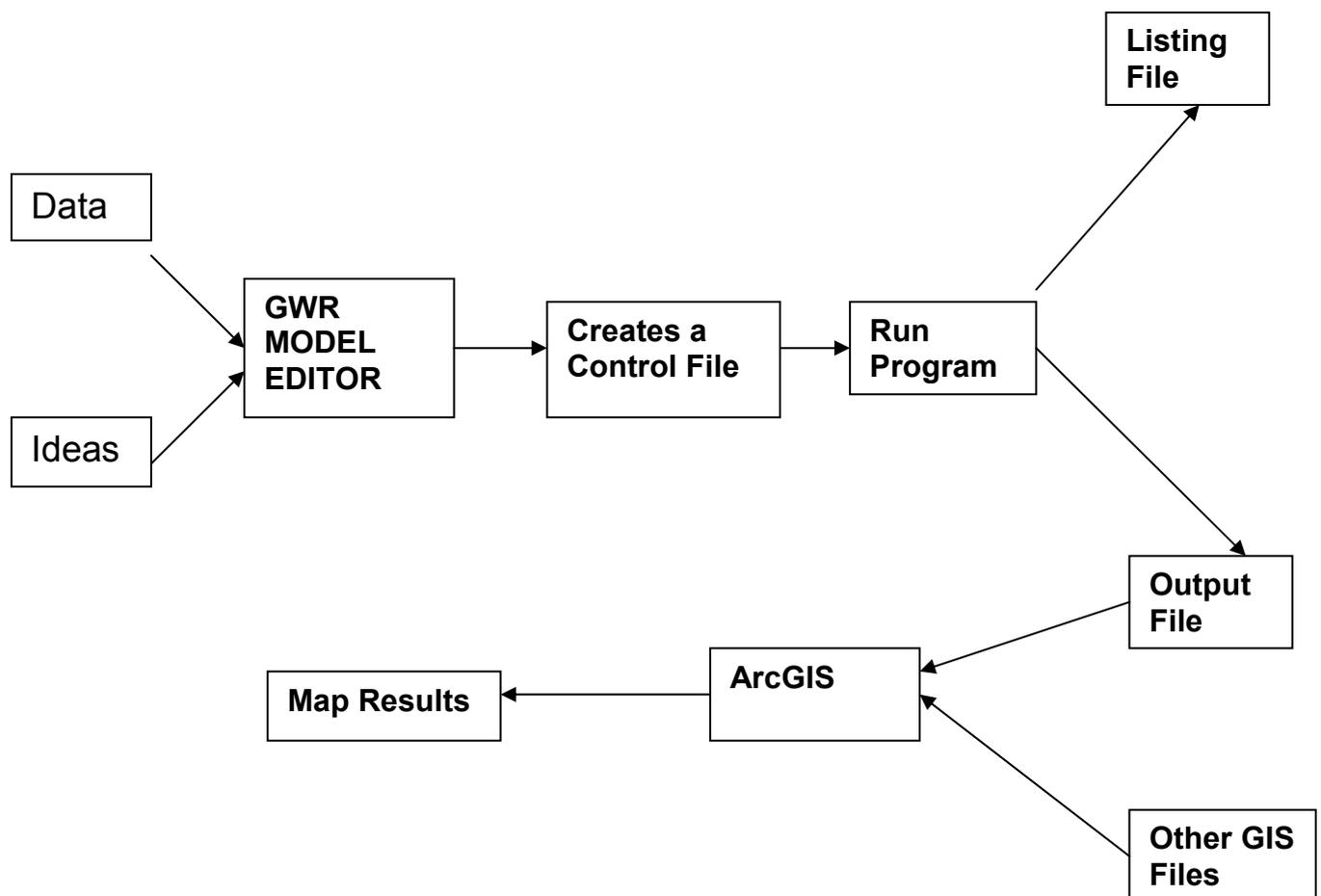
Contents

1	Introduction	1
2	A Primer on Running GWR	3
3	Poisson GWR	19
4	Binary Logit (Logistic) GWR	21
5	Geographically Weighted Descriptive Statistics	23

1. Introduction

1.1 The Basic Operation of GWR 3.0 and its Linkage to GIS

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.

