GWR4 User Manual

GWR4

Windows Application for Geographically Weighted Regression Modelling

Tomoki Nakaya

Updated 7 May 2012
3 June 2009

GWR 4 Development Team

Tomoki Nakaya (Department of Geography, Ritsumeikan University)
Martin Charlton, Paul Lewis (NCG, National University of Ireland)
Stewart Fotheringham (School of Geography & Geosciences, University of St. Andrews)
Chris Brunsdon (Department of Geography, University of Liverpool)

Contents

1. Introduction

What is GWR4?

GWR4 is a new release of a Microsoft Windows-based application software for calibrating geographically weighted regression (GWR) models, which can be used to explore geographically varying relationships between dependent/response variables and independent/explanatory variables. A GWR model can be considered a type of regression model with geographically varying parameters. A conventional GWR is described by the equation

$$y_i = \sum_k \beta_k(u_i, v_i)x_{k,i} + \varepsilon_i \ ,$$

where y_i, $x_{k,i}$, and ε_i are, respectively, dependent variable, *k*th independent variable, and the Gaussian error at location *i*; (u_i, v_i) is the x-y coordinate of the *i*th location; and coefficients $\beta_k(u_i, v_i)$ are varying conditionals on the location. Such modelling is likely to attain higher performance than traditional regression models, and reading the coefficients can lead to a new interpretation of the phenomena under study. An important extension of GWR is its semiparametric formation by mixing globally fixed and geographically varying coefficients.

$$y_i = \underbrace{\sum_k \beta_k(u_i, v_i)x_{k,i}}_{\text{Local terms}} + \underbrace{\sum_l \gamma_l z_{l,i}}_{\text{Global terms}} + \varepsilon_i \quad .$$

where $z_{l,i}$ is the *l*th independent variable with a fixed coefficient γ_l. Such a model may reduce the model complexities and enhance its predictable performance. Using the framework of geographically weighted generalised linear modelling (GWGLM), logistic and Poisson regression models with geographically varying coefficients are also popular for binary or count data modelling. GWR4 enables the fitting of such GWR and GWGLM models with their semiparametric formations, associated statistical tests, and model selections by user-defined data and model settings.

Main features

(1) Semiparametric GWR

As noted above, a most remarkable feature of this release is the function to fit semiparametric GWR models, which allow you to mix globally fixed terms and locally varying terms of explanatory variables simultaneously. The function can be applied to popular types of generalized linear modelling including Gaussian, Poisson, and logistic regressions. Using the semiparametric modelling scheme, a new statistical test of geographical variability on geographically varying coefficients is enabled. It is also possible to use variable selection routines by which variables are automatically selected as either fixed or varying terms by recursive model comparisons.

(2) Interface

A tabbed interface has been introduced to enable modelling sessions to intuitively proceed in a step-by-step manner. Datasets and geographically listwise results can be viewed in separate spreadsheet-like windows. Several popular file types can be used as input data files (space, comma, tab separated text, and dbase IV formats). In addition, Areal key field can be integrated into the output of GWR modelling, enabling you to join your output CSV file to a GIS attribute table via the key field for mapping the result in a GIS environment. GWR4 can be also used by a batch mode without the Windows interface.

(3) Requirements

GWR4 runs on Windows Vista or Windows 7 environments with the .NET Framework 4. The maximum size of data is dependent on your local machine environment. GWR4 dynamically allocates memory for large matrices (n by n, where n is the number of regression points) even for conventional GWR models. Thus using a PC having relatively large memory size (equal to or larger than 4GB) for running GWR4 is recommended. If the PC has multi-core processors, GWR4 automatically uses multi-threading routines to speed-up the computation.

Notes for use of GWR4

(i) GWR4 is copyrighted by the GWR4 development team.

(ii) GWR4 can be freely distributed and used for academic and non-profit purposes. The developers of GWR4 are not responsible for any difficulties that users of the software may encounter.

(iii) When any results using GWR4 are published, the author(s) should clearly state that GWR4 was used.

2. Installation / Uninstallation

< How to install GWR4 >
Download the GWR4 installer, GWR4_setup.msi, and then double-click the icon.

< .NET Framework>
GWR4 works only in MS Windows environments that have the .NET Framework 4 installed. If your PC does not have it installed, a message will pop up suggesting that you download the .Net Framework 4 Client Profile from a Microsoft website. After setting up the .NET Framework 4, you can then retry installation of GWR4.

< When the installer starts … >
Follow the instructions to select the GWR4 installation folder and users. If the installation is successful, a shortcut to the program will appear on your desktop and in the GWR4 program group. You may access the program by clicking this shortcut.

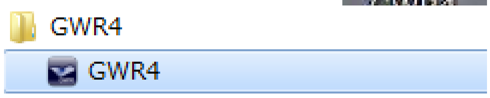
< To uninstall >
To uninstall GWR4 from your local environment, double click the setup file, GWR4_setup.msi, again and select “delete”. Alternatively, you can use the “Uninstall Programs” utility in the Control Panel group in Vista and Windows 7 environments.

3. Starting the program, Exiting the program, and Tab design

< Starting the program >

To start the program, double click the GWR4 shortcut icon on the desktop,

or select it from the GWR4 program group.



< Tab design >

Ensure that there are five tabs labelled Step 1 to Step 5. Click a tab label to move to the corresponding tab page. The first tab page when the program starts is "Step 1: Data>".

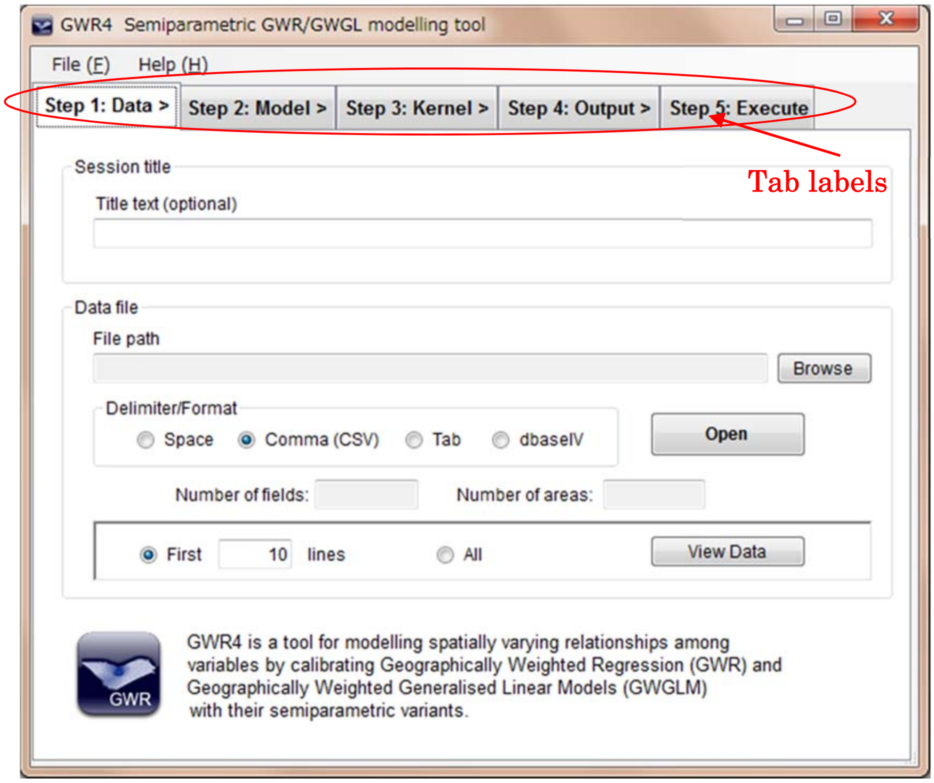
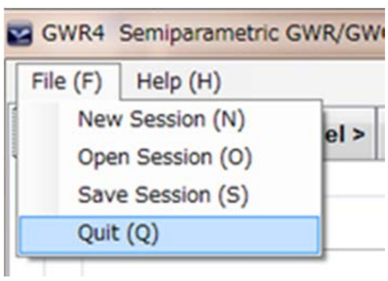
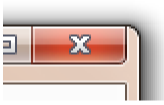


Figure 3.1: GWR4 startup screen

< Exiting the program >

To exit the program, select "Quit(Q)" on the File menu (alternatively, you can press the Alt and "F" keys simultaneously and then press the "Q" key), or you can click the close button in the top-right corner of the window.





Five steps in GWR calibration

We regard a session as the overall process by which settings are used to calibrate a GWR model. To build and proceed through your own session, you can generally follow the following five steps (Figure 4.1). Each step is separated into a tabbed page of the software. As described before, to move to a different page, click the corresponding tab label in the upper part of the window.

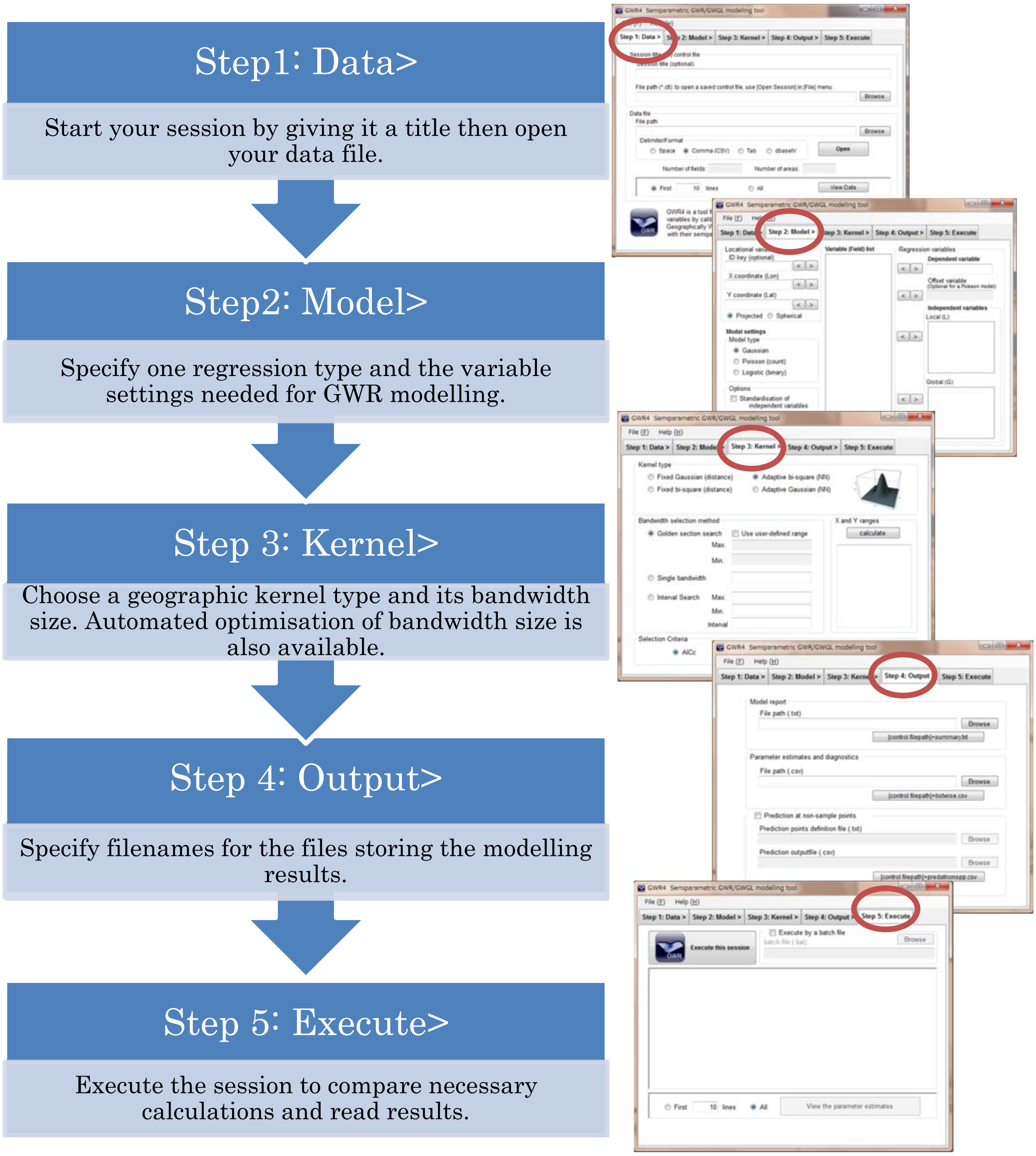


Figure 3.2: Steps in a GWR modelling session with respect to the five tabs

4. Step 1: The Data Tab

Data preparation

< What fields do I have to prepare in my dataset? >
To calibrate a GWR model, you must prepare a tabular dataset that contains fields of dependent and independent variables, and x-y coordinates. Every variable should consist of numeric values, except for Areal key, as an identifier of observation. Areal key is treated as a string field in GWR4.

