Esri International User Conference | San Diego, CA

Technical Workshops | July 2011

Modeling Spatial Relationships using Regression Analysis

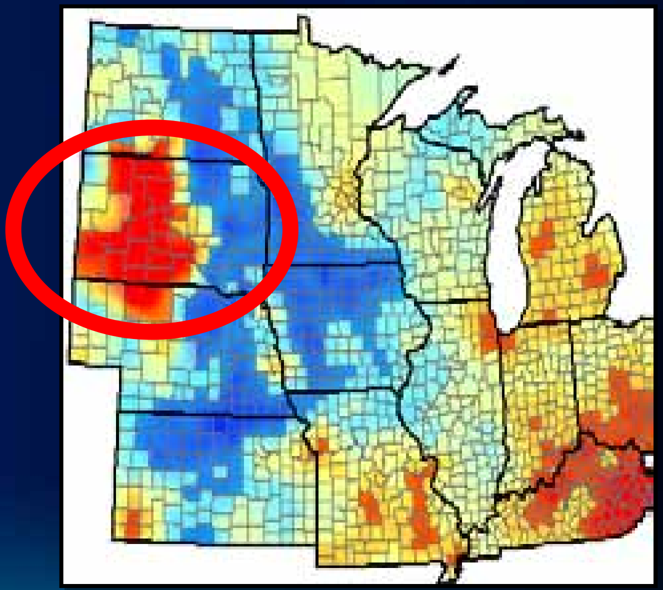
Lauren M. Scott, PhD

Lauren Rosenshein, MS

Mark V. Janikas, PhD

Answering “Why?” Questions

- **Introduction to Regression Analysis**
- Why are people dying young in South Dakota?
  - **Building a properly specified OLS model**
    - **The 6 things you must check!**
  - **Exploring regional variation using GWR**

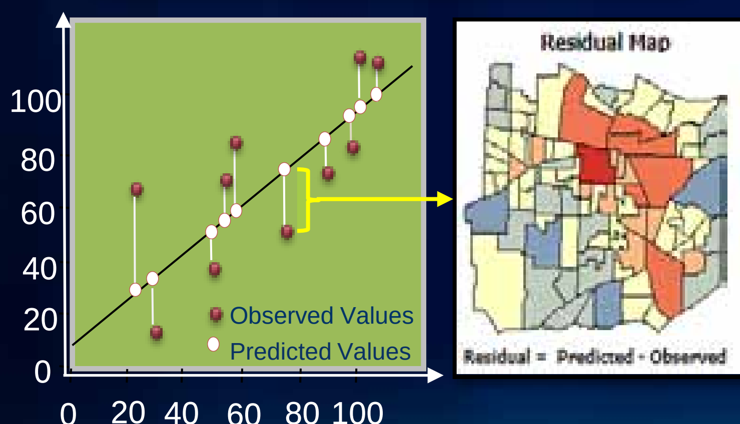


Kindly complete a course evaluation:
www.esri.com/sessionevals

What can you do with Regression analysis?

- Model, examine, and explore spatial relationships
- Better understand the factors behind observed spatial patterns
- Predict outcomes based on that understanding

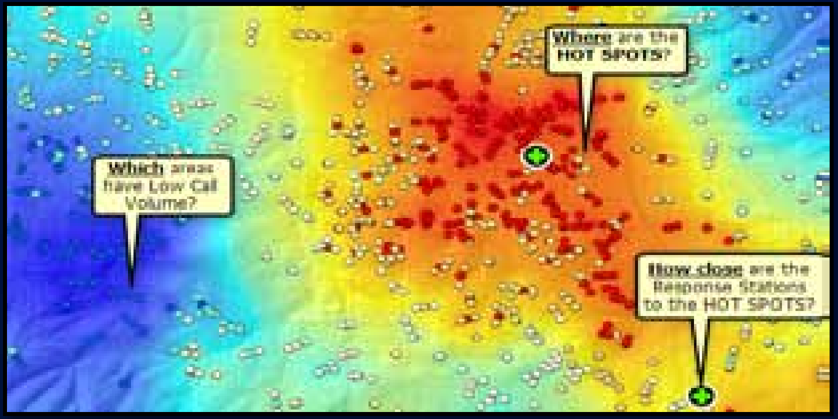
Ordinary Least Square (OLS)



