The Spatial Distribution of Poverty

A geographically weighted regression

—by Sumeeta Srinivasan

Introduction

Problem

How can we explore the spatial distribution of poverty and determine its correlates? This exercise examines data from Sri Lanka. Many quantitative studies use ordinary least squares (OLS) regression to estimate the effect of variables such as ethnicity, proximity to urban areas, elevation, and other indicators of development on poverty rates. This exercise uses a more generalized geographically weighted regression (GWR) model in addition to the OLS model to incorporate the effects of spatial clustering.

Location

Sri Lanka

Time to complete the lab

Three hours

Keywords: poverty estimation; development studies; South Asia; Sri Lanka; ordinary least squares (OLS) regression; weight matrices; weight functions; local and global multicollinearity; geographically weighted regression (GWR)

Prerequisites

- Familiarity in the use of ArcGIS® 10
- Understanding of OLS regression and GWR including diagnostic statistics
- Understanding of spatial statistics such as Moran's I

Data used in this lab

SriLankaCaseStudy (Amarasinghe et al. 2005): A shapefile for Sri Lanka including poverty counts at the Divisional Secretariat level

- Projection: Transverse Mercator
- Scale factor at central meridian*:* 0.999600
- Longitude of central meridian*:* 81
- Latitude of projection origin*:* 0.000000
- False easting*:* 500000.000000
- False northing*:* 0.000000

Student activity

In this lab, you will use ordinary least squares regression and geographically weighted regression to estimate the relationship between poverty and various social, economic, and geographic factors at the Divisional Secretariat level. Other studies have found a strong clustering of residuals—that is, a spatial autocorrelation between the residuals, which is a violation of the model assumptions. There are at least three strategies for dealing with spatial autocorrelation in regression model residuals:

1. Resample until the input variables no longer exhibit statistically significant spatial autocorrelation. While this does not ensure the analysis is free of spatial autocorrelation problems, they are far less likely when spatial autocorrelation is removed from the dependent and explanatory variables. This is the traditional statistician's approach to dealing with spatial autocorrelation and is only appropriate if spatial autocorrelation is the result of data redundancy (the sampling scheme is too fine). (You can perform this test in most statistical packages.)
2. Isolate the spatial and nonspatial components of each input variable using a spatial filtering regression method. Space is removed from each variable but is put back into the regression model as a new variable to account for spatial effects and spatial structure (Getis 2010). (You can do this to some extent in R by using lagged variables.)
3. Incorporate spatial autocorrelation into the regression model using spatial econometric regression methods (Klieber and Zeileis 2008).

The OLS regression model with one predictor variable is of the form

$$y = \beta 0 + \beta 1x1 + \varepsilon$$

where y is the dependent variable and x1 is the explanatory variable. β0 and β1 are parameters to be estimated. ε is a mean zero random error term with constant (but unknown) variance and normally distributed. Assuming that these conditions are satisfied, the parameters can be estimated using OLS. In matrix form, the solution is given by

$$\beta = (XTX)\text{-}1XTY$$

where X is the "design" matrix of explanatory variables and Y is the vector of values of the dependent variable. This matrix formulation generalizes to the case of multiple explanatory variables.

The OLS regression model is a "global" model in that it does not explicitly take spatial location into account, although sometimes coordinates are used as explanatory variables. Global models best describe the overall data relationships in a study area. When these relationships are consistent across the study area, OLS regression estimates these relationships well. When these relationships behave differently in different parts of the study area, however, the regression equation is more of

an average of the mix of relationships present. Where these relationships represent extremes, the global average does not model either extreme well.

When your explanatory variables exhibit regional variation, global models tend to produce unsatisfactory results. There are at least four ways to deal with regional variation in OLS regression models:

1. Include a variable in the model that explains the regional variation. If you see that your model is always overpredicting in the north and underpredicting in the south, for example, add a regional variable set to 1 for northern features and 0 for southern features.
2. Use methods that incorporate regional variation into the regression model such as geographically weighted regression or spatial expansion.
3. Use robust regression standard errors and probabilities to determine whether the regression coefficients are statistically significant. Geographically weighted regression is still recommended.
4. Redefine or reduce the size of the study area so that the processes no longer exhibit regional variation.

One way to incorporate regional variation into the model is to make the regression coefficients depend on location coordinates. If the coefficients depend on location coordinates, the predicted value and the residual will do the same. OLS uses all the data at one time to estimate coefficients, but in that case, the coefficients would not depend on the location coordinates. Alternatively, a "weight matrix" can be used in geographically weighted regression. It would seem logical that the data for close locations would have higher weights than locations farther away, because two locations that are close together should be similar in most characteristics compared to locations that are far away from each other. GWR estimation allows selection of bandwidth to determine which locations should be considered neighbors and therefore more like each other. The bandwidth may either be chosen by the user or estimated using a technique such as cross-validation. The resultant parameter estimates are mapped to examine local variations in the parameter estimates. You might also want to map the standard errors of the parameter estimates to see where the GWR is giving you better estimators.

In this lab exercise, you will

1. Look at overall poverty patterns using descriptive spatial statistics.
2. Fit an OLS regression in ArcGIS to model the relationship between poverty counts and various other variables.
3. Fit a GWR in ArcGIS to see the spatial effects of various developmental variables on poverty.

Prepare your workspace

Data preparation, storage, and backup are basic and crucial when doing a GIS project. It is good practice to store all your data within a single folder on your computer or storage device.

To begin, create a workspace to keep all data for this lab.

1 Create a ***SriLanka*** folder where you will store the results of your work in this lab exercise.

2 Create a ***Data*** folder under the *SriLanka* folder.

Collect and process data

1 Download the *SriLankaCaseStudy* data from `http://gisweb.ciat.cgiar.org/povertymapping/`.

2 Add it to your data workspace (i.e., *SriLanka**Data*).

3 Click the *Geospatial Metadata* link for Sri Lanka. This information is useful in understanding the variables in the shapefile.

It is important to view data before moving on to the analysis because you need to see the variables that you have, the geographic unit available, and the spatial extent of the data.

4 Launch ArcCatalog™.

5 Examine the shapefile that you downloaded for Sri Lanka in the *Data* folder by looking at the metadata, the attribute table, and the geographic units.

Examine poverty patterns in Sri Lanka

Analyze

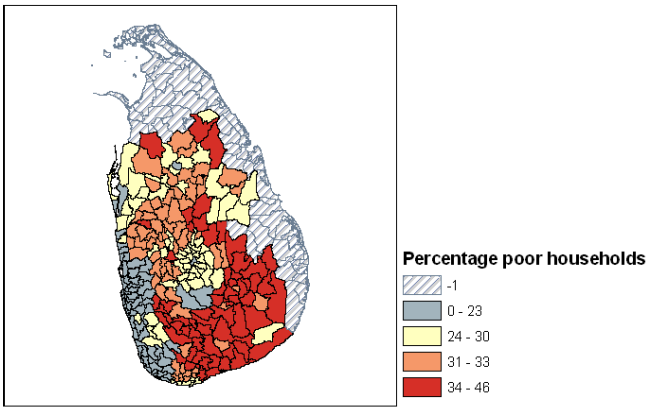
1 Map the dependent and explanatory variables to see the spatial patterns of distribution.

2 Launch ArcMap™ and add the shapefile *SriLankaCaseStudy*.

3 In the table of contents, right-click the shapefile and then click *Properties » Symbology* and then *Quantities » Graduated colors*. Then select the variables you want to map. First select *PCTPOOHH*.

VISUALIZE

Map 1 shows the distribution of the poverty variable *PCTPOOHH*.