< Coordinates >
Both projected and latitude/longitude (lat/lon) decimal degrees coordinates can be used as x-y coordinates in GWR4. However, projected coordinates are superior to decimal degrees in terms of computing time.

< Possible data formats >
GWR4 can open data files in text format delimited by space, comma, or tab; as well as files in dbase IV format (*.dbf). The most popular type is CSV (a text format that uses the comma delimiter).

< dbase IV files >
Since the shapefile spatial data format—a common GIS file format from ESRI Inc.—uses dbase IV as an attribute database linking with geographic objects, the function to read dbase file provides an easy way to use such a GIS data file in GWR4. However, there is one caveat for this function:

In the case of dbase IV, the length of the filename must not exceed 8 characters, due to the restriction on database connection used in this software. Some examples of readable and unreadable filenames are shown below:

Readable filenames: tokyom.dbf, tokyodat.dbf
Unreadable filenames: dublindata.dbf, irelandmortality.dbf

< Field names >
In the case of text formats, the first line (row) of the data file must be the list of field names. (A dbase IV file automatically has the field name lists.)

< Samples of text-based data files >

Two samples of data files using space and comma delimiters are shown below.

Table 4.1: Sample of text data file for GWR4: Tokyo mortality data
(Space delimited text file)

IDnum0	X_CENTROID	Y_CENTROID	db2564	eb2564	OCC_TEC	OWNH	POP65	UNEMP
0	378906.83	17310.41	189	194.572	0.126	0.606	0.104	2.865
1	334095.21	25283.20	95	97.526	0.107	0.671	0.111	3.401
2	378200.19	-877.05	70	83.235	0.106	0.733	0.110	1.724
3	357191.03	29064.39	48	52.392	0.075	0.767	0.140	1.829
4	358056.34	10824.73	65	67.664	0.075	0.812	0.146	1.961
5	366747.61	-3073.12	107	120.745	0.140	0.623	0.078	2.636
6	351099.27	11800.35	65	67.196	0.066	0.824	0.121	1.603
7	377929.98	4635.10	76	87.746	0.130	0.778	0.083	2.438
8	367529.91	20192.51	192	190.255	0.227	0.449	0.101	1.783
9	389231.47	3489.35	27	23.939	0.075	0.821	0.146	2.081
10	389427.64	9290.10	28	23.832	0.088	0.610	0.119	2.589

Note: only the first ten records are shown.

Table 4.2: Sample of text data file for GWR4: Georgia data
(Comma delimited text file: CSV)

```
AreaKey,Latitude,Longitud,TotPop90,PctRural,PctBach,PctEld,PctFB,PctPov,PctBlack,ID,X,Y↓
13001,31.75339,-82.28558,15744,75.60,8.20,11.43,0.64,19.90,20.76,133,941396.60,3521764.00↓
13003,31.29486,-82.87474,6213,100.00,6.40,11.77,1.58,26.00,26.86,158,895553.00,3471916.00↓
13005,31.55678,-82.45115,9566,61.70,6.60,11.11,0.27,24.10,15.42,146,930946.40,3502787.00↓
13007,31.33084,-84.45401,3615,100.00,9.40,13.17,0.11,24.80,51.67,155,745398.60,3474765.00↓
13009,33.07193,-83.25085,39530,42.70,13.30,8.64,1.43,17.50,42.39,79,849431.30,3665553.00↓
13011,34.35270,-83.50054,10308,100.00,6.40,11.37,0.34,15.10,3.49,23,819317.30,3807616.00↓
13013,33.99347,-83.71181,29721,64.60,9.20,10.63,0.92,14.70,11.44,33,803747.10,3769623.00↓
13015,34.23840,-84.83918,55911,75.20,9.00,9.66,0.82,10.70,9.21,24,699011.50,3793408.00↓
13017,31.75940,-83.21976,16245,47.00,7.60,12.81,0.33,22.00,31.33,138,863020.80,3520432.00↓
13019,31.27424,-83.23179,14153,66.20,7.50,11.98,1.19,19.30,11.62,153,859915.80,3466377.00↓
13021,32.80451,-83.69915,149967,16.10,17.00,12.23,1.06,19.20,41.68,85,809736.90,3636468.00↓
13023,32.43552,-83.33121,10430,57.90,10.30,12.60,0.64,18.30,22.36,100,844270.10,3595691.00↓
13025,31.19702,-81.98323,11077,100.00,5.80,9.02,0.33,18.20,4.58,159,979288.90,3463849.00↓
13027,30.84653,-83.57726,15398,65.60,9.10,13.68,1.76,25.90,41.47,169,827822.00,3421638.00↓
13029,32.02037,-81.43763,15438,80.60,11.80,7.22,0.45,13.20,14.85,118,1023145.00,3554982.00↓
13031,32.39071,-81.74391,43125,63.20,19.90,9.56,1.16,27.50,25.95,97,994903.40,3600493.00↓
13033,33.05837,-81.99939,20579,72.30,9.60,10.60,0.43,30.30,52.19,71,971593.80,3671394.00↓
13035,33.28834,-83.95713,15326,73.40,7.20,10.41,0.72,15.60,35.48,65,782448.20,3684504.00↓
13037,31.52793,-84.61891,5013,100.00,10.10,15.94,0.10,31.80,58.89,149,724741.20,3492653.00↓
13039,30.91895,-81.63783,30167,47.10,13.50,4.78,2.14,11.50,20.19,165,1008480.00,3437933.00↓
```

Note: only the first twenty records are shown.

< Missing values >

GWR4 does not have functions for handling missing codes. Records having blank items in the dependent variable field are skipped for the model fitting. In the cases that there are blank items for independent variables, the model fitting fails. Avoid including missing values for your data set by excluding such data or estimated missing values before GWR modelling.

Operations in the data tab page

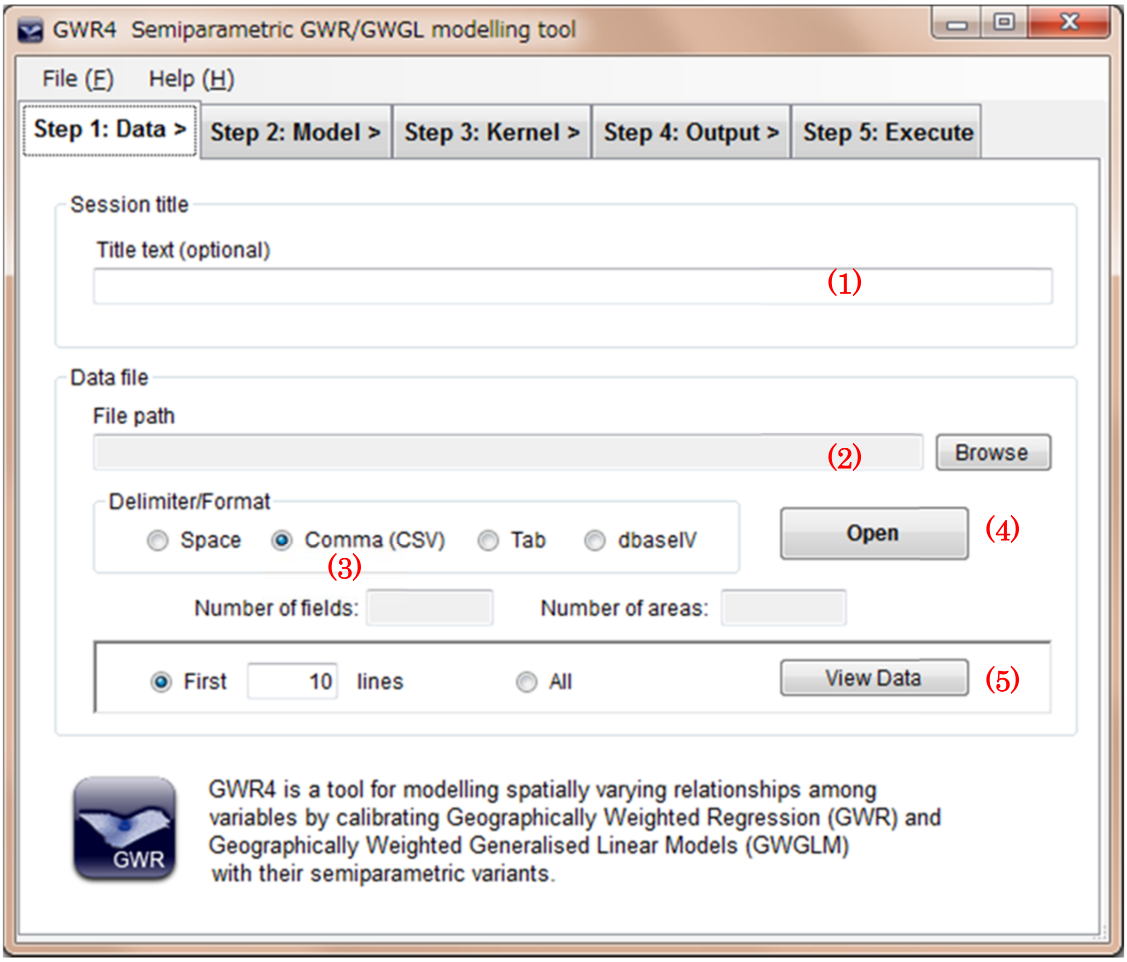


Figure 4.1: Sample screenshot for "Step 1: Data"

The following is the recommended procedure to follow on the "Data" tab page:

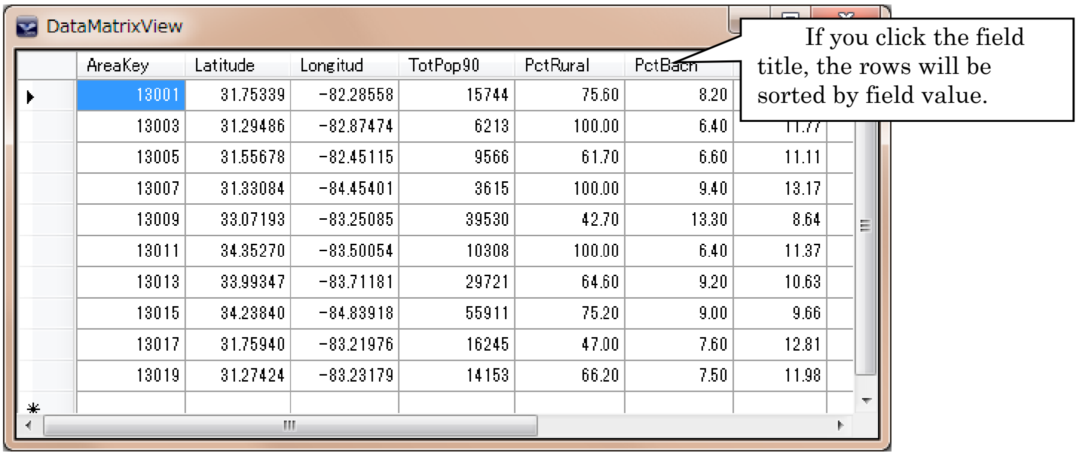
(1) Create a session title that reflects your modelling intention and dataset.