Geographically Weighted Regression (GWR)

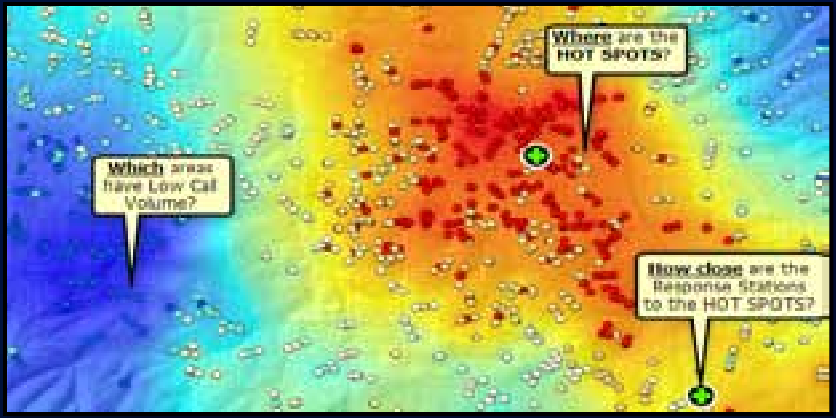
What’s the big deal?

- Pattern analysis (without regression):
  - Are there places where people persistently die young?
  - Where are test scores consistently high?
  - Where are 911 emergency call hot spots?



What's the big deal?

- Pattern analysis (without regression):
  - Are there places where people persistently die young?
  - Where are test scores consistently high?
  - Where are 911 emergency call hot spots?

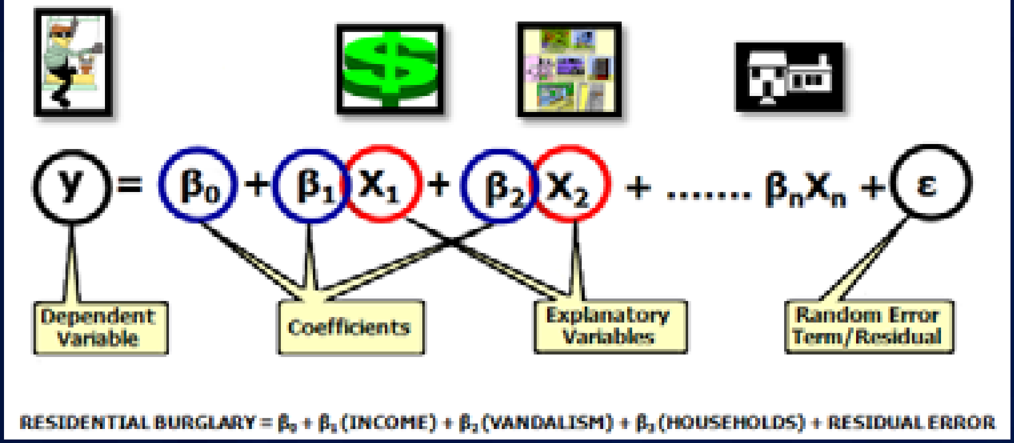


- Regression analysis:
  - Why are people persistently dying young?
  - What factors contribute to consistently high test scores?
  - Which variables effectively predict 911 emergency call volumes?

Why use regression?

- Understand key factors
  - What are the most important habitat characteristics for an endangered bird?
- Predict unknown values
  - How much rainfall will occur in a given location?
- Test hypotheses
  - “Broken Window” Theory: Is there a positive relationship between vandalism and residential burglary?