To run the GWR software you simply need to click on the GWR icon which you will find in whichever folder you have loaded your software. This icon can be moved to the desktop for quick access.



Gwr30.exe

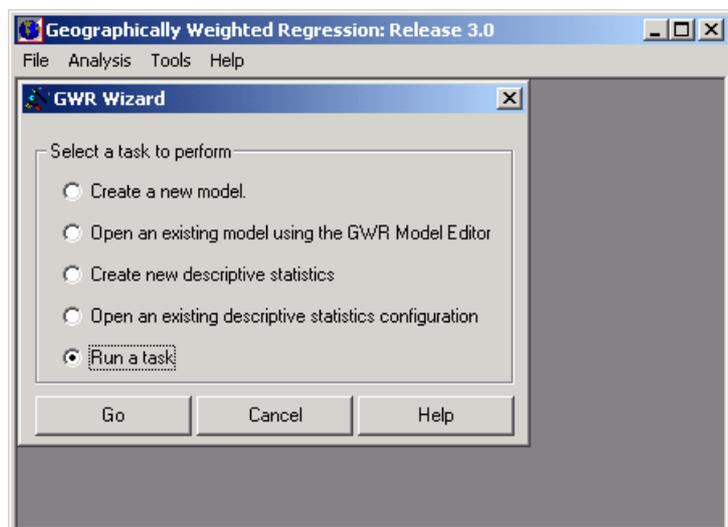
2. A Primer on Running GWR3

2.1 Introduction

This section shows how to set up and run a GWR model using the Visual Basic GWR Model Editor. There are three different varieties of regression model that can be run – Gaussian, Poisson and Binary Logit (or Logistic). The selection of a particular type of regression model is made by simply clicking the appropriate option in a box and whichever model form is chosen, the completion of the model editor follows in much the same way. 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 probably the most frequently encountered application of GWR. At the end of the manual, we discuss the Poisson and Binary Logit options briefly.

Once you have loaded the software using the GWR software loading program supplied on the CD, the GWR software will be located in a folder on your machine. The folder is usually named `GWR3`. There are two binaries in this folder. The name of the GWR Model Editor is `GWR3.exe`. A link can be created to the file `GWR3.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. Then 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. The first task is to 'Create a new model'.



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 help 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 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 are separated by commas
10. All attributes are numeric

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
13. The maximum number of observations is 100,000
14. The maximum number of variables is 35

As an example, here are the first 11 lines of the data file for the georgia educational attainment data supplied on the CD:

ID	Latitude	Longitude	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.

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

PARAM_1 ... PARAM_n	Values of the estimates of the parameters at each regression point. n is one more than the number of independent variables with PARAM_1 containing the values of the intercept term.
SVVAL_1 ... SVVAL_n	Values of the estimates of the standard errors of the parameters at each regression point. The numbering of

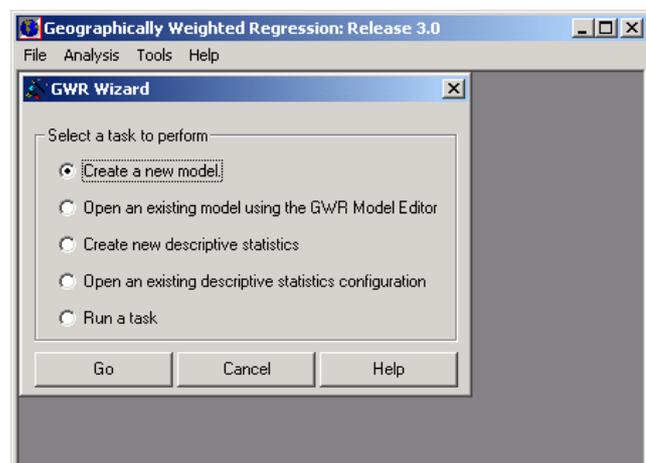
	these variables is as for the parameter estimate variables.
TVAL_1 ... TVAL_n	Pseudo-t values
OBS	Observed y variable value
PRED	Predicted y variable value
RESID	Unstandardised residual
HAT	Leverage value
STDRES	Standardised residual
COOKSD	Cook's Distance
LOCRSQ	Pseudo-R ² values