Map 1: Distribution of *PCTPOOHH*

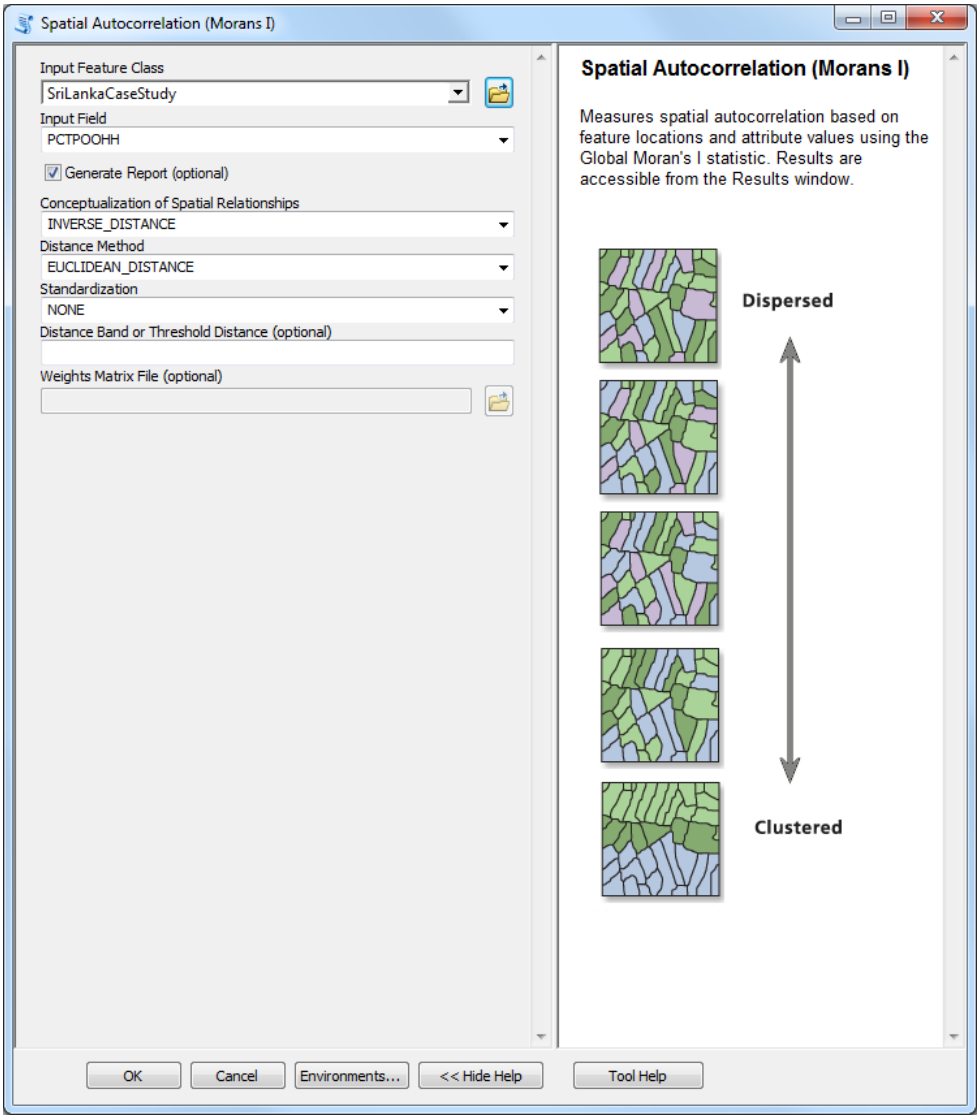
Question 1: *Identify the names of the locations that had the highest and lowest percentages of poor households. Do they appear to be clustered?*

Note that to discuss spatial variation, you should familiarize yourself with the names of the various regions and administrative units. It is extremely helpful to understand the locational context while discussing spatial distribution of phenomena like poverty.

ASSESS POVERTY CLUSTERING PATTERNS IN SRI LANKA

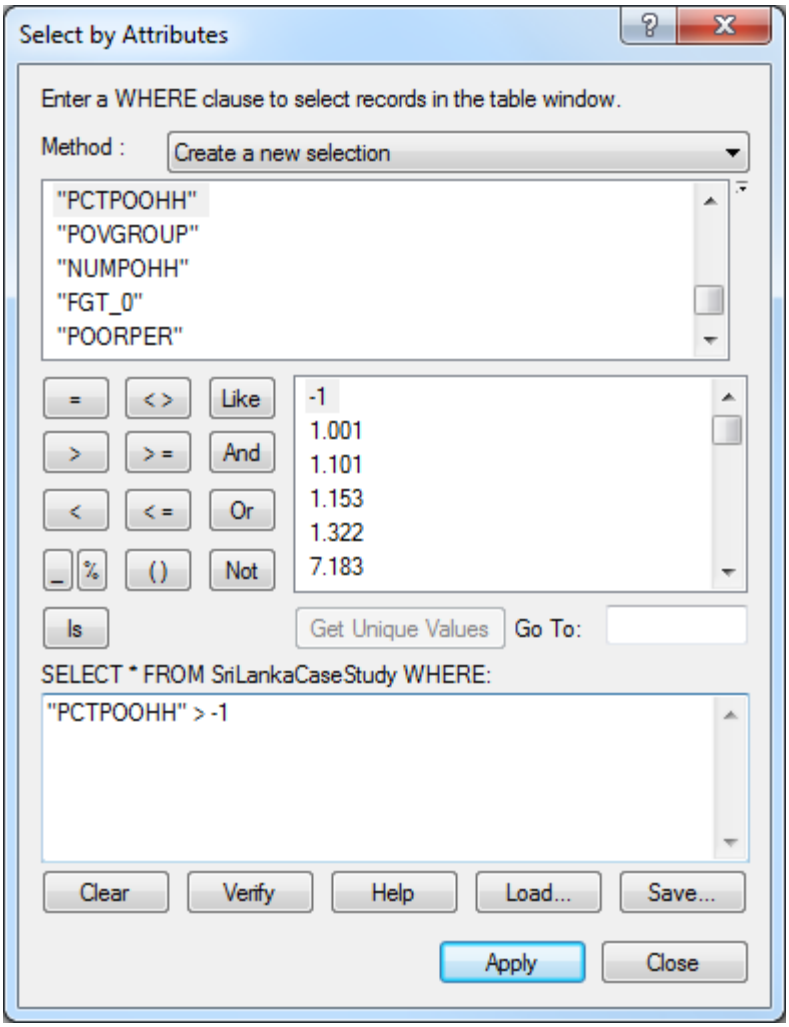
ANALYZE

1 You can assess the extent to which poverty may be clustered by using Moran's I. In ArcToolbox™, expand *Spatial Statistics Tools » Analyzing Patterns* and then double-click *Spatial Autocorrelation (Morans I)*. For *PCTPOOHH*, the Moran's I is 0.68 and suggests a highly clustered pattern in the location of the percentage of poor households in a district. Calculate the Moran's I for other variables of interest. For example, Amarasinghe et al. suggest that low agricultural employment and better access to roads are key characteristics of low-poverty locations.



Question 2: *Report the Moran's I for a few variables of interest.*

2 Note that several locations have no data. Eliminate them by creating a new shapefile for locations that have reported PCTPOOHH greater than -1. Name this shapefile *SriLankaCorrected*. On the main menu, click *Selection » Select By Attributes* to select locations as shown in the figure below. Then, in the table of contents, right-click the layer in the main menu and select *Data » Export Data* to export the selected features to a new shapefile.



3 Before fitting regressions to predict the number of poor households in a district, you first need to generate several new variables. Add four new variables: small holding size per agricultural operator (*SMHOLDPAG*), agricultural operators not owning land (%) (*AGOPWOLA*), acreage in small holdings below 1 acre (%) (*PACSMALL*), and the number of agricultural operators per household (*NUMAGHH*).

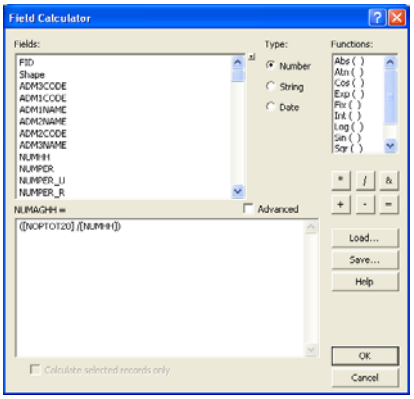
4 Open the attribute table of the shapefile (right-click it in the table of contents and click *Open Attribute Table*).

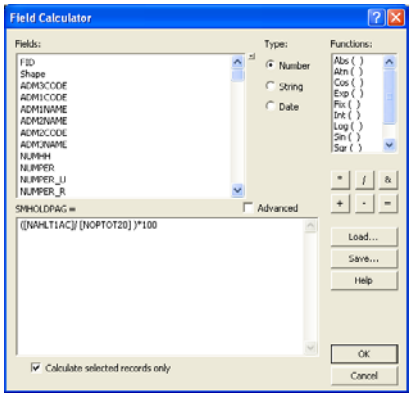
5 Click *Table Options* to then add fields, making sure to select the attribute type *Float*.