(2) Input the filename of your dataset in this textbox by clicking the rightmost "Browse" button to open a file dialog box.

(3) Select the appropriate format or delimiter that separates each value in a row of a text data file.

(4) Click the "Open" button. Both "Number of fields" and "Number of areas" will automatically appear. (Note: "Number of areas" refers to the number of data records/observations in the data file.)

(5) You can view your dataset in a spreadsheet-like window by clicking the "View data" button. The number of rows (lines) in the gridded view can be controlled using the options to the left of the button (default is first 10 lines).



AreaKey	Latitude	Longitud	TotPop90	PctRural	PctBacn	
13001	31.75339	-82.28558	15744	75.60	8.20	
13003	31.29486	-82.87474	6213	100.00	6.40	11.77
13005	31.55678	-82.45115	9566	61.70	6.60	11.11
13007	31.33084	-84.45401	3615	100.00	9.40	13.17
13009	33.07193	-83.25085	39530	42.70	13.30	8.64
13011	34.35270	-83.50054	10308	100.00	6.40	11.37
13013	33.99347	-83.71181	29721	64.60	9.20	10.63
13015	34.23840	-84.83918	55911	75.20	9.00	9.66
13017	31.75940	-83.21976	16245	47.00	7.60	12.81
13019	31.27424	-83.23179	14153	66.20	7.50	11.98

Figure 4.2: Data view window

5. Step 2: The Model Tab

If the data file you specified on the "Data" tab page is successfully opened, field names will appear in the "Variable (Field) list" list box in the middle of the "Model" tab page. If there is no field name or the listings in the list box are insufficient, return to the "Data" tab page, then press the "Open" button after selecting a proper data file and format/delimiter option.

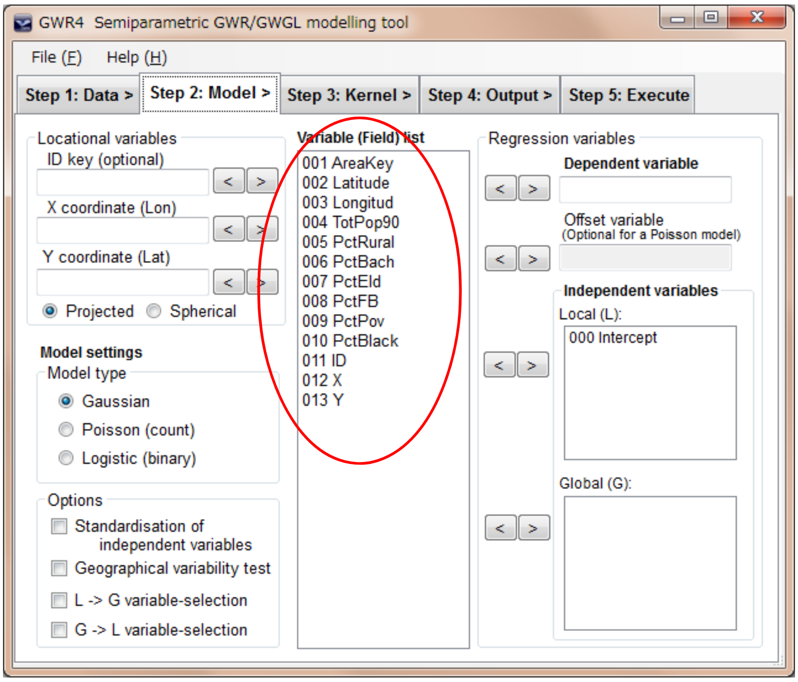


Figure 5.1: Screenshot of the model tab page just after opening a data file

Basic operations: using an example of Gaussian GWR

< Gaussian GWR model >
It is essential to ensure that appropriate options are selected for your model structure and variables in this "Model" tab page. A conventional (Gaussian) GWR model is described as

$$y_i = \sum_k \beta_k(u_i, v_i) x_{k,i} + \varepsilon_i \quad ,$$

where y_i, $x_{k,i}$, and ε_i are, respectively, dependent variable, *k*th independent variable, and the Gaussian error at the location *i*; (u_i, v_i) is the x-y coordinate of the *i*th location; and coefficients $\beta_k(u_i, v_i)$ are varying conditional on the location. Usually, the first variable is constant by setting $x_{0,i} = 1$, after which $\beta_0(u_i, v_i)$ becomes a geographically varying "intercept" term.

< Fitting a Gaussian GWR model >
To fit a Gaussian GWR model, at least three kinds of operations should be done:

(1) < variable settings >: By using the "<" and the ">" buttons, select and move the appropriate field of dependent variable to the "dependent variable" text box. In a similar fashion, move the fields of independent variables to "Local" from the variable (field) list.

The item "000 Intercept" is included in the "Local" list box for geographically varying terms as a default condition. By moving the "intercept" list item to the variable (field) list, the intercept can be fixed as zero in the model (no intercept model). However, in most cases, it is recommended that "intercept" not be removed from your model.

(2) < coordinate specification >: Similarly, select and move appropriate fields from the variable (fields) list to "x coordinate variable" and "y coordinate variable"; then select the appropriate coordinate option (either "projected" or "spherical").

"Projected" is typically used for coordinates projected onto an orthogonal two-dimensional space, such as UTM coordinates. "Spherical" should be selected if x-y coordinates are stored in a Lat-Lon format (decimal degrees). The spherical distance is more accurate if the geographic extent of the study area is wide (for example, continental scale). However, spherical distance calculation is computationally more expensive than Euclidian distance calculation applied to "Projected" coordinates. For most city and regional scale applications, projected coordinates are suitable in terms of the balance between accuracy and computational load.

(3) < model type >
Select the "Gaussian" option in the model type panel.

< Example of Gaussian GWR >
The following screenshot shows an example of a Gaussian GWR model using the Georgia sample data with the following specifications:

$$\begin{aligned}\mathrm{PctBatch}_i &= \beta_0(X_i,Y_i)+\beta_1(X_i,Y_i)\mathrm{PctRural}_i \\ &+\beta_2(X_i,Y_i)\mathrm{PctPov}_i+\beta_3(X_i,Y_i)\mathrm{PctBlack}_i+\varepsilon_i\end{aligned}$$

where X_i *and* Y_i are projected x-y coordinates in this example.

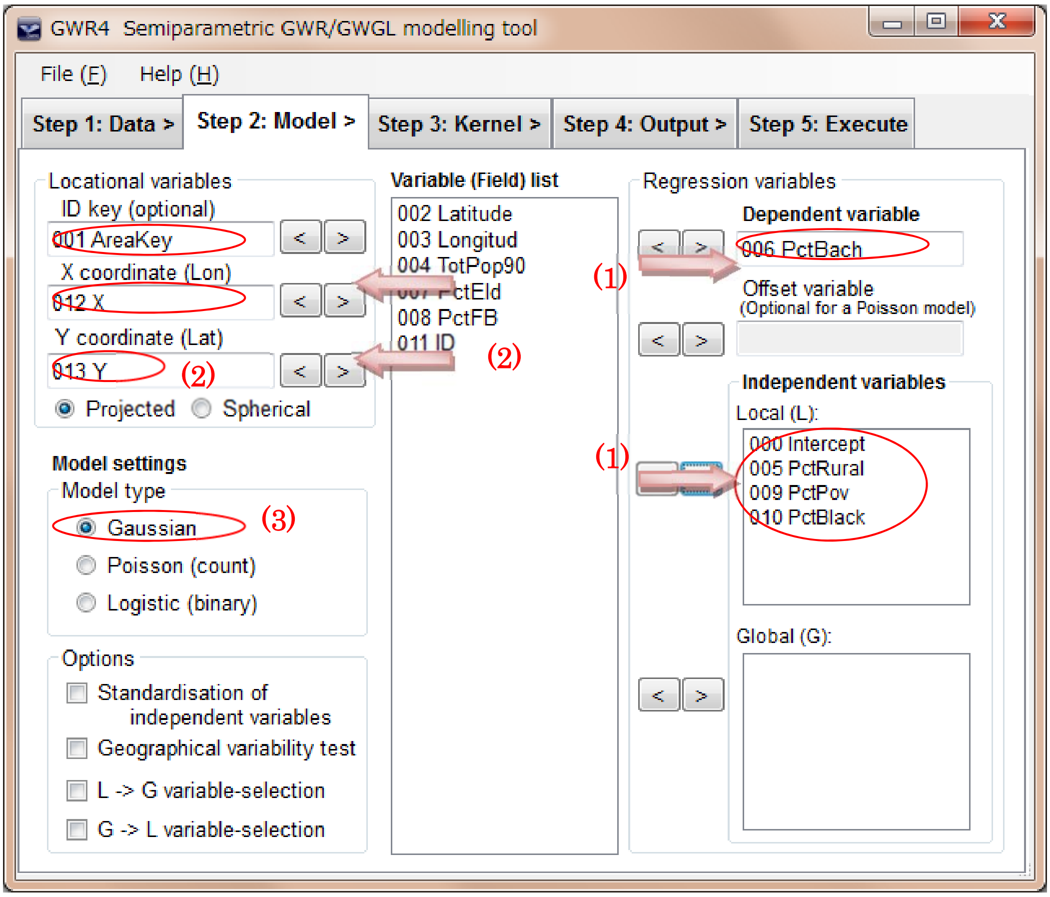


Figure 5.2: Screenshot of the model tab page for a simple Gaussian GWR

< ID Key option >

If you have a locational ID field (such as place names or regional codes), you can include it for later use in your results. Using this ID key, you can join your resulting output files containing local estimates to other tables (such as GIS attribute tables).

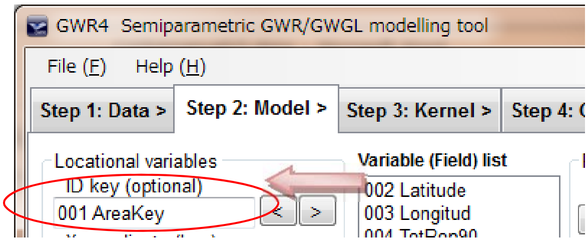


Figure 5.3: ID key option

Semiparametric GWR

< Semiparametric GWR >

A semiparametric Gaussian GWR model is described as

$$y_i = \underbrace{\sum_k \beta_k(u_i, v_i) x_{k,i}}_{\text{Local terms}} + \underbrace{\sum_l \gamma_l z_{l,i}}_{\text{Global terms}} + \varepsilon_i \quad .$$

where $z_{l,i}$ is the lth independent variable with a fixed coefficient γ_l. Thus, the model mixes geographically local and global terms. This kind of model has several aliases, such as "mixed model" and "partial linear model".