Three types of output format are available

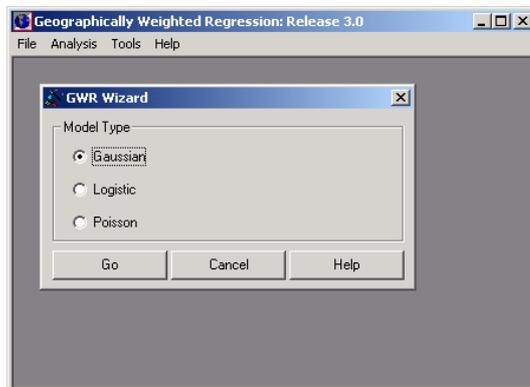
1. ArcINFO uncompressed export format. This may be imported into Arc/INFO to create a point coverage (where the coordinates of each point are those of the regression points). On a PC the coverage can be created using ArcToolBox. The filetype is .e00.
2. Comma-separated-variable format. This may be imported 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 10⁴. 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.

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



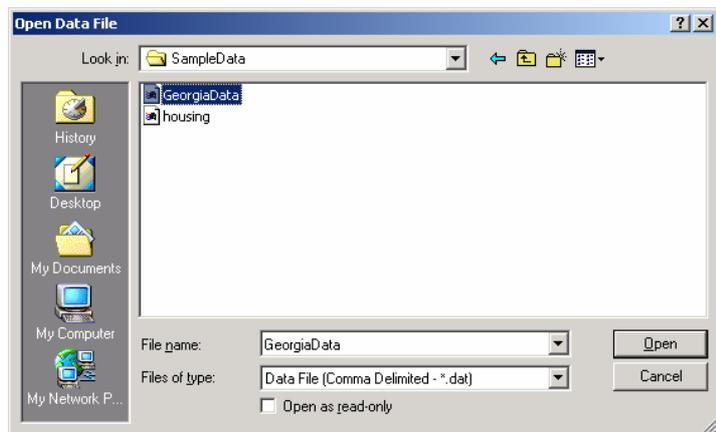
'Go' to proceed with the new model, 'Cancel' to close the Wizard, or 'Help' to obtain some assistance on what to do next.



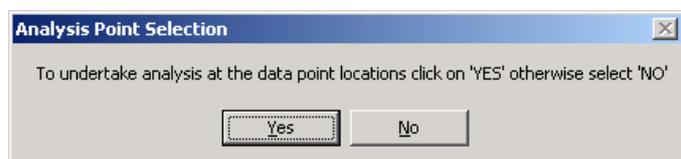
In this example, we have checked 'Create a new model' and clicked on 'Go'. We next need to determine what type of GWR model we wish to fit. In many cases this will be a Gaussian model. Select Gaussian and then 'Go'

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.

The form shown on the right will now appear - it is a standard Windows type 'File Open' form. In this case, there are two data files in the data folder. Click on the relevant data file name to highlight it and click on 'Open' to proceed.

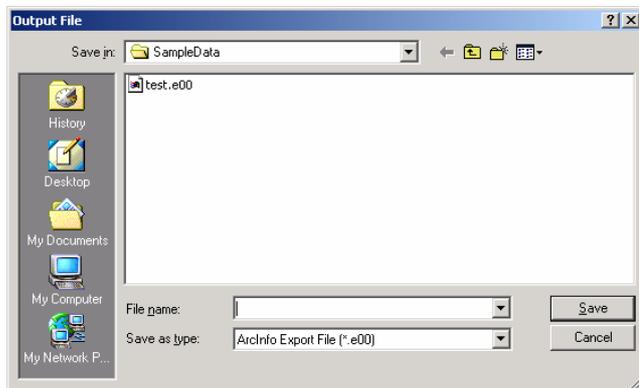


For the Gaussian option only, GWR estimates may be produced at locations other than those at which data are