6 Add fields *SMHOLDPAG*, *AGOPWOLA*, *PACSMALL*, and *NUMAGHH*.

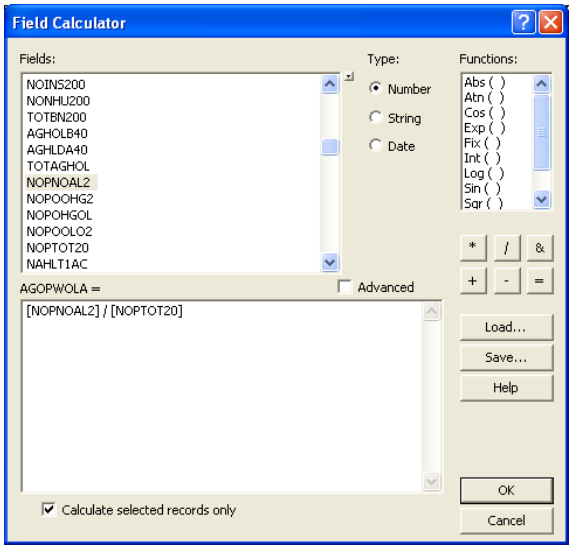
7 Calculate the field *NUMAGHH* by right-clicking the column for it and then clicking *Field Calculator* to get a new window as shown below. The number of agricultural operators per household *NUMAGHH* is the total number of agricultural operators *NOPTOT20* divided by the number of households *NUMHH*.

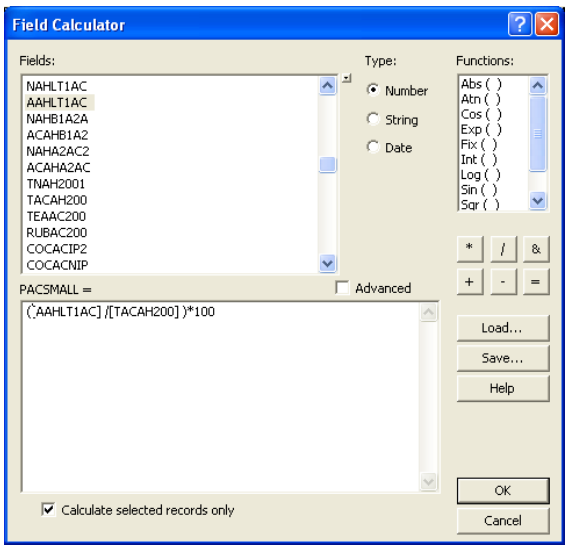
Make sure that you do not have any districts that have 0 values for the denominator before you do the field calculations.





8 *SMHOLDPAG* is the number of holdings below one acre divided by the total number of agricultural operators (*NAHLT1AC/NOPTOT20*) as a percentage, *AGOPWOLA* is the number of agriculture operators not owning any land divided by the total number of agricultural operators (*NOPNOAL2/NOPTOT20*), and *PACSMALL* is the acreage of agriculture holdings of all classes divided by the acreage of agriculture holdings with extent less than one acre (*AAHLT1AC/ TACAH200*) as a percentage. In each case, make sure that the records where the denominator is 0 are not selected.



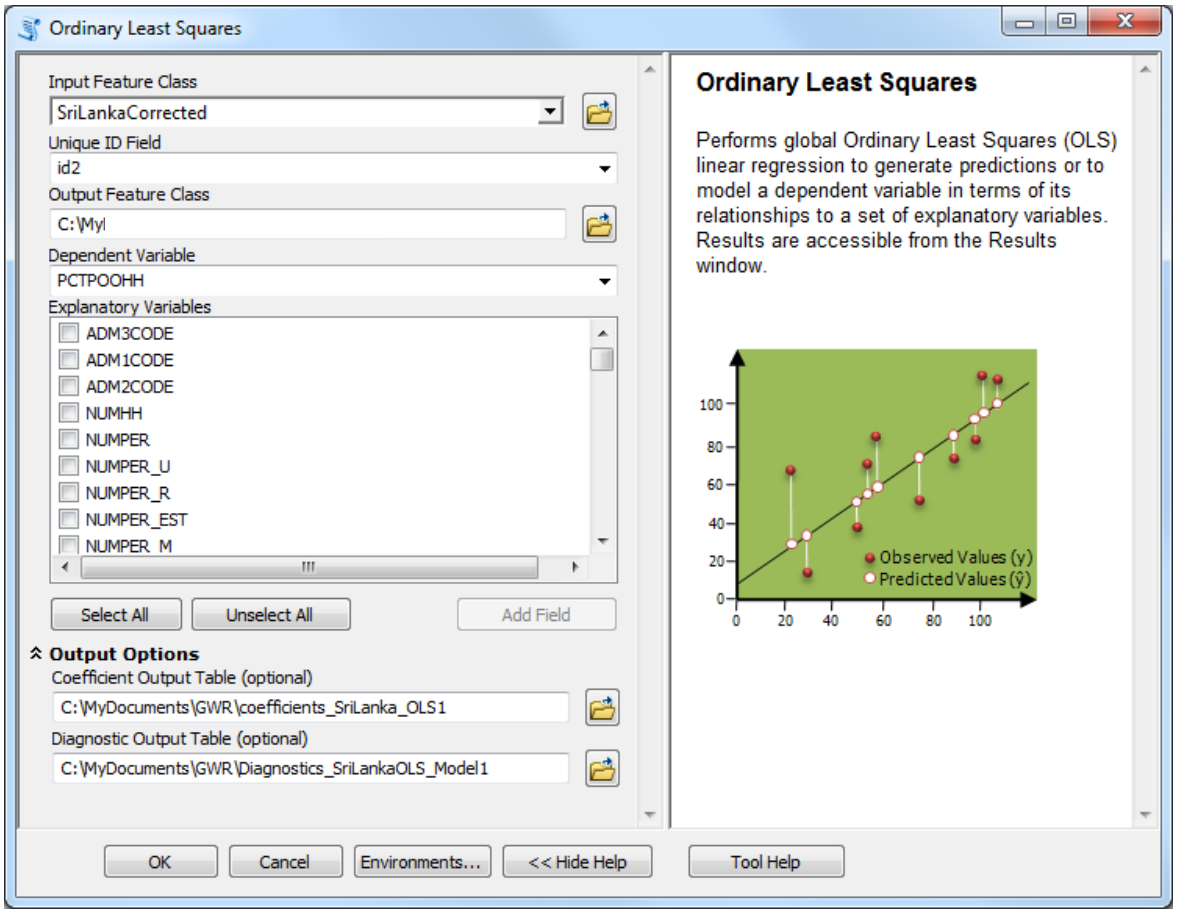


9 Now you can estimate the coefficients for an OLS regression of the percentage of poor households. In ArcToolbox, expand *Spatial Statistics Tools » Modeling Spatial Relationships* and then double-click *Ordinary Least Squares* to get an *Ordinary Least Squares* window. (See the following figure for an example of how to assign dependent variables.) Use *PCTPOOHH* as the dependent variable. Explanatory variables are as follows:

- Yala rainfall season: *YALARF*
- Maha rainfall season: *MAHARF*
- Paddy area under major irrigation: *ASSWMAJ3*
- Paddy area under minor irrigation: *ASSWMIN3*
- Distance to roads: *DISROADS*
- Distance to town: *DISTOWN*
- SMHOLDPAG, AGOPWOLA, PACSMALL, and NUMAGHH

Note that you need a unique ID, *id2*, for each record in your shapefile for ArcGIS to carry out an OLS calculation.

10 Click *OK* to run the function. The results are shown on the next page.



coefficients_SriLanka_OLS1

OID	Field1	Variable	Coef	StdError	t_Stat	Prob	Robust_SE	Robust_t	Robust_Pr
0	0	Intercept	14.696919	4.610992	3.187366	0.001639	5.344666	2.749829	0.006419
1	0	ASSWMAJ3	-0.000322	0.000113	-2.840593	0.004895	0.000123	-2.628643	0.009125
2	0	ASSWMIN3	-0.000448	0.000241	-1.856373	0.064638	0.000189	-2.37042	0.018554
3	0	DISROADS	-0.310127	0.342001	-0.906803	0.365416	0.302136	-1.02645	0.305713
4	0	DISTOWN	0.079159	0.045514	1.739228	0.083294	0.043254	1.830103	0.068488
5	0	MAHARF	-0.021496	0.004775	-4.501778	0.000013	0.004714	-4.56043	0.00001
6	0	YALARF	0.012102	0.005682	2.129844	0.034201	0.005566	2.17422	0.030663
7	0	SMHOLDPA	0.085394	0.087854	0.971997	0.332028	0.092158	0.926601	0.355061
8	0	AGOPWOLA	5.359811	5.825823	0.920009	0.358488	6.058761	0.884638	0.377233
9	0	PACSMALL	-0.111693	0.198094	-0.563836	0.573402	0.182872	-0.610769	0.541938
10	0	NUMAGHH	19.04301	2.193115	8.683089	0	2.386349	7.979975	0