The intercept term is usually specified as a varying term since other varying coefficients often cause a variation in the intercept. However, it is also possible to assign the intercept as a fixed term in the interface.

< Variable settings for semiparametric models >

To fit a semiparametric model, most of the operations needed are the same as those used in the case of traditional GWR models. One additional thing is to specify the global term by moving independent variables for global terms to the "Global" box.

< Example of semiparametric Gaussian GWR >

The following equation is an example of a semiparametric Gaussian GWR model using the Georgia sample data with the following specifications:

$$\begin{aligned}\text{PctBatch}_i &= \beta_0(X_i,Y_i)+\beta_1(X_i,Y_i)\text{PctRural}_i \\ &+\beta_2(X_i,Y_i)\text{PctPov}_i+\beta_3(X_i,Y_i)\text{PctBlack}_i \\ &+\gamma_1\text{PctEld}_i+\gamma_2\text{PctFB}_i+\varepsilon_i\end{aligned}$$

It should be noted that the coefficients for PctEld and PctFB, γ_1 and γ_2 ,are both geographically invariable, like the coefficients of the conventional regression model.

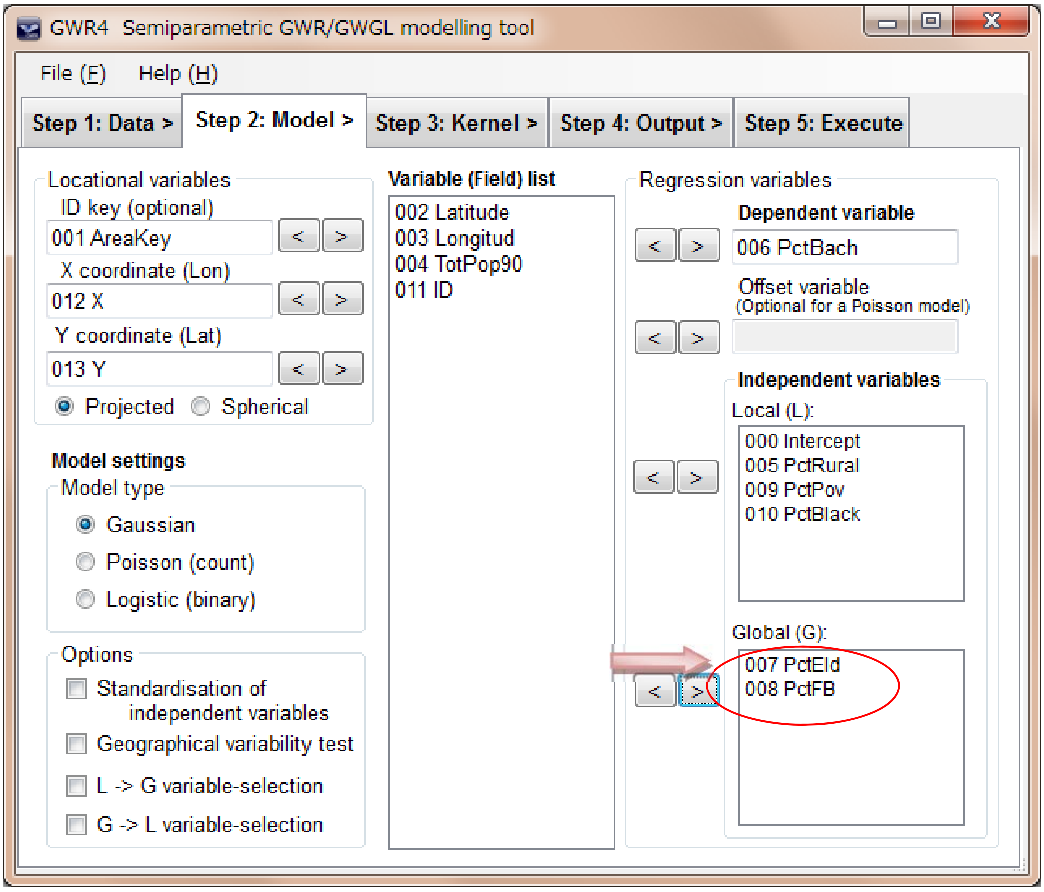


Figure 5.4: Specifying global terms

Extensions of GWR: GWGLM

< GWR for count and binary outcome >
A Gaussian error term is suitable for modelling numerical responses. However, in the case of modelling count or binary (dichotomous) responses, other model types of generalised linear modelling, particularly logistic and Poisson regression, are quite popular. As a natural extension of GWR, we can theoretically derive geographically weighted generalised linear models (GWGLM). In GWR4, geographically weighted Poisson regression (GWPR) and geographically weighted logistic regression (GWLR) can be fitted for modelling count and binary outcome, respectively, with geographically varying coefficients.

< Fitting a GWPR or GWLR >
To fit a GWPR or GWLR, simply switch the model type option to select “Poisson” or “Logistic”.

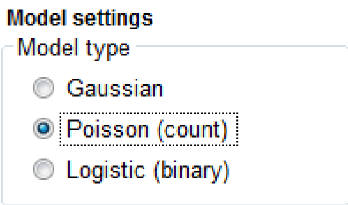


Figure 5.5: Model type settings

Semiparametric GWPR and GWLR using both global and local terms can be fitted by specifying terms with fixed coefficients, similar to the semiparametric Gaussian GWR case.

Geographically weighted Poisson regression (GWPR)

A GWPR model and its semiparametric variant are shown as

GWPR: $$y_i \sim \text{Poisson}\left[N_i \exp\left(\sum_k \beta_k(u_i,v_i)x_{k,i}\right)\right],$$

Semiparametric GWPR: $$y_i \sim \text{Poisson}\left[N_i \exp\left(\sum_k \beta_k(u_i,v_i)x_{k,i} + \sum_l \gamma_l z_{l,i}\right)\right]$$

The dependent variable should be an integer that is greater than or equal to zero. N_i is the offset variable at the ith location. This term is often the size of the population at risk or the expected size of the outcome in spatial epidemiology. In cases where the "offset variable" box is left blank, N_i becomes 1.0 for all locations.

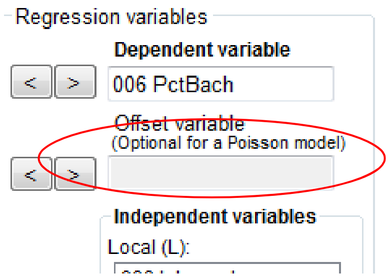


Figure 5.6: Offset variable box (optional)

Geographically weighted logistic regression (GWLR)

A GWLR model is shown as

$$y_i \sim \text{Bernoulli}[p_i]$$
$$\text{logit}(p_i) = \sum_k \beta_k(u_i, v_i)x_{k,i}$$

The dependent variable must be 0 or 1. p_i is the modelled probability that the dependent variable becomes one. Its semiparametric variant is described as

$$y_i \sim \text{Bernoulli}[p_i]$$
$$\text{logit}(p_i) = \sum_k \beta_k(u_i, v_i)x_{k,i} + \sum_l \gamma_l z_{l,i} \quad .$$

< Example of semiparametric GWPR >

The following equation is an example of the semiparametric GWPR model using Tokyo mortality sample data with the following specifications:

$$
\begin{aligned}
&\mathrm{db2564}_i \sim \mathrm{Poisson}\left[\mathrm{eb2564}_i \exp\left(\mu_i\right)\right] \\
&\quad \mu_i = \beta_0(X_i,Y_i) + \beta_1(X_i,Y_i)\mathrm{OCC\_TEC}_i + \beta_2(X_i,Y_i)\mathrm{OWNH}_i \\
&\qquad + \gamma_1 \mathrm{POP65}_i + \gamma_2 \mathrm{UNEMP}_i
\end{aligned}
.
$$

where db2564 and eb2564 are the observed and expected size of deaths in region *i*, respectively. Note that eb2564 is the offset variable. The intercept and the two coefficients of OCC_TEC and OWNH are geographically varying but the other two coefficients of POP65 and UNEMP are fixed in the entire study region. The following screenshot shows the settings that fit the aforementioned model.

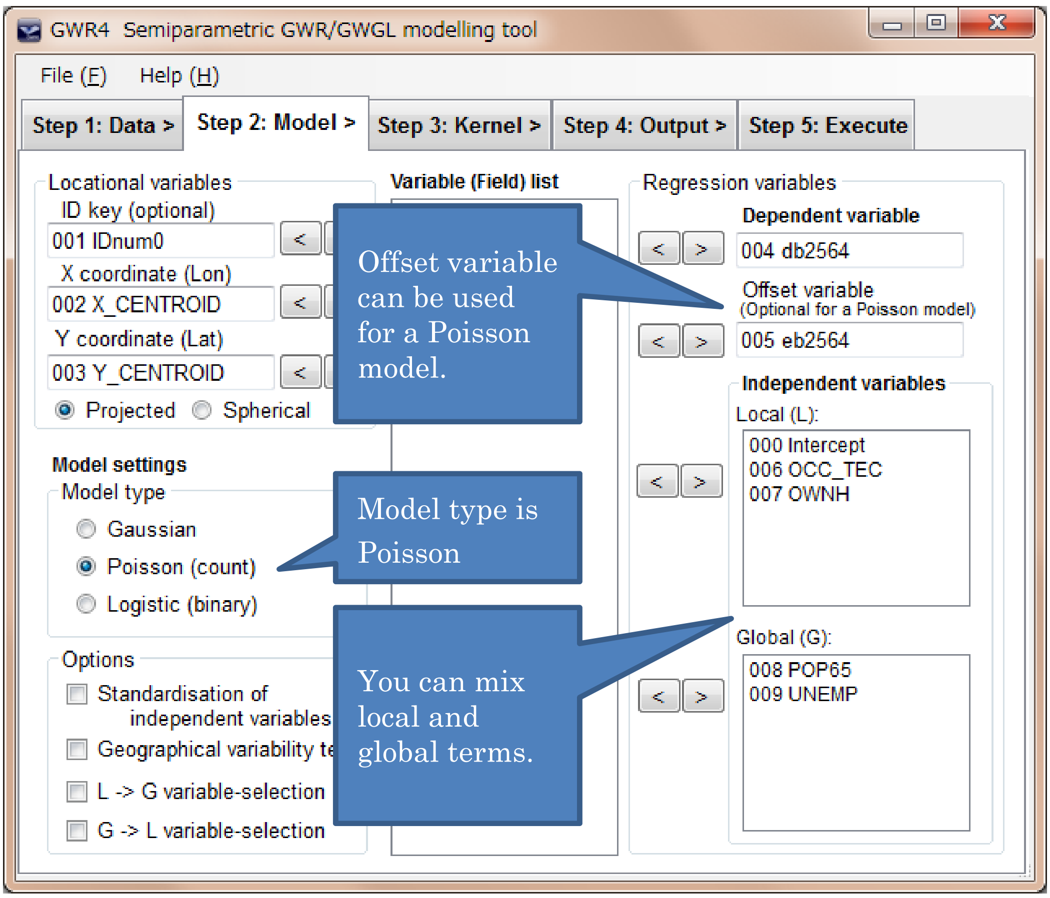


Figure 5.7: Screenshot of the model tab page for a semiparametric GWPR using Tokyo mortality sample data

Modelling Options

There are several advanced options for GWR model fitting. Check an option in the “Options” panel to enable the corresponding additional computation.