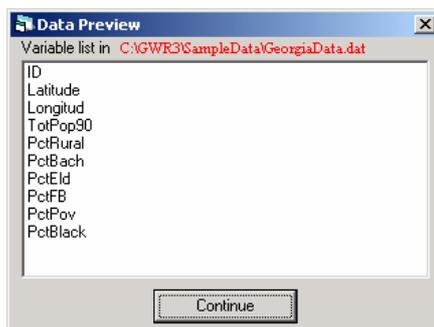
sampled. Locations where observations are recorded are referred to as data points (or as sample points) and the locations at which the estimates are produced as regression points. In most instances, the regression points and the data points will be the same. However, there is an option in GWR3.0 to produce 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. This file must contain three values per line: ID, XCoord YCoord and must **not** contain any variables names on the first line. Note that using this second option means the automatic bandwidth selection and a range of diagnostic statistics will not be available.

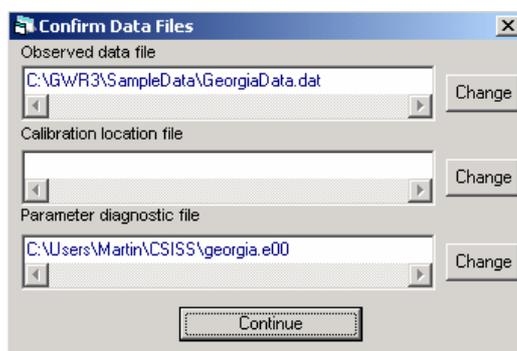


Next, the name of the file into which the results will be written must be specified. This file can be in one of several formats (comma-separated variable, ArcInfo uncompressed export, and MapInfo Interchange). The user also needs to specify the appropriate filetype - .e00 for an

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.

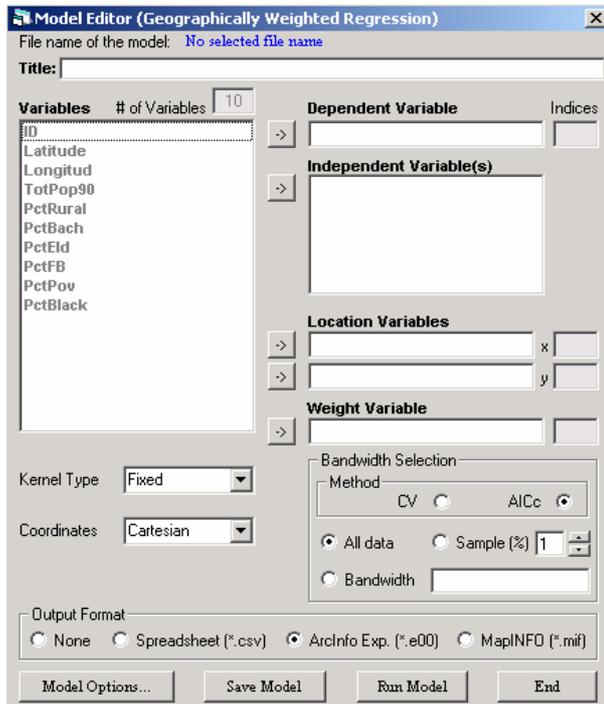


The Data Preview window allows you to check that you have loaded the correct file - it lists the variable names which it has found in the first line of the file and gives you the location of the file.



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.



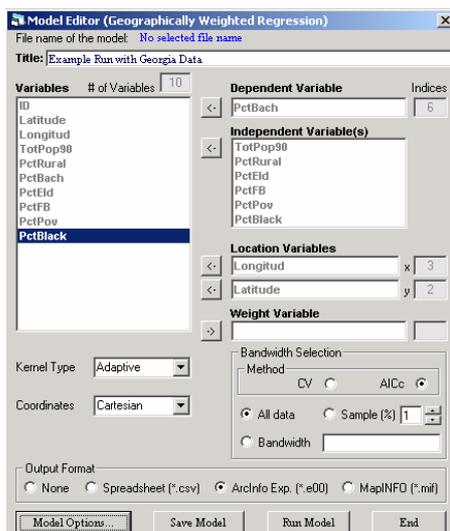
The Model Editor Window appears next and is shown on the left. It allows a GWR model to be created, saved and run. The `title` box allows the user to input a title which will then appear in the output listing. The list of `variables` is read automatically from the comma-separated list on the first line 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* (see below).