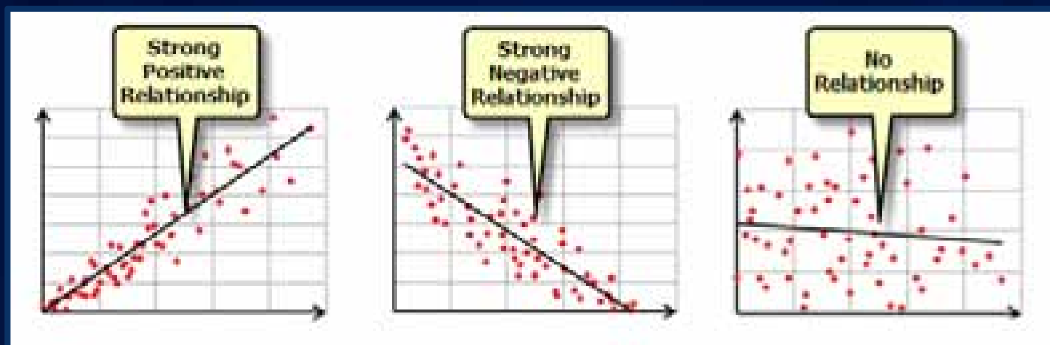
Regression analysis terms and concepts



- *Dependent variable* (Y): What you are trying to model or predict (e.g., residential burglary).
- *Explanatory variables* (X): Variables you believe cause or explain the dependent variable (e.g., income, vandalism, number of households).
- *Coefficients* (β): Values, computed by the regression tool, reflecting the relationship between explanatory variables and the dependent variable.
- *Residuals* (ε): The portion of the dependent variable that isn't explained by the model; the model under- and over-predictions.

Regression model coefficients

- **Coefficient sign (+/-) and magnitude reflect each explanatory variable's relationship to the dependent variable**



Intercept	1.625506
INCOME	-0.000030
VANDALISM	0.133712
HOUSEHOLDS	0.012425
LOWER CITY	0.136569

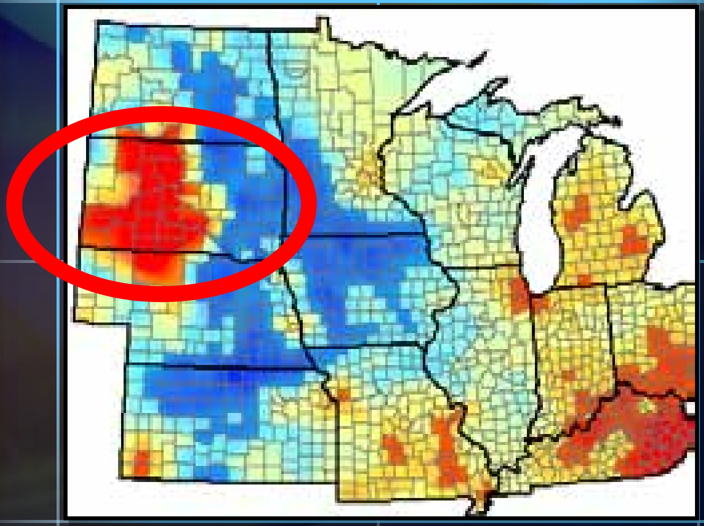
The asterisk * indicates the explanatory variable is statistically significant