Diagnostics_SriLankaOLS_Model1

OID	Field1	Diag_Name	Diag_Value	Definition
0	0	AIC	1614.111868	Akaike's Information Criterion: A relative measure of performance used to compare models; the smaller AIC indicates the superior model.
1	0	AICc	1615.433902	Corrected Akaike's Information Criterion: second order correction for small sample sizes.
2	0	R2	0.50612	R-Squared, Coefficient of Determination: The proportion of variation in the dependent variable that is explained by the model.
3	0	AdjR2	0.485369	Adjusted R-Squared: R-Squared adjusted for model complexity (number of variables) as it relates to the data.
4	0	F-Stat	24.389872	Joint F-Statistic Value: Used to assess overall model significance.
5	0	F-Prob	0	Joint F-Statistic Probability (p-value): The probability that none of the explanatory variables have an effect on the dependent variable.
6	0	Wald	266.788894	Wald Statistic: Used to assess overall robust model significance.
7	0	Wald-Prob	0	Wald Statistic Probability (p-value): The computed probability, using robust standard errors, that none of the explanatory variables have an effect on the dependent variable.
8	0	K(BP)	41.363093	Koenker's studentized Breusch-Pagan Statistic: Used to test the reliability of standard error values when heteroskedasticity (non-constant variance) is present.
9	0	K(BP)-Prob	0.00001	Koenker (BP) Statistic Probability (p-value): The probability that heteroskedasticity (non-constant variance) has not made standard errors unreliable.
10	0	JB	5.275213	Jarque-Bera Statistic: Used to determine whether the residuals deviate from a normal distribution.
11	0	JB-Prob	0.071532	Jarque-Bera Probability (p-value): The probability that the residuals are normally distributed.
12	0	Sigma2	36.353722	Sigma-Squared: OLS estimate of the variance of the error term.

Question 3: *Look at the fitted OLS regression of the percentage of poor households using the explanatory variables shown. What do the results mean? How would you interpret the coefficients of the significant variables? Also, interpret the diagnostics. For example, the statistically significant Koenker diagnostic test means that there may be nonstationarity in your model that GWR can account for. It is therefore appropriate to use a GWR model here.*

Question 4: *Map the residuals. Do they appear to be clustered? What is the Moran's I for the residuals?*

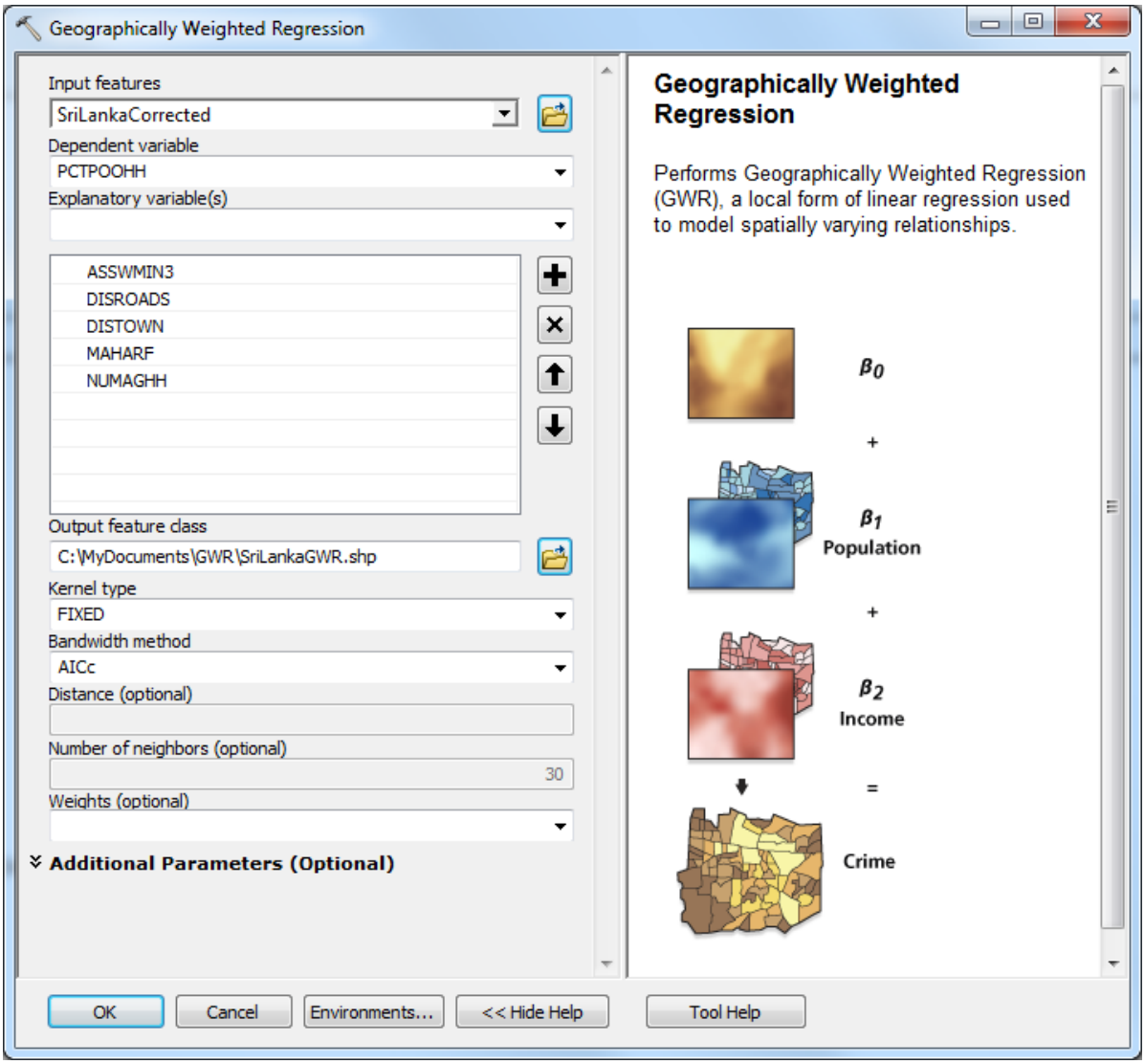
Note that the residuals showed positive spatial autocorrelation based on the global Moran's I. One way to deal with this problem is to run a GWR to see if the effects of the coefficients are different over the study area. Also see `www.esri.com/news/arcuser/0111/findmodel.html` for more explanations on interpreting and refining your results.

11 To run GWR within ArcToolbox, expand *Spatial Statistics Tools » Modeling Spatial Relationships* and then double-click *Geographically Weighted Regression* to get a window of the same name. Within this window, select the explanatory and dependent variables you used to estimate your OLS model in the last section.

12 Initially, you will not be able to run GWR because of multicollinearity. To check for global multicollinearity, remove explanatory variables that have large variance inflation factor (VIF) values (for example, above 7.5), which may be redundant. You can find the VIF values in your results (on the main menu, click *Geoprocessing » Results*).

13 Finding local multicollinearity is more difficult. Create a thematic map for each of the explanatory variables and look for areas with little or no variation in values. Also, avoid using dummy/binary variables, variables reflecting categorical/nominal data, or variables with only a few possible values.

14 To successfully run GWR, use the following variables: *ASSWMIN3*, *DISROADS*, *DISTOWN*, *MAHARF*, and *NUMAGHH*. You should map the coefficient surfaces as well as the residuals.



15 In addition to regression residuals, the output feature class table (which is the same type as the input) includes fields for observed and predicted y-values, condition number (Cond), local R^2, residuals, and explanatory variable coefficients and standard errors. Unlike OLS, where you interpret the coefficients for all the data, here you interpret the coefficient based on locations.

Note

Condition number: This diagnostic evaluates local collinearity. In the presence of strong local collinearity, results become unstable. Results associated with condition numbers larger than 30 may be unreliable.

Local R^2: These values range between 0.0 and 1.0 and indicate how well the local regression model fits observed y-values. Very low values indicate that the local model is performing poorly. Mapping the local R^2 values to see where GWR predicts well and where it predicts poorly may provide clues about important variables that might be missing from the regression model.

Predicted: These are the estimated (or fitted) y-values computed by GWR.

Residuals: To obtain the residual values, the fitted y-values are subtracted from the observed y-values. Standardized residuals have a mean of 0 and a standard deviation of 1. A cold-to-hot rendered map of standardized residuals is automatically added to the table of contents when GWR is executed in ArcMap.