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. 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, 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`.

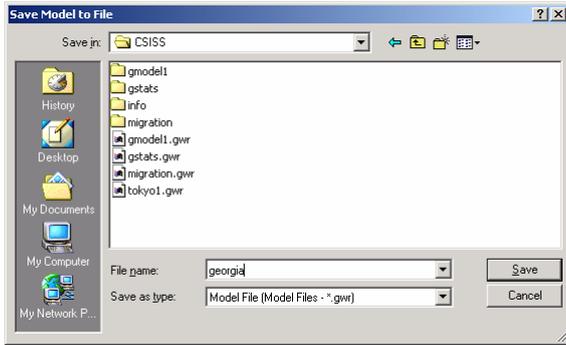
If your coordinates are in some projected coordinate system (UTM , for example) then the Coordinate Type should be Cartesian. If your measurements are in Lat/Lon, then select Spherical. If you have Lat/Lon coordinates, but your study area is in a relatively low latitude, then you can use Cartesian as the type. With Spherical coordinates, the distance computations in the geographical weighting use Great Circle distances.

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`, `Leung`, 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). Note that although the Leung test appears in the model editor, it is very cumbersome for large data sets and we have disabled it. Clicking on this option therefore generates an error message.

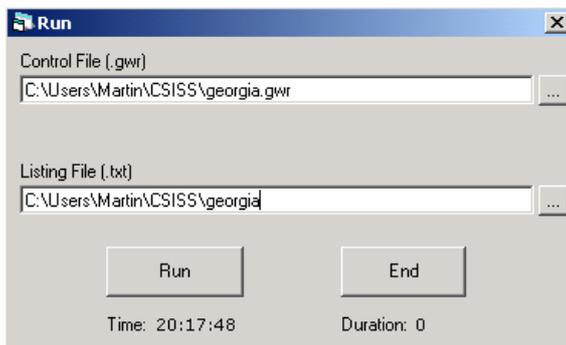


A completed example of the GWR Editor is shown on the left. The dependent variable is the proportion of the county population with education to degree level. Suppose we are interested to see how this is related to total population within each county, the percentage rural, the percentage elderly, the 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 `Longitude` (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.

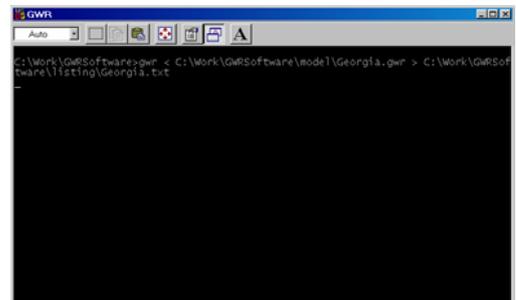


Before the model can be run, it must be saved. Clicking on **Save Model** will open the standard window shown on the left which depicts the contents of the model folder where the model control files are stored. Type the name of the file in the **Filename** box or click on an existing filename and then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

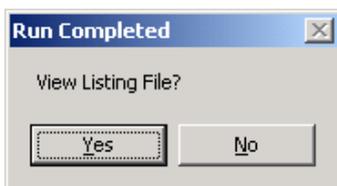


Once the model has been saved, it can be run. Simply click on the **Run** button in the Model Editor window and this brings up the form shown on the left. A name must be specified for the **Model Listing File (.txt)**. This file will be placed in the listing folder. To specify a filename click on the **...** 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 on the right.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 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.

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

2.6 Printed Outputs

```

*****
* Geographically Weighted Regression *
* Release 3.0.1 *
* Dated: 09/21/2009 *
* *
* Martin Charlton, Chris Brunsdon *
* Research Fellow *
* (c) University of Newcastle upon Tyne *
*****

** Program limits:
** Maximum number of variables..... 36
** Maximum number of observations... 12000
** Maximum number of fit locations... 12000

Example Run with Georgia Data
** Observed data file: C:\GWR3\SampleData\GeorgiaData.dan
** Prediction location file: Estimation at sample point locations
** Result output file: C:\Users\Martin\GIS\georgia.g00

** Variables in the data file...
ID Latitude Longitude TotPop90 PctRural PctEld PctFB PctPov PctBlack
PctPov PctBlack

** Dependent (y) variable.....PctEld
** Easting (x-coord) variable.....Longitude
** Northing (y-coord) variable.....Latitude
** No weight variable specified
** Independent variables in your model...
TotPop90 PctRural PctEld PctFB PctPov PctBlack
View another listing file Close
  
```

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 ($e^T e$), the number of parameters in the global model, the standard error of the estimate (σ), 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 $\alpha=0$.