Summary of OLS Results

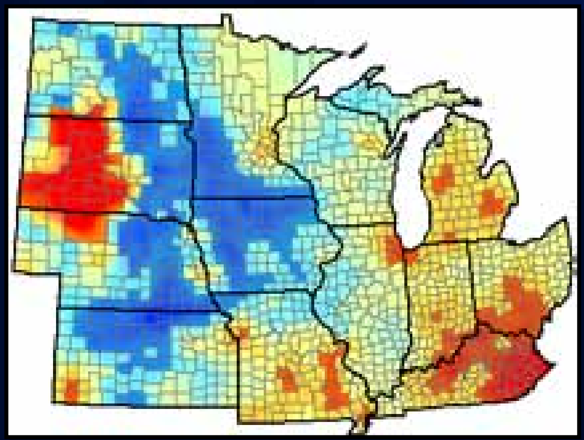
Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	1.625506	0.132184	12.297312	0.000000*	0.141894	11.455814	0.000000*	--------
INCOME	-0.000030	0.000003	-9.714560	0.000000*	0.000003	-10.200131	0.000000*	1.312936
VANDALISM	0.133712	0.007818	17.103675	0.000000*	0.013133	10.181417	0.000000*	1.306447
HOUSEHOLD	0.012425	0.001213	10.245426	0.000000*	0.002155	5.764924	0.000000*	1.315309
LOWERCITY	0.136569	0.122861	1.111567	0.266500	0.112888	1.209771	0.226562	1.318318

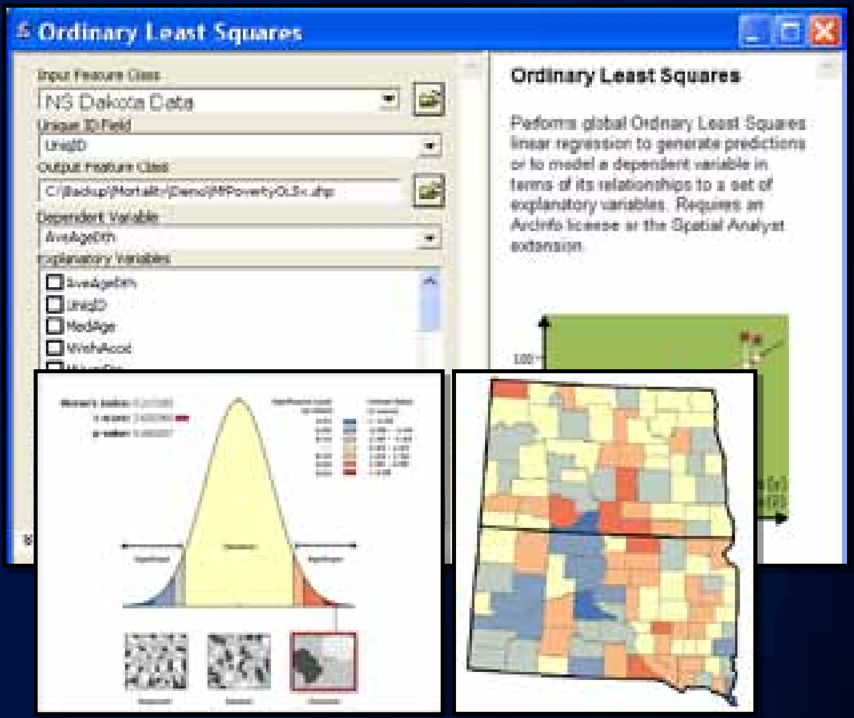
OLS Regression

Why are people dying young in South Dakota?



OLS analysis





Why are people dying young in South Dakota?

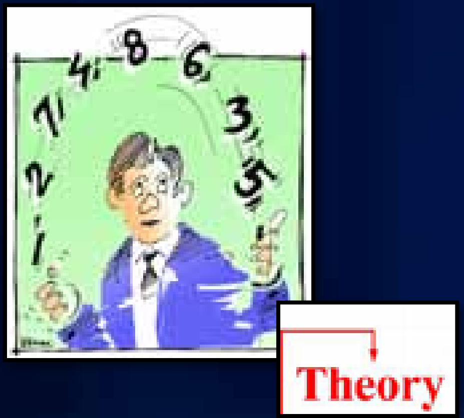
Do economic factors explain this spatial pattern?