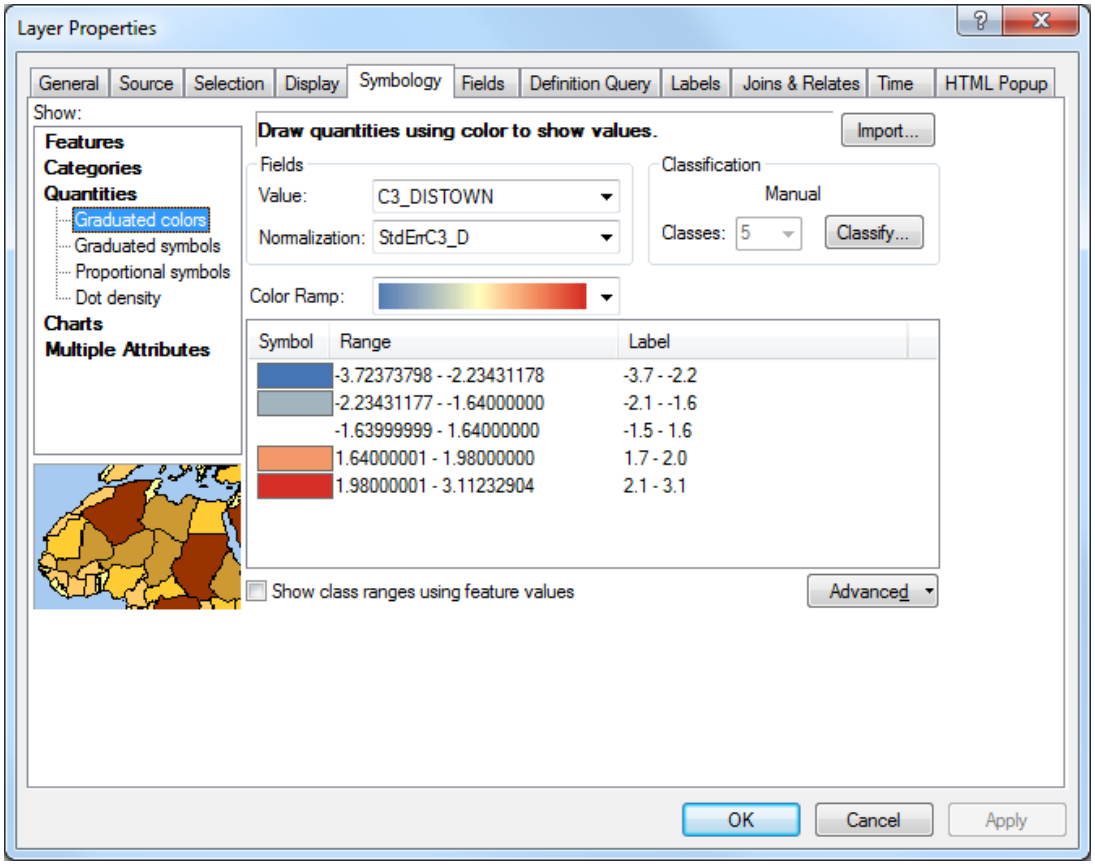
Coefficient standard error: These values measure the reliability of each coefficient estimate. Confidence in these estimates is higher when standard errors are small in relation to the actual coefficient values. Large standard errors may indicate problems with local collinearity.

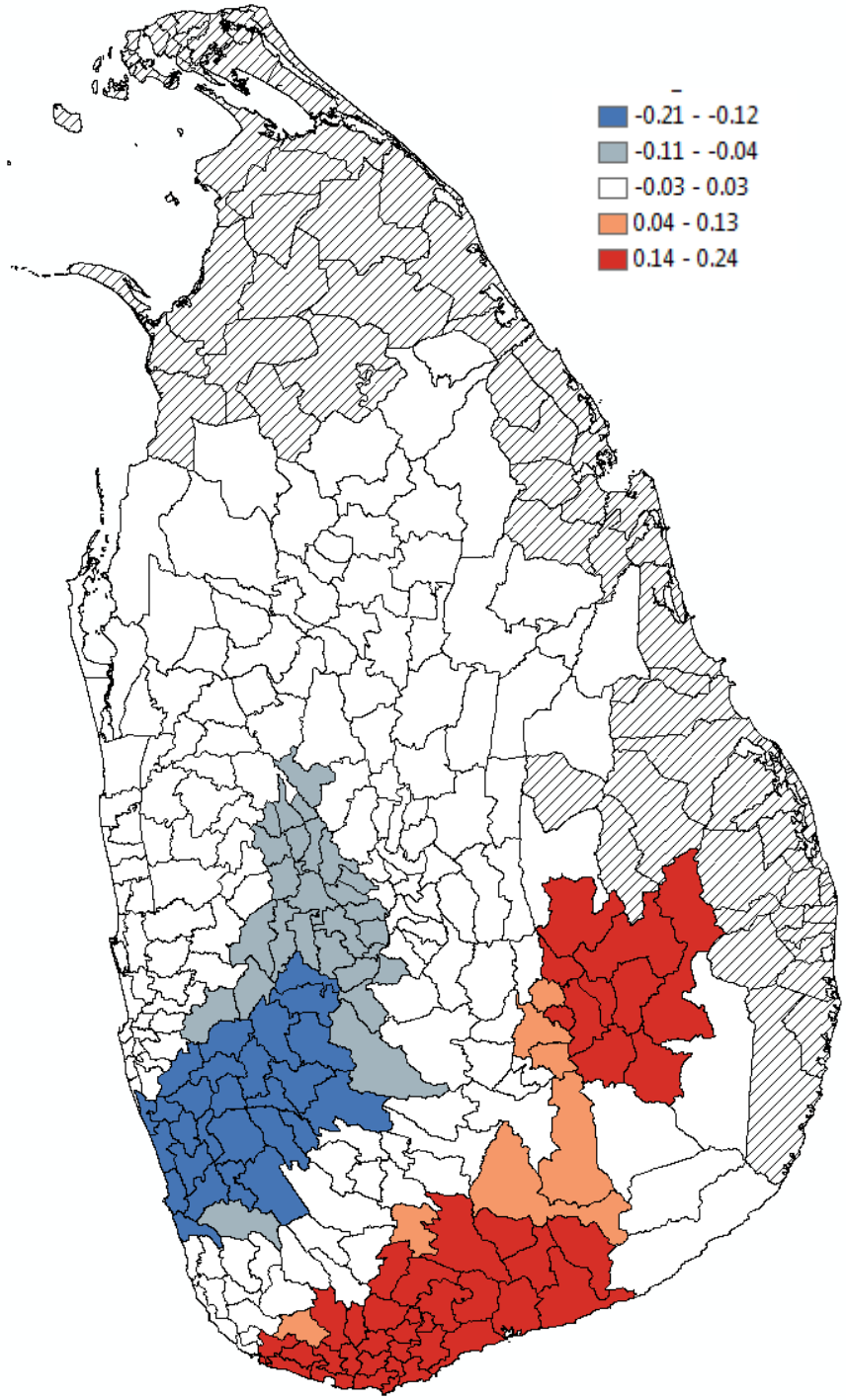
Examine the coefficient raster surfaces created by GWR (and with polygon data, a graduated color rendering of the feature-level coefficients) to better understand regional variation in the model explanatory variables. You can use your understanding of this variation to inform policy:

- Statistically significant global variables that exhibit little regional variation inform regionwide policy.
- Statistically significant global variables that exhibit strong regional variation inform local policy.
- Some variables may not be globally significant, because in some regions they are positively related, and in others they are negatively related.

VISUALIZE

Map 2 shows the spatial distribution of one of the explanatory variables, *DISTOWN*, for districts that have significant coefficients. To do this, you need to right-click the *GWR* layer in the table of contents, click *Properties » Symbology* and then *Quantities » Graduated colors*, and calculate the T statistics as shown in the figure below. The map suggests that there is geographic variation in the coefficients.





Map 2: Coefficient variation for distance to towns by district

Based on the coefficients, it appears that districts in the southern districts of Matara, Hambantota, and parts of Galle and Ratnapura (colored red) have a higher percentage of households in poverty when they are close to towns. However, districts in southwestern Sri Lanka, such as Kalutara, Colombo, and Kegalle (colored blue), have lower poverty when they are close to towns. Note that these maps mask locations with insignificant T statistics at the 10 percent level of significance (that is, the absolute value of the coefficient value/standard error is higher than 1.64). Compare this to

the global OLS coefficient, which is positive for all districts no matter where they are located, suggesting that locations close to towns everywhere have higher poverty. This would imply that proximity to towns has a very different effect depending on where the district is located within the country.