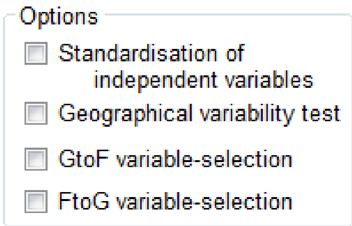


Figure 5.8: Modelling options

Standardisation

If this option is checked, all of the independent variables are standardised by z-transformation so that each variable has zero mean and one standard deviation. It is useful for interpreting estimated coefficients under the same metric. In some cases, standardisation makes iterative computation of model fitting faster.

Geographical variability test

< What is this test? >

Geographical variability for each varying coefficient is tested by model comparison. For testing the geographical variability of the *k*th varying coefficient, a model comparison is carried out between the fitted GWR and a model in which only the *k*th coefficient is fixed while other coefficients are kept as they are in the fitted GWR model. Let us term the two models original and switched GWR model, respectively. If the original GWR is better than the compared switched GWR model by a model comparison criterion such as AICc, we can judge that the *k*th coefficient is certainly varying over space. The test routine repeats this comparison for each geographically varying coefficient. The model comparison indicator used for this test is the same as the one used for bandwidth selection (specified in the “Kernel” tab page). The default indicator is AICc.

For example, imagine that a simple GWR model with one independent variable and intercept is fitted:

(Fitted model: original GWR model)

$$y_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)x_{1,i} + \varepsilon_i$$

To test the geographical variability of $\beta_0(u_i, v_i)$ and $\beta_1(u_i, v_i)$, we compare the following fixed constant and fixed slope models, respectively, with the fitted model.

(Fixed constant model: switched GWR model for testing local constant)

$$y_i = \beta_0 + \beta_1(u_i, v_i)x_{1,i} + \varepsilon_i$$

(Fixed slope model: : switched GWR model for testing local slope)

$$y_i = \beta_0(u_i, v_i) + \beta_1 x_{1,i} + \varepsilon_i$$

If the fixed slope model outperforms the fitted model, this suggests that the coefficient of $\beta_1(u_i, v_i)$ significantly varies over space. For a Gaussian model, statistics for F test is also shown. The test is applied to each term specified in the "Local" list box.

For computational simplicity, the compared models are fitted with the same bandwidth as the fitted model. Since the size of the best bandwidth of the simpler model with less number of effective parameters tends to be larger than that of the larger model, this geographical variability test is likely to be conservative (the originally fitted model is likely to outperform the counterpart model). However, the difference in bandwidth between the fitted and the compared model tends to be small so that this test is effectively useful for inspecting the geographical variability in many situations.

< How do I read the result? >

GWR4 will report on the difference between the original GWR model and the switched model in terms of model comparison indicators, if the test option is enabled. The result table consists of rows of local terms with a "Diff of Criterion" column, which shows the difference in model comparison indicator between the original GWR model and the switched GWR model. If the switched GWR model attains a statistically better fit, the value of the model comparison indicator is smaller than that of the original GWR model so that "Diff of Criterion" becomes a positive value, suggesting no spatial variability in the highlighted local term.

< Example of geographical variability test >

Imagine the example case of simple Gaussian GWR:

$$\mathrm{PctBatch}_i = \beta_0(X_i, Y_i) + \beta_1(X_i, Y_i)\mathrm{PctRural}_i$$
$$+\beta_2(X_i, Y_i)\mathrm{PctPov}_i + \beta_3(X_i, Y_i)\mathrm{PctBlack}_i + \varepsilon_i$$

According to the test results (Table 6.8), the coefficient of PctPov is not significantly varying so the term can be change to a global term. In the case of Table 6.8, the model comparison was conducted using AICc. It should be noted that if the difference of AIC or AICc is greater than two, we can consider that the model performance is significantly different. Thus, the intercept coefficients of PctRural and PctBlack are indicated as varying significantly in the table.

Table 5.1: Example of geographical variability tests of local coefficients

```
*************************************************************************
 Geographical variability tests of local coefficients
*************************************************************************
 Variable           F                DOF for F test  DIFF of Criterion
-------------------- ----------------- --------------- -----------------
Intercept                 5.367001    1.860  145.487       -6.069357
PctRural                  4.303917    2.191  145.487       -4.720029
PctPov                   1.976232    1.591  145.487        0.440477
PctBlack                  7.985049    1.662  145.487       -9.870218
-------------------- ----------------- --------------- -----------------
Note: positive value of diff-Criterion (AICc, AIC, BIC/MDL or CV) suggests no
spatial variability
```

LtoG / GtoL variable selection.

Like the geographical variability test, for each term in the “Local” list box the LtoG variable selection routine conducts a series of model comparison tests between the originally fitted model and a model in which a varying term has been changed to a fixed term with other terms remaining unchanged. Then, if the best of the compared model outperforms the original, the term used by the compared model is fixed as the result of the first step. Given the condition that the term is fixed, a series of model comparisons are successively repeated for each of the remaining varying terms. This variable selection from the geographically varying terms to the fix ones is repeated until there is no candidate of such a term being changed or no improvement is gained by changing any term.

Unlike the geographical variability test, bandwidth selection is applied for each compared model as specified in the “Kernel” tab page.

If there are varying terms that you do not want changed to fixed terms, you may identify them by clicking the terms in the “Local” list box. In other words, highlighted terms in the list box are assumed not to be candidates for switching from varying to fixed term.

GtoL variable selection is the reverse of LtoG variable selection. In this case, each term in the “Global” list box is a candidate for switching from fixed to varying term. The same logic of recursive variable selection of term switching from fixed to varying term is conducted.

Warning: these model selection routines may take quite a long time to compute.

6. Step 3: The Kernel Tab

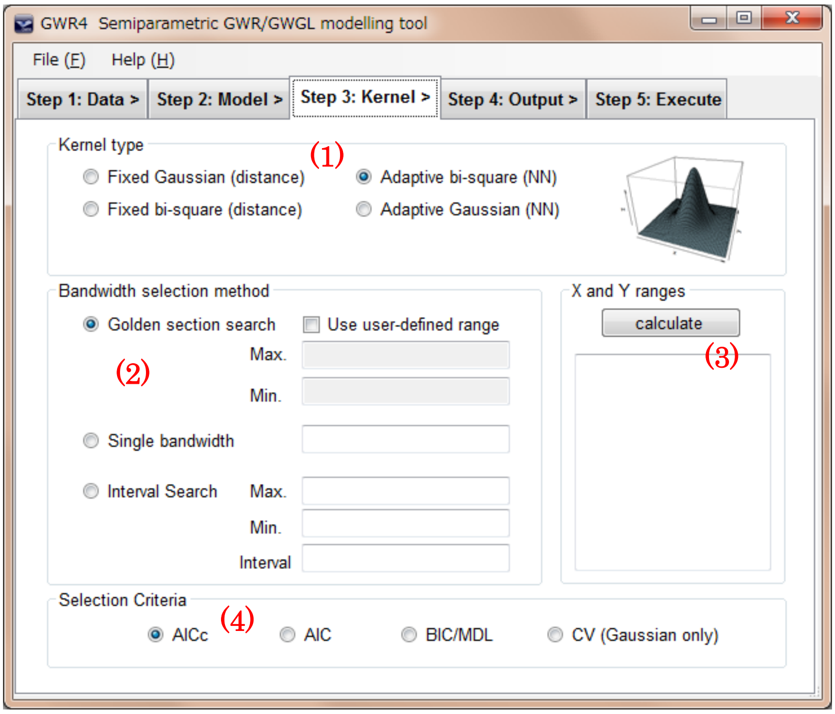


Figure 6.1: Screenshot of the default kernel tab

< What to do in this tab page>

In this tab page, a user is required to select the kernel function for geographical weighting to estimate local coefficients, its bandwidth size, and model selection criteria that are necessary for finding the best bandwidth and for comparing it with other modelling results using the same data. The default settings on this tab page can be used for most modelling cases.

(1) < kernel function type >: Choose one of the four available options for geographical kernel weighting.

(2) < bandwidth selection method >: Choose one of the three available options for bandwidth size selection. A larger bandwidth will estimate geographically smoother coefficients for local terms. The “golden section search” option can be used to automatically determine the best bandwidth size.

(3) < checking the coordinates ranges (optional) >: This is an optional function. If you would like to know the ranges of the x-y coordinates of your data for selecting bandwidth sizes, click the “calculate” button.

(4) < selection criteria >: Choose one of the four available options for model comparison criteria, which are mainly used for bandwidth selection.

Possible fixed and adaptive kernel functions for geographical weighting

If you select a fixed kernel, the geographic extent for local model fitting to estimate geographically local coefficients is constant over space. On the other hand, adaptive kernel changes such a local extent by controlling the k-th nearest neighbour distance for each regression location.

Classic options of geographic kernel type for GWR are "Gaussian fixed kernel" and "Adaptive bi-square kernel". Gaussian kernel weight continuously and gradually decreases from the centre of the kernel but never reaches zero. Gaussian kernel is suitable for fixed kernels since it can avert or mitigate the risk of there being no data within a kernel. Bi-square kernel has a clear-cut range where kernel weighting is non-zero. It is suitable for when you want to clarify local extents for model fitting. In the case of adaptive kernel, the number of areas included in the kernel is kept constant so that using bi-square kernel is secure.

Table 6.1: Four kernel type options available in GWR4

Fixed Gaussian	$w_{ij} = \exp(-d_{ij}^2 / \theta^2)$
Fixed bi-square	$w_{ij} = \begin{cases} (1 - d_{ij}^2/\theta^2)^2 & d_{ij} < \theta \\ 0 & d_{ij} > \theta \end{cases}$
Adaptive bi-square	$w_{ij} = \begin{cases} (1 - d_{ij}^2/\theta_{i(k)})^2 & d_{ij} < \theta_{i(k)} \\ 0 & d_{ij} > \theta_{i(k)} \end{cases}$
Adaptive Gaussian	$w_{ij} = \exp(-d_{ij}^2 / \theta_{i(k)}^2)$

Notes: i is the regression point index; j is the locational index;

w_{ij} is the weight value of observation at location j for estimating the coefficient at location i.

d_{ij} is the Euclidean distance between i and j;

θ is a fixed bandwidth size defined by a distance metric measure.

$\theta_{i(k)}$ is an adaptive bandwidth size defined as the k th nearest neighbour distance.

However, you may want to use different combinations of kernel types (fix or adaptive) and kernel functions (Gaussian or bi-square). For example, you may know that regression points are evenly distributed and using a bi-square function is secure even for a fixed kernel option. Or, in the case of GWLR where outcome distribution is unbalanced (e.g. a case-control study of a rare disease), bi-square kernel is not secure even for adaptive kernel and you may want to try Gaussian adaptive kernel as a securer option.

Bandwidth selection routines

< Golden section search >

To automatically search for the optimal bandwidth size, there are two options; golden-section search and interval search. In most cases, it is expected that golden-section search will efficiently identify the optimal bandwidth size. You can limit the search range for golden-section search by turning on the "Use user-defined range" switch. The default lower limit for golden section search is set to roughly keep 40 degrees of freedom for local regression fitting. In some cases, it is better to set the minimum size for the search range at a large value, such as 100 (see Figure 7.1 as an example). If the best bandwidth is estimated as either the max or min of the search ranges, GWR4 gives a warning message. In such a case, it is recommended that different min and/or max search ranges be tried.