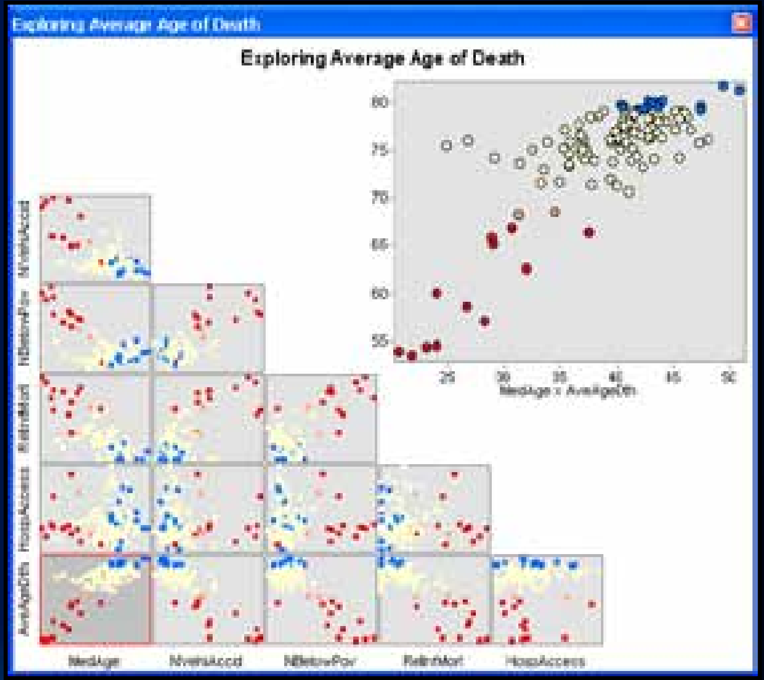
Poverty rates explain 66% of the variation in the average age of death dependent variable: Adjusted R-Squared [2]: 0.659

However, significant spatial autocorrelation among model residuals indicates important explanatory variables are missing from the model.

Build a multivariate regression model

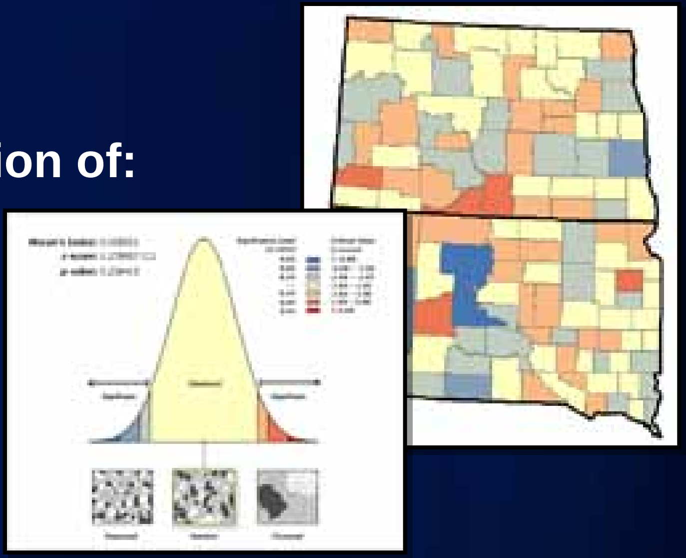
- Explore variable relationships using the scatterplot matrix
- Consult theory and field experts
- Look for spatial variables
- Run OLS (this is iterative)
- Use Exploratory Regression





Our best OLS model

- Average Age of Death as a function of:
  - Poverty Rates
  - Vehicle Accidents
  - Lung Cancer
  - Suicide
  - Diabetes
- This model tells 86% of the story… and the over and under predictions aren't clustered!



But are we done?

Check OLS results

1 Coefficients have the expected sign.

2 No redundancy among explanatory variables.

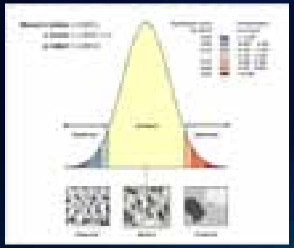
3 Coefficients are statistically significant.