Question 5: *Attach the coefficient maps for some interesting coefficients and discuss the trends you see for each of them. Suggest how this variation over the region you are examining may affect your interpretation of the regression. Also include some discussion of coefficient standard error values in comparison to the actual coefficient values. Interpret the maps of the predicted values you estimated in conjunction with the standardized residuals map. What do they suggest about the poverty patterns in Sri Lanka?*

Submit your work

Submit answers to questions 1–5 along with relevant screen captures. Describe how these maps and statistics could inform economic policy.

Credits

Data

Data in this activity used and displayed in images under license from International Water Management Institute (IWMI), `http://www.iwmi.org`

Instructor resources

Additional information

This lab should be a follow-up to a basic spatial statistics lab and a supplement to any discussion of regression in a spatial context. You might want to discuss the implications of different bandwidths as they might affect the GWR results and tie it to the use of bandwidths in kernel density estimation (see the spatial statistics lab exercise). Questions that pertain to policy, such as, What does the varying coefficient mean for a better understanding of poverty for policy? are especially important for improving students' understanding of the methods. Data for Ecuador (available at the poverty mapping website listed below) could be used to conduct similar types of analyses.

Answers to questions

Question 1: *Identify the names of the locations that had the highest and lowest percentages of poor households. Do they appear to be clustered?*

Answer:

ADM3NAME	PCTPOOHH	POVGROUP	NUMPOHH
Thimbirigasyaya	1.001	1	680
Colombo	1.101	1	953
Dehiwala-Mount Lavinia	1.153	1	621
Sri Jayawardanapura Kotte	1.322	1	367
Moratuwa	7.183	2	3465
Kolonnawa	8.619	2	3243
Maharagama	9.688	2	4463
Kesbewa	10.059	2	4624
Kaduwela	10.51	2	3922
Homagama	10.873	2	4644
Negombo	11.047	2	3842
Hanwella	11.057	2	2832
Kelaniya	11.722	2	3774
Wattala	12.036	2	4039
Katana	12.317	2	6863
Galle Four Gravets	12.593	2	2207
Gampaha	12.675	2	5973
Padukka	12.844	2	1571
Ja-Ela	13.55	2	4445
Mahara	14.453	2	5457
Biyagama	14.459	2	3232
Minuwangoda	14.705	2	5647

Figure 1: Lowest percentage of poor households

ADM3NAME	PCTPOOHH
Kahawatta	38.002
Godakawela	38.058
Thissamaharama	38.267
Eheliyagoda	38.448
Weeraketiya	38.719
Imbulpe	38.944
Angunakolapelessa	39.375
Ayagama	39.539
Elapatha	39.7
Okewela	39.747
Katuwana	40.089
Sooriyawewa	40.124
Kolonna	40.302
Soranathota	40.567
Haldummulla	40.873
Kalawana	40.985
Weligepola	41.307
Opanayaka	41.332
Mahiyanganaya	43.558
Meegahakivula	44.061
Kandaketiya	44.389
Rideemaliyadda	45.65

Figure 2: Highest percentage of poor households

Question 2: *Report the Moran's I for a few variables of interest.*

Answer: The percentage of poor households was highly clustered in Sri Lanka. So were ASSWMIN3 and ASSWMAJ3 (paddy area under irrigation, either major or minor). So were the distances to towns and roads. All showed significant and high positive autocorrelation based on the Moran's I.

coefficients_SriLanka_OLS1

OID	Field1	Variable	Coef	StdError	t_Stat	Prob	Robust_SE	Robust_t	Robust_Pr
0	0	Intercept	14.696919	4.610992	3.187366	0.001639	5.344666	2.749829	0.006419
1	0	ASSWMAJ3	-0.000322	0.000113	-2.840593	0.004895	0.000123	-2.628643	0.009125
2	0	ASSWMIN3	-0.000448	0.000241	-1.856373	0.064638	0.000189	-2.37042	0.018554
3	0	DISROADS	-0.310127	0.342001	-0.906803	0.365416	0.302136	-1.02645	0.305713
4	0	DISTOWN	0.079159	0.045514	1.739228	0.083294	0.043254	1.830103	0.068488
5	0	MAHARF	-0.021496	0.004775	-4.501778	0.000013	0.004714	-4.56043	0.00001
6	0	YALARF	0.012102	0.005682	2.129844	0.034201	0.005566	2.17422	0.030663
7	0	SMHOLDPA	0.085394	0.087854	0.971997	0.332028	0.092158	0.926601	0.355061
8	0	AGOPWOLA	5.359811	5.825823	0.920009	0.358488	6.058761	0.884638	0.377233
9	0	PACSMALL	-0.111693	0.198094	-0.563836	0.573402	0.182872	-0.610769	0.541938
10	0	NUMAGHH	19.04301	2.193115	8.683089	0	2.386349	7.979975	0

Diagnostics_SriLankaOLS_Model1

OID	Field1	Diag_Name	Diag_Value	Definition
0	0	AIC	1614.111868	Akaike's Information Criterion: A relative measure of performance used to compare models; the smaller AIC indicates the superior model.
1	0	AICc	1615.433902	Corrected Akaike's Information Criterion: second order correction for small sample sizes.
2	0	R2	0.50612	R-Squared, Coefficient of Determination: The proportion of variation in the dependent variable that is explained by the model.
3	0	AdjR2	0.485369	Adjusted R-Squared: R-Squared adjusted for model complexity (number of variables) as it relates to the data.
4	0	F-Stat	24.389872	Joint F-Statistic Value: Used to assess overall model significance.
5	0	F-Prob	0	Joint F-Statistic Probability (p-value): The probability that none of the explanatory variables have an effect on the dependent variable.
6	0	Wald	266.788894	Wald Statistic: Used to assess overall robust model significance.
7	0	Wald-Prob	0	Wald Statistic Probability (p-value): The computed probability, using robust standard errors, that none of the explanatory variables have an effect on the dependent variable.
8	0	K(BP)	41.363093	Koenker's studentized Breusch-Pagan Statistic: Used to test the reliability of standard error values when heteroskedasticity (non-constant variance) is present.
9	0	K(BP)-Prob	0.00001	Koenker (BP) Statistic Probability (p-value): The probability that heteroskedasticity (non-constant variance) has not made standard errors unreliable.
10	0	JB	5.275213	Jarque-Bera Statistic: Used to determine whether the residuals deviate from a normal distribution.
11	0	JB-Prob	0.071532	Jarque-Bera Probability (p-value): The probability that the residuals are normally distributed.
12	0	Sigma2	36.353722	Sigma-Squared: OLS estimate of the variance of the error term.

Figure 3: Results of OLS regression in ArcGIS

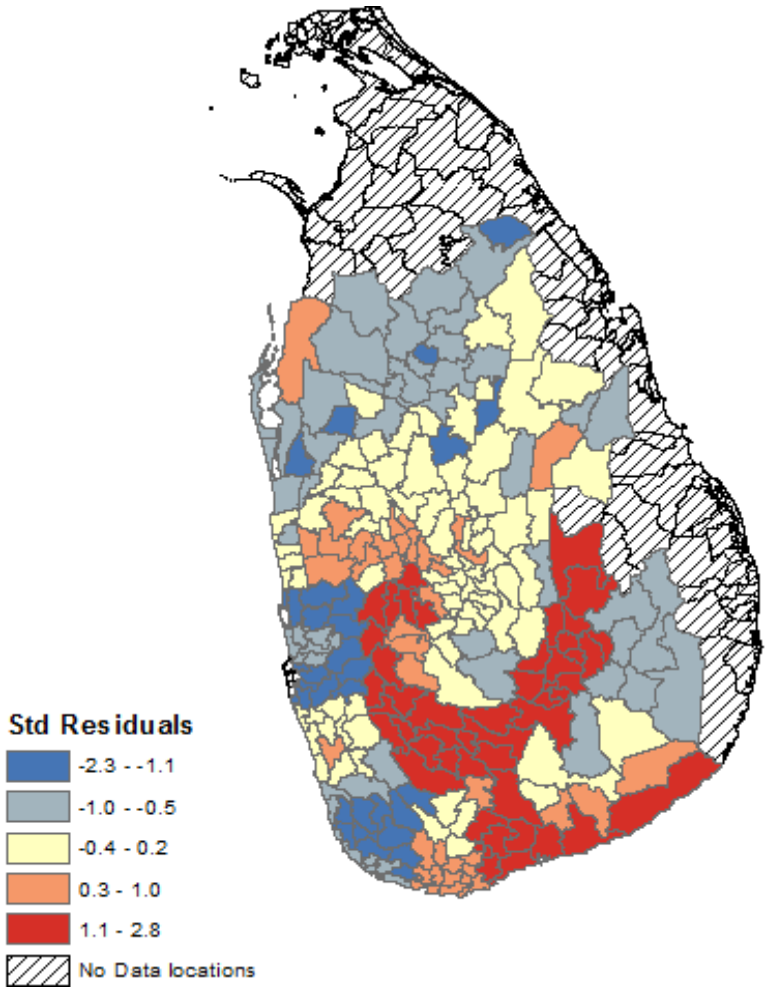
Question 3: *Look at the fitted OLS regression of the percentage of poor households using the explanatory variables shown. What do the results mean? How would you interpret the coefficients of the significant variables? Also, interpret the diagnostics. For example, the statistically significant Koenker diagnostic test means that there may be nonstationarity in your model that GWR can account for. It is therefore appropriate to use a GWR model here.*