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. 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 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. 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 ± 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 $2 \times \text{S.E.}$ 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.

```

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

Tests based on the Monte Carlo significance test
procedure due to Hope [1968, JR SB, 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

```

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.

3 Poisson GWR

To this point, we have assumed that the dependent data being analysed within a GWR framework are continuous and have a Gaussian (Normal) distribution. We now mention two other types of GWR that can be performed with GWR 3.0. Poisson Regression is used when the response variable refers to counts of some phenomenon and the covariates are either continuous or binary measurements. Typical examples of count data might include numbers of people with a particular disease, numbers of crimes in an area or numbers of derelict houses in a neighbourhood. In GWR such data are modelled using a Poisson regression model. Each observation has an integer-valued response variable, a number of explanatory variables, and two locational variables (X and Y coordinates). If you wish to model counts which relate to some underlying areal population, for example, the number of children aged 0–14 with leukaemia, you should use what is referred to in the Poisson regression literature as ‘an offset variable’. In this case, the offset variable would be the number of 0–14 children in each spatial unit. There will be more detail about this below. You should not attempt to model data which is continuous, such as the educational attainment data we have used in a previous workshop, in a Poisson regression framework – use Gaussian regression for this. The Poisson GWR model is set up using the GWR Model Editor as in previous examples.

A Poisson model takes the form

$$y_i^* = O_i e^{(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})}$$

in which, on taking logs of both sides, becomes

$$\log y_i^* = \log O_i + \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}$$

The O_i is known as the offset – in the above equation $\log O_i + \beta_0$ form the constant term.

In Poisson GWR, as with Gaussian GWR, the resulting parameter estimates β_i^* are specific to each location i . Unlike Gaussian GWR, however, the model is fitted using a technique known as iteratively reweighted least squares. This has the following implications. First, the fitting technique is iterative and so a typical GW Poisson regression will take approximately 5 times as long to run as an

equivalent Gaussian regression. Second, to compute the standard errors, the observed counts are required, so parameter estimates may only be obtained at the data points.

The output file contains the parameter estimates from the Poisson model and their standard errors as well as the exponentials of the parameter estimates and the standard errors of these exponentiated values. Positive parameter values when exponentiated are greater than unity, a parameter of zero when exponentiated yields a value of unity, and negative parameter values when exponentiated are less than unity. In all cases, the exponentiated values are positive.

There is one issue to note with Poisson GWR when the dependent variable consists of large numbers of zeros (such as a count of rare events for example). In this case, in the search for an optimal bandwidth low values of a bandwidth can result in the situation where all the values of the dependent variable are zero for a particular regression point. The model will then generate an error and will stop. The solution is to manually insert a series of bandwidths and examine the results at different bandwidths, selecting the model that performs best.

4 Binary Logit (Logistic) GWR

Binary Logit Regression (also known as Logistic Regression) is used when the response variable is binary and the covariates are either continuous or binary measurements. Typical examples of binary data might include yes/no, alive/dead above/below. In GWR such data are modelled using a logistic regression model. You will find the term binary or dichotomous used as names for the 0/1 data – note, however, that these data are not binomial, and the model is not binomial. In GWR each observation has a 1/0 valued response variable, a number of explanatory variables, and some some locational variables. The problem is set up using the GWR Model Editor as before.