Summary of OLS Results

Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	86.082979	0.875151	98.363521	0.000000*	0.813152	105.863324	0.000000*	--------
NVEHIACCID	-110.520016	12.213013	-9.049366	0.000000*	14.544464	-7.598769	0.000000*	2.351229
NSUICIDE	-138.221155	18.180324	-7.602788	0.000000*	29.800993	-4.638139	0.000011*	1.556498
NLUNGCANC	-47.045741	12.076316	-3.895703	0.000172*	13.536130	-3.475568	0.000732*	1.051207
NDIABETES	-33.429850	13.805975	-2.421405	0.017044*	14.732174	-2.269173	0.025148*	1.400358
NBELOWPOV	-14.408804	3.633873	-3.965137	0.000134*	4.125643	-3.492499	0.000692*	3.232363

OLS Diagnostics

Input Features:	DakotaData.shp	Dependent Variable:	AVEAGEDTH
Number of Observations:	119	Akaike's Information Criterion (AICc) [2]:	527.985211
Multiple R-Squared [2]:	0.870551	Adjusted R-Squared [2]:	0.864823
Joint F-Statistic [3]:	151.985705	Prob(>F), (5,113) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	496.057428	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [5]:	21.590491	Prob(>chi-squared), (5) degrees of freedom:	0.000626*
Jarque-Bera Statistic [6]:	4.207198	Prob(>chi-squared), (2) degrees of freedom:	0.122017



4 Residuals are normally distributed.

5 Residuals are not spatially autocorrelated.

6 Strong Adjusted R-Square value.

1. Coefficient signs

- Coefficients should have the expected signs.

Notes on Interpretation

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

Summary of OLS Results

Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	86.082979	0.875151	98.363521	0.000000*	0.813152	105.863324	0.000000*	--------
NVEHIACCID	-110.520016	12.213013	-9.049366	0.000000*	14.544464	-7.598769	0.000000*	2.351229
NSUICIDE	-138.221155	18.180324	-7.602788	0.000000*	29.800993	-4.638139	0.000011*	1.556498
NLUNGCANC	-47.045741	12.076316	-3.895703	0.000172*	13.536130	-3.475568	0.000732*	1.051207
NDIABETES	-33.429850	13.805975	-2.421405	0.017044*	14.732174	-2.269173	0.025148*	1.400358
NBELOWPOV	-14.408804	3.633873	-3.965137	0.000134*	4.125643	-3.492499	0.000692*	3.232363

OLS Diagnostics

Input Features:	DakotaData.shp	Dependent Variable:	AVEAGEDTH
Number of Observations:	119	Akaike's Information Criterion (AICc) [2]:	527.985211
Multiple R-Squared [2]:	0.870551	Adjusted R-Squared [2]:	0.864823
Joint F-Statistic [3]:	151.985705	Prob(>F), (5,113) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	496.057428	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [5]:	21.590491	Prob(>chi-squared), (5) degrees of freedom:	0.000626*
Jarque-Bera Statistic [6]:	4.207198	Prob(>chi-squared), (2) degrees of freedom:	0.122017

2. Coefficient significance

- Look for statistically significant explanatory variables.

Probability
0.000000 *
0.000000 *
0.000000 *
0.000172 *
0.017044 *
0.000134 *

Robust_Prob
0.000000 *
0.000000 *
0.000011 *
0.000732 *
0.025148 *
0.000692 *

Notes on Interpretation

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

* Statistically significant at the 0.05 level

Summary of OLS Results

Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	86.082979	0.875151	98.363521	0.000000*	0.81315[illegible]	[illegible].863324	0.000000*	--------
NVEHIACCID	-110.520016	12.213013	-9.049366	0.000000*	14.5444[illegible]	[illegible].598769	0.000000*	2.351229
NSUICIDE	-138.221155	18.180324	-7.602788	0.000000*	29.8009[illegible]	[illegible].638139	0.000011*	1.556498
NLUNGCANC	-47.045741	12.076316	-3.895703	0.000172*	13.536130	-3.475568	0.000732*	1.051207
NDIABETES	-33.429850	13.805975	-2.421405	0.017044*	14.732174	-2.269173	0.025148*	1.400358
NBELOWPOV	-14.408804	3.633873	-3.965137	0.000134*	4.125643	-3.492499	0.000692*	3.232363