Answer: The student should note that ASSWMAJ3, ASSWMIN3, DISTOWN, MAHARF, and YALARF were all significant at the 10 percent p values. The signs should be interpreted. For example, the negative sign on ASSWMAJ3 and ASSWMIN3 suggests that as more area is allocated to paddy, there is less poverty, all other aspects remaining the same, but on the other hand, the positive sign on DISTOWN suggests that as distance to towns increases, there is more poverty, all else remaining the same. Distance to roads, DISROADS, is not significant in affecting poverty. The positive sign on MAHARF suggests that higher rainfall leads to lower levels of poverty, all else being equal.

The R^2 suggests that about 51 percent of the variance is explained by this model, and the adjusted R^2, which accounts for the number of variables used, drops slightly to 48 percent. The FStat is significant.

Question 4: *Map the residuals. Do they appear to be clustered? What is the Moran's I for the residuals?*

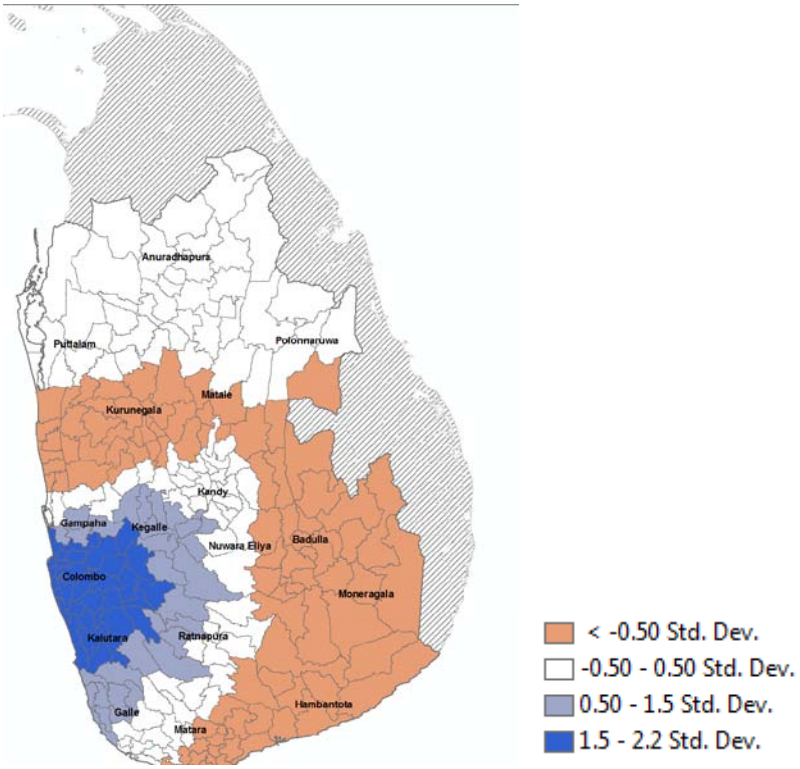
Answer: The residuals showed positive spatial autocorrelation based on the global Moran's I, which was high and significant (about 0.6). One way to deal with this problem is to run a GWR to see if the effects of the coefficients are different over the study area. Also see `www.esri.com/news/arcuser/0111/findmodel.html` for more explanations on interpreting and refining your results.



Map 3: Standardized residuals

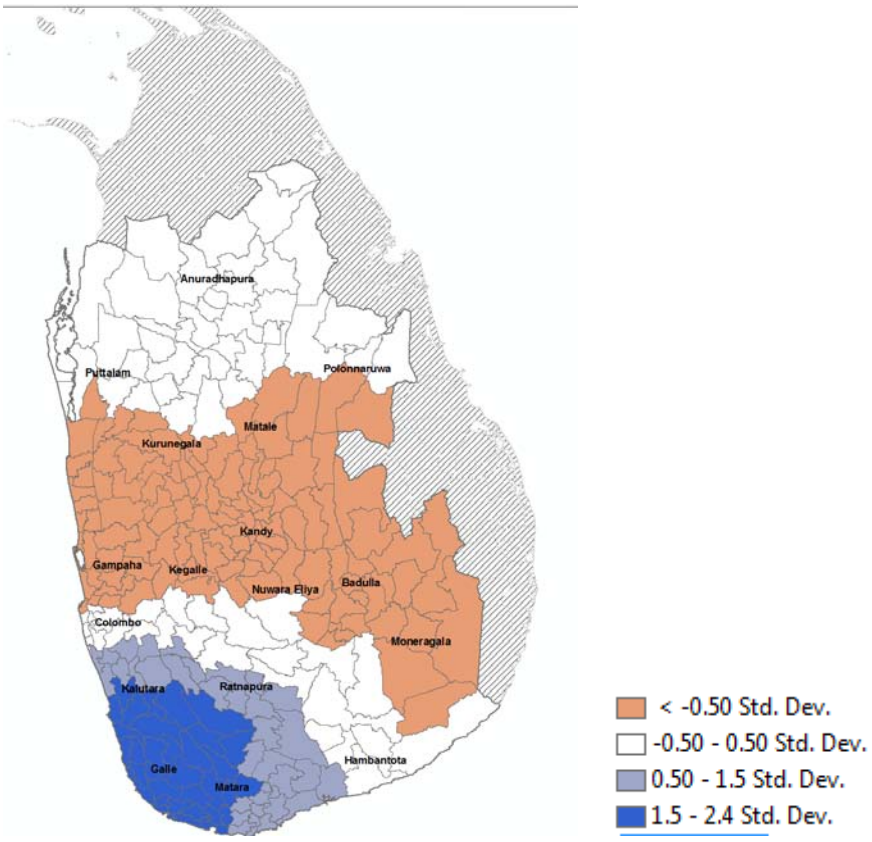
Question 5: *Attach the coefficient maps for some interesting coefficients and discuss the trends you see for each of them. Suggest how this variation over the region you are examining may affect your interpretation of the regression. Also include some discussion of coefficient standard error values in comparison to the actual coefficient values. Interpret the maps of the predicted values you estimated in conjunction with the standardized residuals map. What do they suggest about the poverty patterns in Sri Lanka?*

Answer: The lab discusses distance to towns, which shows variation in the sign of the coefficient depending on location in either south or southwest, unlike the OLS, which suggests that the farther the location is from a town, the higher the poverty percentage.



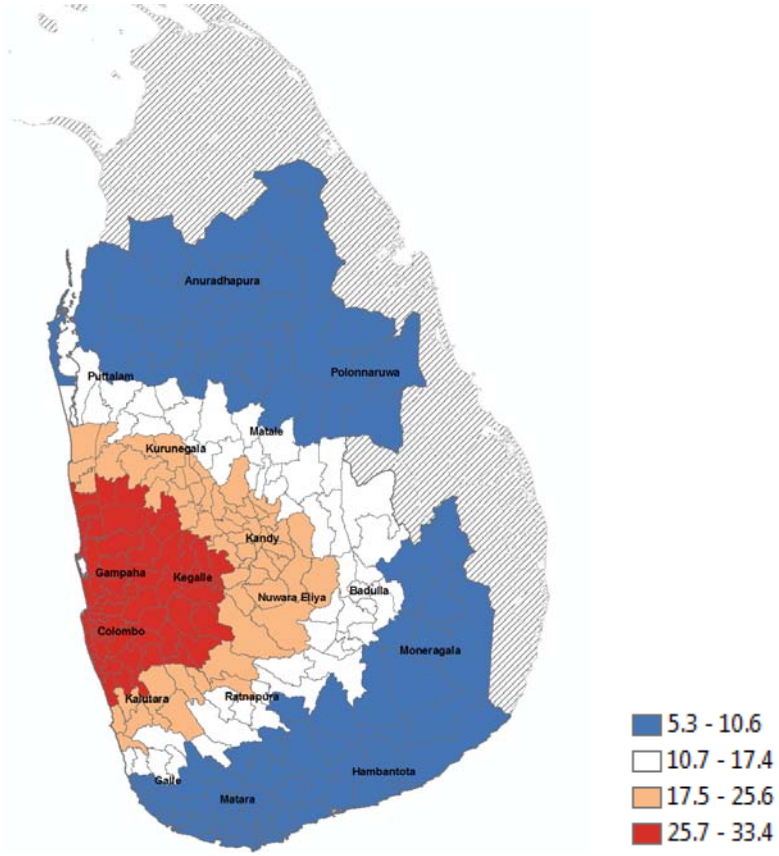
Map 4: GWR coefficients for distance to roads mapped by district

This figure maps the coefficient for the distance to roads. Again, you can see different effects depending on the location. In many locations in the Hambantota, Moneragala, Badulla, Matale, and Kurunegala districts, as distance to roads increases, there is a lower percentage of households in poverty, though the effect is small. However, closer to Colombo and Kalutara, the longer distance from roads would mean a higher number of households living in poverty.