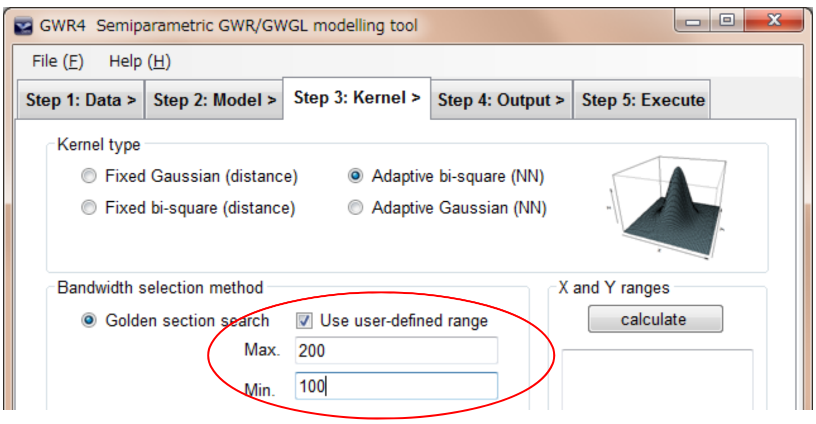


Figure 6.1: Example of golden section search use with user-defined range from 100 to 200 nearest neighbours and using an adaptive kernel)

For user-defined bandwidth specification, the number should be a distance metric (e.g. 15,000 m) without the metric symbol in the case of fixed kernel, while the number should be the number of nearest neighbours to be used for the local model fitting process in the case of adaptive kernel (e.g., if you input 35, the 35th nearest neighbour distance is used for each local model fitting).

< Interval search >

Interval search is a simple exhaustive search using a regular interval size bandwidth within a pre-specified range. Compared to the golden section search, the interval search is more intuitive and robust.

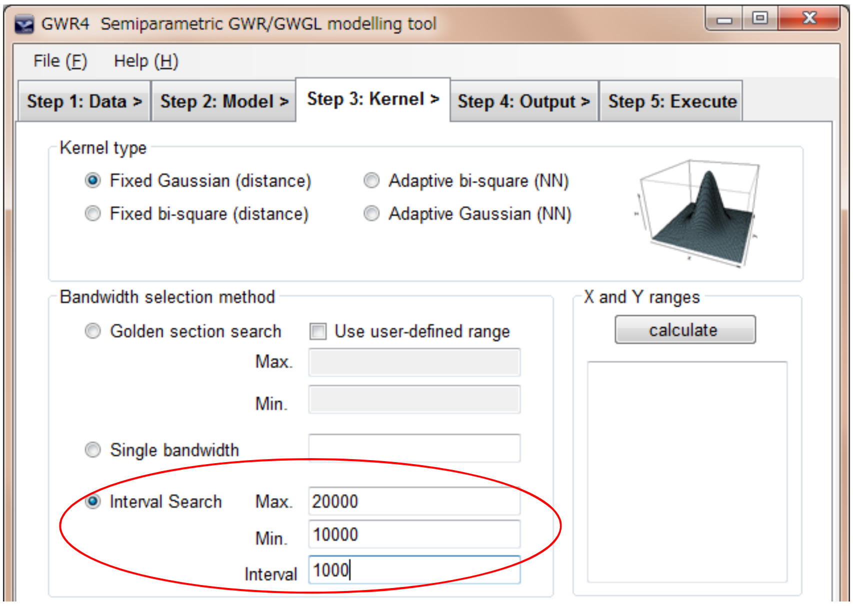


Figure 6.2: Example of interval search use from 10,000 to 20,000 m in 1,000 m steps (10,000, 11,000, 12,000, …, 19,000, 20,000) and using a fixed kernel

< Single bandwidth >

The simplest option is a single bandwidth size where you can provide a specific number for the bandwidth size.

Selection criteria

In the golden section and interval searches, the optimal bandwidth size is determined by means of comparison of model selection indicators with different bandwidth sizes. The criterion is also used for several modelling options described previously.

< AICc and AIC >
The default option is AICc (small sample bias corrected AIC) which is the most suitable in terms of statistical prediction for local Gaussian regression modelling where the local degree of freedom is likely to be small. Classic AIC tends to choose smaller bandwidths by which geographically varying coefficients are likely to be undersmoothed. The bias correction for Poisson and logistic models by AICc is not theoretically justified. However, AICc empirically provides better results even for Poisson regression.

< BIC/MDL and CV >
BIC/MDL is a new option that tends to choose larger bandwidth sizes. The indicator is appropriate for arguing the degree of complexity of the process to be analysed rather than statistical prediction of unobserved outcomes. Another classic but robust indicator, cross validation (CV) is applicable only to Gaussian models.

7. Step 4: The Output Tab

Session Control File

A user should input the name of your session control file in the top red-coloured textbox by clicking the rightmost “Browse” button to open a file dialog box. All of your settings, including filenames, model specifications, and modelling options will be saved in this simple text-formatted file when the session is run or the “save session” command in the File menu is issued.

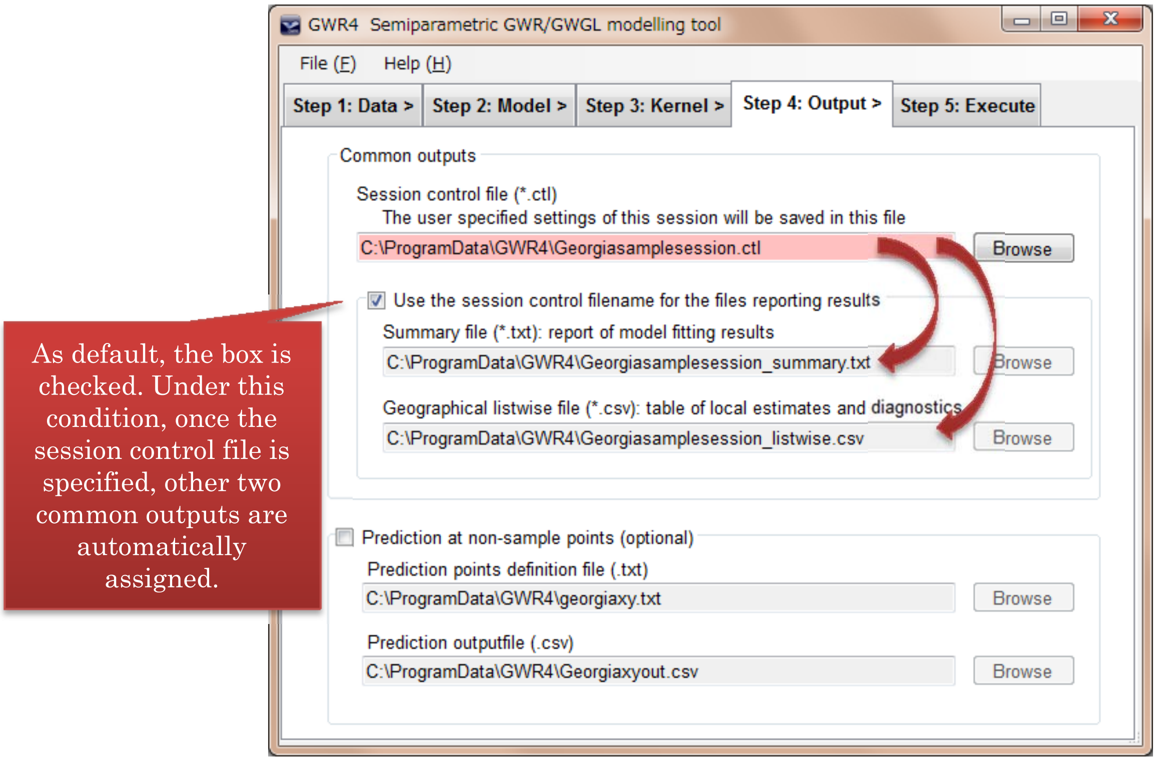


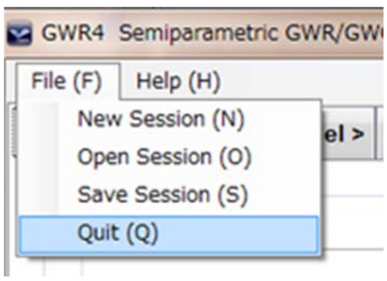
Figure 7.1: Sample screenshot of “Step 4: Output”

You may also save the current setting, open a stored session control file, or initialise the current setting at any time by using the File menu, as outlined below:

“New session” initialises the current session. Every setting reverts to its default.

“Open session” opens an existing session file (*.ctl file).

“Save session” creates a text file with a .ctl extension.



Common output files

In all cases, the modelling result is summarised in “Summary file” which is a plain-text file, and “Geographically listwise file” which lists estimated local statistics, including geographically varying coefficients and local diagnostic indicators, is saved in a CSV file. As default, the two common outputs are automatically assigned with the filenames and filepaths which the session control file is currently using. If the box of “Use the session control filename …” is unchecked, you can provide other filenames and filepaths for the two common output files.

The “Prediction at non-regression points” option

When this box is checked, you can provide a list of x-y coordinates for estimating local geographical coefficients and local goodness-of-fit indicators. This can be used for displaying a gridded surface of geographically varying coefficients or diagnosis of cross-validation design.

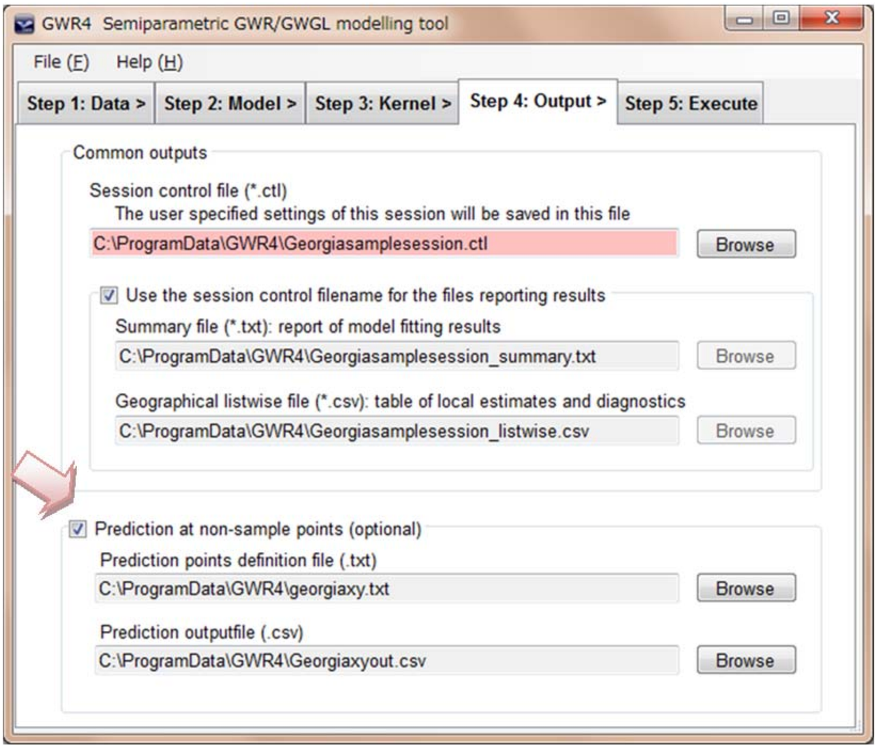


Figure 7.2: Sample screenshot of “Step 4: Output” when “Prediction at non-regression points” is turned on.

A prediction point definition file must be a text file in which each row contains x and y coordinates, and assumes that the first column lists x coordinates and the second column lists y coordinates. The first row is assumed to be used for field name listing. Any delimiter of space, comma, and tab can be used and there is no need to specify which delimiter is used. A sample is given below.