OLS Diagnostics

Input Features:	DakotaData.shp	Dependent Variable:	AVEAGEDTH
Number of Observations:	119	Akaike's Information Criterion (AICc) [2]:	527.985211
Multiple R-Squared [2]:	0.870551	Adjusted R-Squared [2]:	0.864823
Joint F-Statistic [3]:	151.985705	Prob(>F), (5,113) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	496.057428	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [5]:	21.590491	Prob(>chi-squared), (5) degrees of freedom:	0.000626*
Jarque-Bera Statistic [6]:	[illegible]	Prob(>chi-squared), (2) degrees of freedom:	[illegible]

Koenker(BP) Statistic [5]: 38.994033 Prob(>chi-squared) (5) degrees of freedom: 0.000626 *

Check for variable redundancy

- Multicollinearity:
  - Term used to describe the phenomenon when two or more of the variables in your model are highly correlated.
- Variance inflation factor (VIF):
  - Detects the severity of multicollinearity.
  - Explanatory variables with a VIF greater than 7.5 should be removed one by one.

3. Multicollinearity

- **Find a set of explanatory variables that have low VIF values.**
- **In a strong model, each explanatory variable gets at a different facet of the dependent variable.**

Notes on Interpretation

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

[1] Large VIF (> 7.5, for example) indicates explanatory variable redundancy.

VIF

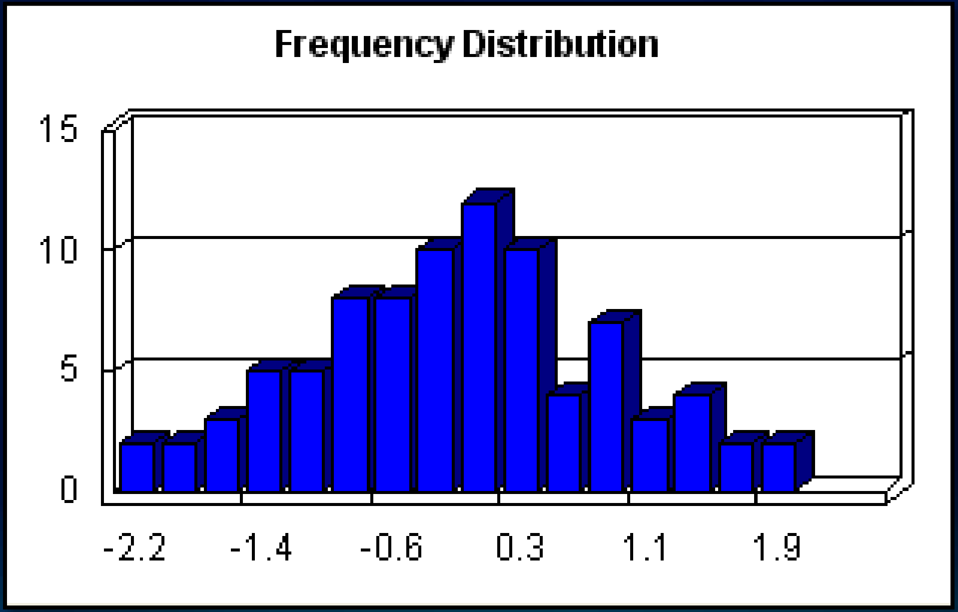
2.351229
1.556498
1.051207
1.400358
3.232363