Mathematically, the situation is that we need a model form that predicts y_i , where $i = 1 \dots n$, as a value in the interval 0–1 based on a set of explanatory variables, $1 \dots k$. Such a model is,

$$y_i^* = \frac{e^{(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})}}{1 + e^{(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})}},$$

where y_i^* is the predicted value of y_i ,

because, as $\exp(\beta_0 + \beta_1 x_{1i} \dots + \beta_k x_{ki}) \rightarrow 0$, $y_i^* \rightarrow 0$

and as $\exp(\beta_0 + \beta_1 x_{1i} \dots + \beta_k x_{ki}) \rightarrow \infty$, $y_i^* \rightarrow 1$

Note that

$$1 - y_i^* = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})}}$$

so that

$$\frac{y_i^*}{1 - y_i^*} = e^{(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})}$$

and

$$\log \left[\frac{y_i^*}{(1 - y_i^*)} \right] = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}$$

The term on the left hand side of the equation is known as the logit transformation and this produces a linear function in terms of the right hand side of the equation.

In logistic GWR, as with Gaussian GWR, the local parameter estimates β_i^* are specific to each location i . The model calibration, however, is more complicated than in OLS regression and the model is fitted using a technique known as iteratively reweighted least squares. This has two implications for our analysis. First, the fitting technique is iterative, and a typical problem will take approximately 5 times as long to run as a Gaussian GWR. Second, to compute the standard errors, the observed 0/1 values are required, so parameter estimates may only be obtained at the data points – that is, we no longer have the option of producing local parameter estimates at points other than the data points.

The output file from logistic GWR contains the parameter estimates and their standard errors from and also the exponentials of these values which might be useful for mapping. Positive parameter values when exponentiated are greater than unity, a parameter of zero yields an exponentiated value of unity, and negative parameter values when exponentiated are less than unity. In all cases, the exponentiated values are positive.

The issue identified above with the Poisson model in situations where the dependent variable consists of large numbers of zeros applies more frequently to Binary Logit GWR. Here, the dependent variable consists of either zeros or ones and the situation where locally all of the dependent variables have the same value is more likely to occur. Again, the model will then generate an error and will stop. The solution is to manually insert a series of bandwidths and examine the results at different bandwidths, selecting the model that performs best.

5 Geographically Weighted Descriptive Statistics

The geographically weighted mean is the starting point for thinking about geographically weighted statistics. Let us consider the arithmetic mean – its formula is:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

This is simply the sum of the values making up a batch of numbers divided by the size of the batch. More generally, we can consider a weighted mean:

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

where the w_i s are the weights. Here we multiply each value by its weight, and divide by the sum of the weights. In the case that each observation has a weight of unity, then this formula and the one above are equivalent.

In many cases the weights are integers, but they may also be non-integer numbers. In this case, we can use weights generated from the same geographical weighting scheme that we have used for geographically weighted regression. Rather than being a whole-map statistic, a geographically weighted mean is available at a particular location, say, u . Thus the formula for a geographically weighted mean at location u is:

$$\bar{x}(u) = \frac{\sum_{i=1}^n w(u)_i x_i}{\sum_{i=1}^n w(u)_i}$$

$w(u)_i$ is the geographical weight of the i th observation relative to the location u . The weights may be generated using a fixed radius or an adaptive kernel.

By analogy the local geographically weighted variance is

$$\sigma^2(u) = \frac{\sum_{i=1}^n w(u)_i (x_i - \bar{x}(u))^2}{\sum_{i=1}^n w(u)_i}$$

and the locally weighted standard deviation is the square root of this. Notice that the mean here is the geographical mean around point u and NOT the global mean of the data.

The GWR software currently supplied (GWR3.0) allows the user to compute geographically weighted means, variances, and standard deviations for a set of input data, and for either a fixed or an adaptive kernel. However, as there is no concept here of an optimal bandwidth, the bandwidth must be supplied by the user. If a fixed kernel is used, the bandwidth must be in the same units as the coordinates on the input data; if an adaptive kernel is used, the bandwidth is the number of objects to include in the local sample. If the bandwidth is very small, the degree of smoothing from the weighting scheme will be very small: the local means will approach the original data values and the variances will be very small. A zero bandwidth may cause premature and possibly inelegant termination of the program. The larger the bandwidth, the greater will be the degree of smoothing in the resulting geographically weighted statistic. With a fixed kernel the bandwidth can be as large as you wish, although anything greater than the study area width will result in an almost similar set of means. With the adaptive kernel, the bandwidth should not be greater than the number of observations in your dataset.