Table 7.1: Sample data of prediction point definition file

```
x,  y
599500,142200
575400,167200
530300,177300
524100,170300
426900,514600
```

The output file is automatically formatted as a CSV, comma delimited, text file.

8. Step 5: The Execute Tab

The Execute button

When you are ready to execute the session to fit your GWR model, click the "Execute this session" button to begin the model fitting computation. Running status information will appear during the computation but it may proceed slightly slowly.

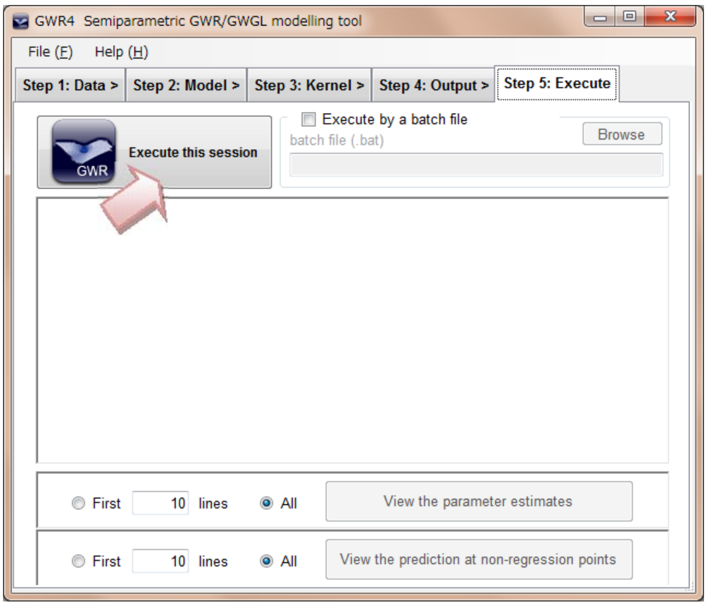


Figure 8.1: Sample screenshot of "Step 5: Execute"

When the computation ends, a results-summary appears in the large textbox. The content includes your modelling settings, global model result, best bandwidth, model diagnostic information of the GWR model, and fixed coefficients if you used them for the fitted model. These are automatically saved in the summary output file specified in the previous tab page.

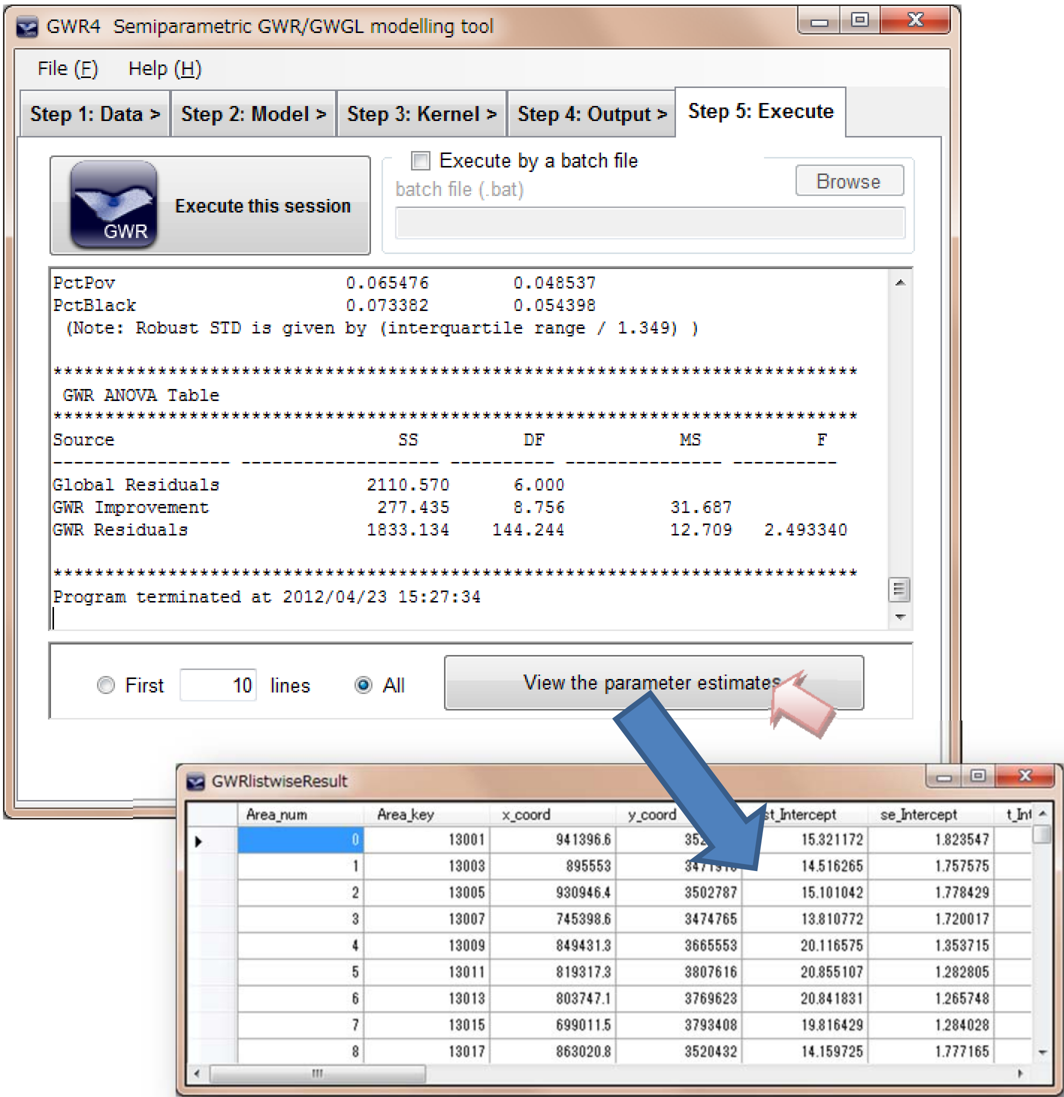


Figure 8.2: Sample screenshot of the output when model fitting ends

In addition, the “View geographically listwise result” button is now enabled. When you click it, a gridded window appears to show geographically listwise information, including local estimates of coefficients and model diagnostic information, in a tabular format. The information is stored in the CSV file of the geographically listwise result specified in the previous tab page.

Fields in a listwise output

The fields in a listwise table are as follows:

-------------------------------------from here---
Area_num: the sequential ID number of the location (automatically assigned),
Areal_key: if you selected this in the model tab, the field will be included.
x_coord: x coordinate of the regression points (data observations)
y_coord: y coordinate of the regression points (data observations)
est_intercept: estimate of the local constant term
se_intercept: standard error of the local constant term
t_intercept: pseudo t value (estimate / standard error) of the local constant term
#loop for each independent variable with geographically varying coefficient

```
{
    est_variablename: local estimate of coefficient
    se_variablename: local standard error
    t_variablename: local pseudo t value
}
```

#loop end
y: observed value of the dependent variable
yhat: predicted dependent variable
residual: y – yhat
std_residual: studentised residual
localR2: local R square
influence: influence indicator
cooksD: Cook's D indicator

-------------------------------------end---

In the case of GWGLM, std_residual, localR2, and cooksD are not included, while local percent deviance explained is included as the local goodness-of-fitness defined as below:

$$pdev_i = 1 - dev_i / nulldev_i$$

where dev_i is the locally weighted deviance of the fitted model and $nulldev_i$ is the locally weighted deviance of the null model having only a constant term at location i.

When you turn on the “prediction at non-regression points” switch in the Output tab page, another button, “View prediction at non-regression points” is enabled. Clicking the button will open another spreadsheet-like window. The field information is the same as that stored in the prediction output file but the values are coefficients or local diagnostic indicators at non-regression points provided in the “Output” tab page.

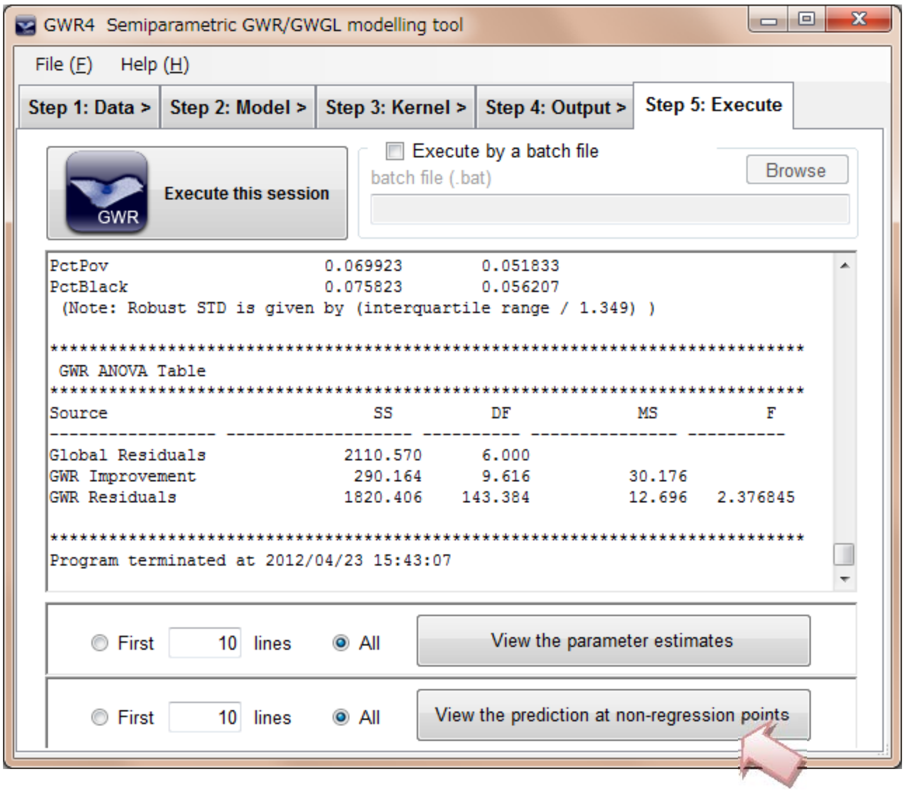


Figure 8.3: Sample screenshot of the output when model fitting ends for a case where the “prediction at non-regression” switch was turned on.

Handling a session and batch mode (optional)

By using the graphical user interface (GUI) based interface, the necessary settings can be assigned interactively. A session control setting is automatically saved when the session is run by clicking the “Execute this session” button. Alternatively, you can save the control file at any time by selecting “Save session” in the File menu. The session file is also be used in this application’s batch mode. Open a command prompt and then type “sgwrwin” plus the desired session filename. The application will then carry out the computations necessary to calibrate the model described in the session file without using the GUI.

The full path to sgwrwin.exe is needed in order for the program to be executed via the command prompt.

e.g. > "C:/Program Files (x86)GWR4/sgwrwin" mysession.ctl
(Note: ">" is the command prompt.)

If the GWR4 install folder is included in the PATH environment variable, you can omit the install folder for the execution at the command prompt.

e.g.   > set path=%path%;"c:¥program files¥GWR4"
        > sgwrwin mysession.ctl

The easiest way to create such a batch file is to use the interface of the Execute tab. With the "Execute by a batch file" box checked, you can specify the path and name of the batch file. Then, by clicking the "Execute this session" button, you can create the batch file and then choose to execute it immediately or at a later time.