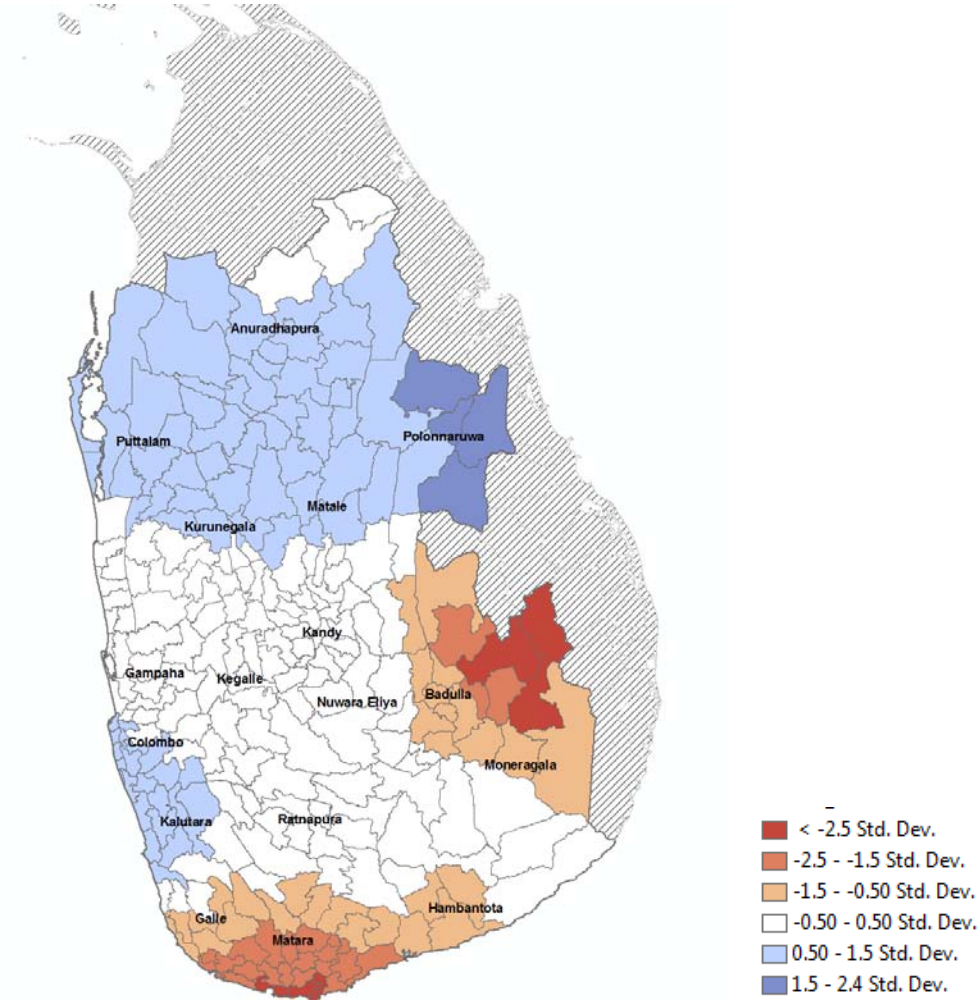
Map 5: GWR coefficients for paddy area under minor irrigation mapped by district

In the case of paddy area under minor irrigation, in locations in the Galle, Kalutara, Matara, and Ratnapura districts, the effect of more land devoted to paddy agriculture slightly increases the percentage of households in poverty. In the orange shaded districts in central Sri Lanka, including Kandy and Kegalle, there is an opposite, though still small, effect. Note that coefficient values vary from -0.001 to 0.006.



Map 6: GWR coefficient for percentage of agricultural households by district

The effect of agricultural households is always positive in that, as the percentage of agricultural households in a location increases, the percentage of households in poverty also increases, all else being equal. However, the effect is stronger in Colombo, Kegalle, and Gampaha compared to the northern and southern districts of Anuradhapura, Polonnaruwa, Moneragala, Galle, Matara, and Hambantota, as the numbers indicate.



Map 7: GWR coefficient for MAHARF by district

The map showing coefficients of the MAHARF indicates that the effect of rainfall in the Maha season (winter) has different effects depending on location. In Badulla, Moneragala, Matara, Galle, and Hambantota, the higher the rainfall, the lower the percentage of households in poverty, all else remaining the same. However, in the north in Anuradhapura, Matale, Polonnaruwa, and Puttalam and the southwest near Colombo and Kalutara, increased rainfall means a higher percentage of poverty.

Compare these coefficient maps to the OLS results. In the OLS results, an irrigated paddy cultivation area had a slight negative overall effect on poverty, but the coefficient could vary depending on where the cultivation occurred. In fact, the effect of increasing the area was positive (increased percentage of households in poverty) in the south. Proximity to roads did not appear to have an effect on poverty in the OLS model. The GWR estimate, however, found that there is significant spatial variation in this coefficient and that it could lead to lower poverty in parts of Sri Lanka. Likewise, the rainfall in the Maha season and the percentage of agricultural households also appear to show different effects on poverty depending on where in Sri Lanka they are located.

Thus, a poverty reduction policy that rejected urbanization because it was correlated with higher percentages of households in poverty would be misguided, because GWR suggests that the spatial

variation of poverty depending on distance to towns is considerable. Some locations benefit from being close to towns, whereas others do not. Likewise, increasing the area of paddy cultivation because the OLS results suggest that this decreased the percentage of households in poverty would be misguided in the southern districts of Galle, Matara, and Kalutara.

References and further reading

Amarasinghe, Upali, Madar Samad, and Markandu Anputhas. 2005. "Spatial Clustering of Rural Poverty and Food Insecurity in Sri Lanka." *Food Policy* 30 (5–6): 493–509.

Casettil, Emilio. 1997. "The Expansion Method, Mathematical Modeling, and Spatial Econometrics." *International Regional Science Review* 20: 9–33.

Fotheringham, A. Stewart, Chris Brunsdon, and Martin Charlton. 2002. *Geographically Weighted Regression: The analysis of spatially varying relationships.* Chichester, England: John Wiley.

Fox, John. 2002. *An R and S-Plus Companion to Applied Regression.* Thousand Oaks, CA: Sage Publications.
http://www.spatialfiltering.com/Workshop/RSupportPage.htm

Getis, Arthur. 2010. "Spatial Filtering in a Regression Framework: Examples using data on urban crime, regional inequality, and government expenditures." In *Perspectives on Spatial Data Analysis*. Berlin: Springer.

Harris, P., A. Stewart Fotheringham, R. Crespo, and Martin Charlton. 2010. "The Use of Geographically Weighted Regression for Spatial Prediction: An evaluation of models using simulated data sets." *Mathematical Geosciences* 42: 657–680.

Klieber, Christian, and Achim Zeileis. 2008. *Applied Econometrics with R*. New York: Springer.

Some useful web links on GWR

http://eprints.ncrm.ac.uk/90/1/MethodsReviewPaperNCRM-006.pdf

http://ncg.nuim.ie/ncg/GWR/GWR_Tutorial.pdf

Data information

Sri Lanka poverty data is documented and available at
http://gisweb.ciat.cgiar.org/povertymapping/. (Registration is required but free.)

Metadata for this data is found at
http://gisweb.ciat.cgiar.org/povertymapping/metadata/srilankacasestudy_faq.htm.