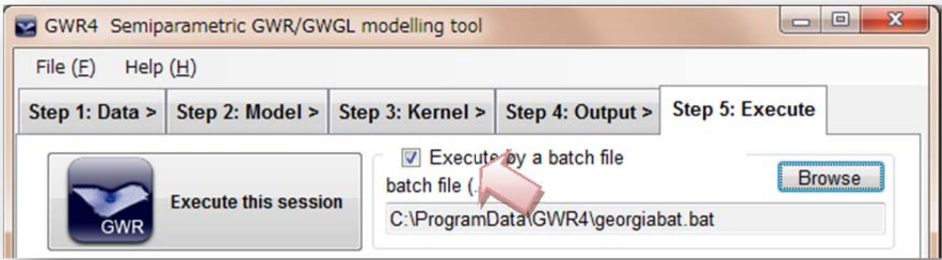


Figure 8.4: Sample screenshot of the process to create a batch file.

Example output

Imagine an example of Gaussian GWR with two local terms and two global terms using the Georgia data sample:

$$\begin{aligned}\text{PctBatch}_i &= \beta_0(X_i,Y_i)+\beta_1(X_i,Y_i)\text{PctRural}_i \\ &+\beta_2(X_i,Y_i)\text{PctPov}_i+\beta_3(X_i,Y_i)\text{PctBlack}_i+\varepsilon_i\end{aligned}$$.

When the model is fit with the geographical variability test, the adaptive kernel function, the golden section search for finding the optimal bandwidth size, and AICc as the model indicator for selecting the optimal bandwidth, the output file will resemble the one below:

```
------------------------------from here-----------------------------------------
******************************************************************************
*           Semiparametric Geographically Weighted Regression
*                     Release 1.0.70 (GWR 4.0.70)
*                          7 May 2012
*              (Originally coded by T. Nakaya: 1 Nov 2009)
*
*            Tomoki Nakaya(1), Martin Charlton(2), Paul Lewis(2),
*              A. Stewart Fotheringham (3), Chris Brunsdon (4)
*                    (c) GWR4 development team
* (1) Ritsumeikan University, (2) National University of Ireland, Maynooth,
*        (3) University of St. Andrews, (4) University of Liverpool
******************************************************************************

Program began at xxxxx 16:25:07

**************************************************
Session: Georgia data
Session control file: C:¥ProgramData¥GWR4¥Georgiasamplesession.ctl
******************************************************************************
```

The session title and control file specified in the Data tab

```
Data filename: C:¥ProgramData¥GWR4¥GeorgiaData.csv
Number of areas/points: 159
```

The data file and the number of records (areas/points)

```
Model settings---------------------------------
Model type: Gaussian
Geographic kernel: adaptive bi-square
Method for optimal bandwidth search: Golden section
Criterion for optimal bandwidth: AICc
Number of varying coefficients: 4
Number of fixed coefficients:   2
```

The "Model settings" section reflects your specification of GWR model in the Model tab page.

Two local terms and two global terms are used.

```
Modelling options-------------------------------
```

```
Standardisation of independent variables: OFF
Testing geographical variability of local coefficients: On
Local to Global Variable selection: OFF
Global to Local Variable selection: OFF
Prediction at non-regression points: On

Variable settings--------------------------------
Area key: field1: AreaKey
Easting (x-coord): field12 : X
Northing (y-coord): field13: Y
Cartesian coordinates: Euclidean distance
Dependent variable: field6: PctBach
Offset variable is not specified
Intercept: varying (Local) intercept
Independent variable with varying (Local) coefficient: field5: PctRural
Independent variable with varying (Local) coefficient: field9: PctPov
Independent variable with varying (Local) coefficient: field10: PctBlack
Independent variable with fixed (Global) coefficient: field7: PctEld
Independent variable with fixed (Global) coefficient: field8: PctFB
Number of prediction at non-regression points   19
*****************************************

*****************************************************************************
  Global regression result
*****************************************************************************

  < Diagnostic information >
Residual sum of squares:              2110.569589
Number of parameters:                      6
 (Note: this num does not include an error variance term for a Gaussian model)
ML based global sigma estimate:            3.643353
Unbiased global sigma estimate:            [illegible]
Log-likelihood:                         862.366074
Classic AIC:                           876.366074
AICc:                                  877.107796
BIC/MDL:                               897.848403
CV:                                    15.331159
R square:                               0.588429
Adjusted R square:                         0.572182

Variable                Estimate    Standard Error     t(Est/SE)
-------------------- -------------- -------------- --------------
```

Advanced options in the Model tab page

"Variable settings" summarises which fields of your data are assigned for GWR variables.

If you estimate coefficients at non-sample points used the Output tab option, the number of non-sample points is displayed.

Global model means traditional regression model with fixed coefficients.

```
Intercept                17.243732       1.753292       9.835062
PctRural                 -0.070323       0.013579      -5.178928
PctPov                  -0.255236       0.072477      -3.521617
PctBlack                  0.049114       0.026485       1.854437
PctEld                   0.011448       0.129535       0.088377
PctFB                    1.852471       0.306830       6.037452

***********************************************
  GWR (Geographically weighted regression) band
***********************************************
Bandwidth search <golden section search>
  Limits: 52,  159
 Golden section search begins...
 Initial values
  pL           Bandwidth:    52.000 Criterion:
  p1           Bandwidth:    92.870 Criterion:
  p2           Bandwidth:   118.130 Criterion:
  pU           Bandwidth:   159.000 Criterion:
 iter    1 (p2) Bandwidth:   118.130 Criterion:
 iter    2 (p1) Bandwidth:   118.130 Criterion:    87         Diff:     15.611
 iter    3 (p2) Bandwidth:   118.130 Criterion:    870.444 Diff:      9.648
 iter    4 (p1) Bandwidth:   118.130 Criterion:    870.444 Diff:      5.963
 iter    5 (p1) Bandwidth:   114.444 Criterion:    870.374 Diff:      3.685
 iter    6 (p1) Bandwidth:   112.167 Criterion:    870.289 Diff:      2.278
 iter    7 (p2) Bandwidth:   112.167 Crit
Best bandwidth size 112.000
Minimum AICc      870.289

**************************************
  GWR (Geographically weighted regression) result
*****************************************************************************
  Bandwidth and geographic ranges
Bandwidth size:                 112.166731
Coordinate                Min              Max           Range
--------------- --------------- --------------- ---------------
X-coord          635964.300000  1059706.000000   423741.700000
Y-coord         3401148.000000  3872640.000000   471492.000000

  Diagnostic information
Residual sum of squares:         1812.11963
Effective number of parameters (model: trace(S)):                13.513005
```

If golden section search is selected, the iterative steps to identify the optimal bandwidth size are reported.

For each step, the current best bandwidth size and model comparison indicator, and the difference in bandwidth size for the current and previous iterative steps.

The best bandwidth size and its model comparison indicator. Since the model is fitted using an adaptive kernel, the nearest 112 samples are used to estimate the local coefficients.

The ranges of the geographic coordinates of sample observations are reported.

```
Effective number of parameters (variance: trace(S'S)):          11.248290
Degree of freedom (model: n - trace(S)):                    145.486995
Degree of freedom (residual: n - 2trace(S) + trace(S'S)):        143.222281
ML based sigma estimate:            3.375941
Unbiased sigma estimate:            3.557035
Log-likelihood:                   838.124834
Classic AIC:                      867.150843
AICc:                             870.288971
BIC/MDL:                          911.689864
CV:                               15.670568
R square:                          0.646628
Adjusted R square:                   0.607426
```

Model diagnostic indicators of the fitted GWR model. For example, AICc of GWR (870.3) is clearly smaller than that of the global regression model (877.1).

```
*********************************************
 << Fixed (Global) coefficients >>
*********************************************
 Variable           Estimate       Standard Error  t(Estimate/SE)
-------------------- --------------- --------------- ---------------
PctEld                -0.085045       0.197299      -0.431044
PctFB                  1.637489       0.345444       4.740252
```

If you specify global terms in the Model tab page, estimates of global term coefficients are tabulated.

```
***********************************************************
 << Geographically varying (Local) coefficie
*********************************************
Estimates of varying coefficients have been saved in the following file.
   Listwise output file: C:¥Users¥tomoki¥Documents¥defaultGWRlistwise.csv
```

Estimates and local diagnostic indicators are stored in the CSV file.

```
Summary statistics for varying (Local) coefficients
Variable                   Mean           STD
-------------------- --------------- -------------
Intercept              17.701755       2.789388
PctRural               -0.078407       0.020370
PctPov                -0.164434       0.052453
PctBlack                0.036442       0.050993
```

Basic statistics of estimated coefficients of local terms

```
Variable                    Min            Max          Range
-------------------- --------------- --------------- ---------------
Intercept              13.492842       21.890966       8.398124
PctRural               -0.112511       -0.041550       0.070960
PctPov                -0.281232       -0.073793       0.207439
PctBlack               -0.049286        0.139190       0.188476
```

```
Variable             Lwr Quartile        Median    Upr Quartile
-------------------- --------------- --------------- ---------------
Intercept               14.964686      18.011195      20.355776
PctRural                -0.096953      -0.078303      -0.063119
PctPov                  -0.204178      -0.155714      -0.129319
PctBlack                -0.003732       0.035008       0.073043

Variable            Interquartile R     Robust STD
-------------------- --------------- ---------------
Intercept                5.391090       3.996360
PctRural                 0.033834       0.025081
PctPov                   0.074858       0.055492
PctBlack                 0.076775       0.056913
 (Note: Robust STD is given by

******************************                                          ****
 GWR ANOVA Table
******************************                                          ****
```

Using GWR ANOVA, a simple test to find out if the global (traditional) regression model and the GWR model have the same statistical performance (the same size of error variance) can be carried out.

```
Source                           SS          DF            MS          F
----------------- ------------------- ----------
Global Residuals             2110.570      6.000
GWR Improvement               298.450      9.778
GWR Residuals                1812.120    143.222

*************************************************
 Geographical variability tests of local coeffici
*************************************************              *********
```

If you select the option for geographical variability test in the Model tab page, the tabulated report shows the test results. A positive “DIFF of Criterion”, especially greater than or equal to two, suggests that the local term is better to be assumed as global.

```
 Variable          F              DOF for F test  DIFF of Criterion
-------------------- ------------------ ---------------- -----------------
Intercept                 5.367001    1.860  145.487        -6.069357
PctRural                  4.303917    2.191  145.487        -4.720029
PctPov                    1.976232    1.591  145.487         0.440477
PctBlack                  7.985049    1.662  145.487        -9.870218
-------------------- ------------------ ---------------- -----------------
Note: positive value of diff-Criterion (AICc, AIC, BIC/MDL or CV) suggests no spatial
variability

*****************************************************************************
Program terminated at xxx 16:25:07
------------------------------ end ----------------------------------------
```

9. References

Most of the theories underlying GWR and the definition of diagnostic indicators are explained in Fotheringham et al. (2002), except for the details of backfitting semiparametric modelling in this software, which is different from that explained in the book. See the details in Nakaya et al. (2005). The core routine to fit semiparametric GWGLM of GWR4 was originally developed in C for a research on disease associative mapping (Nakaya et al., 2005).

Comprehensive information on GWR:

Stewart Fotheringham, AS, Brunsdon, C, Charlton, M, 2002, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Wiley, 282 pages. ISBN: 978-0-471-49616-8

Semiparametric modelling of GWGLM:

Nakaya, T., Fotheringham, S., Brunsdon, C. and Charlton, M. (2005): Geographically weighted Poisson regression for disease associative mapping, *Statistics in Medicine* 24, 2695-2717.

Compared to the previous version (GWR3.x), GWR4 provides almost same results for traditional GWR modelling. A few corrections have been made with regards to calculation methods for local diagnostic statistics, including local sigma and local R square.