Summary of OLS Results

Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	86.082979	0.875151	98.363521	0.000000*	0.813152	105.863324	0.000000*	--------
NVEHIACCID	-110.520016	12.213013	-9.049366	0.000000*	14.544464	-7.598769	0.000000*	2.351229
NSUICIDE	-138.221155	18.180324	-7.602788	0.000000*	29.800993	-4.638139	0.000011*	1.556498
NLUNGCANC	-47.045741	12.076316	-3.895703	0.000172*	13.536130	-3.475568	0.000732*	1.051207
NDIABETES	-33.429850	13.805975	-2.421405	0.017044*	14.732174	-2.269173	0.025148*	1.400358
NBELOWPOV	-14.408804	3.633873	-3.965137	0.000134*	4.125643	-3.492499	0.000692*	3.232363

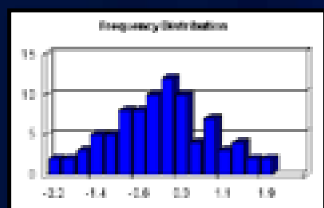
Checking for model bias

- The residuals of a good model should be normally distributed with a mean of zero
- The Jarque-Bera test checks model bias



4. Model bias

- **When the Jarque-Bera test is statistically significant:**
  - **The model is biased**
  - **Results are *not* reliable**
  - **Often indicates that a key variable is missing from the model**



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

[6] Significant p-value indicates residuals deviate from a normal distribution.

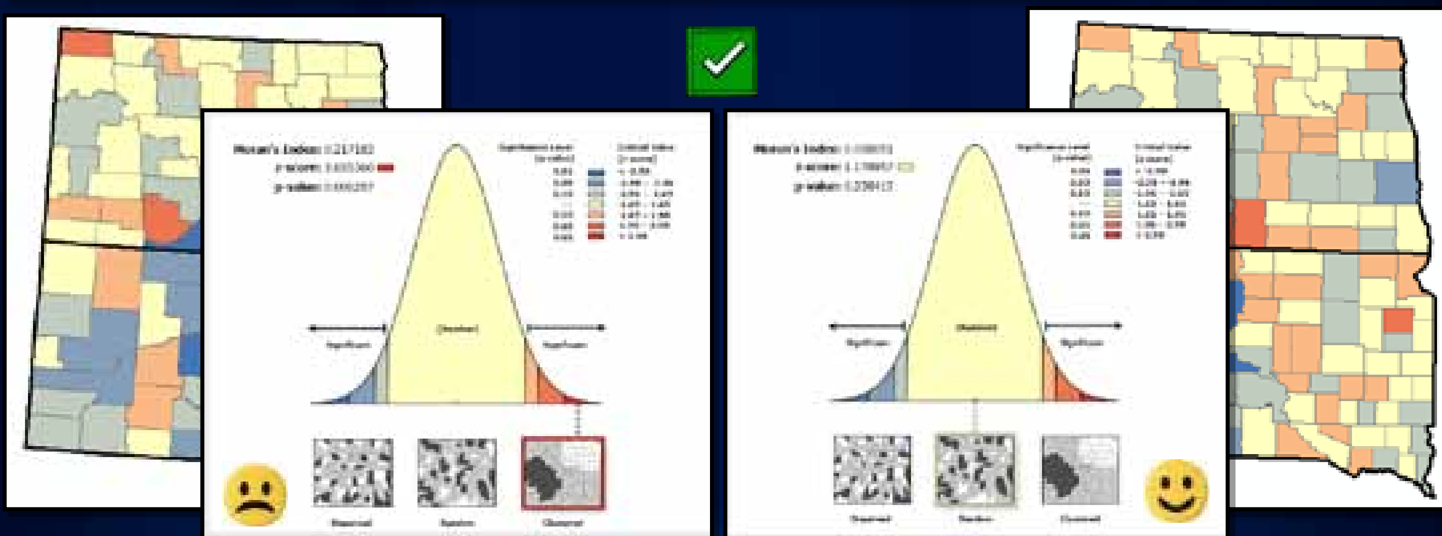
Jaque-Bera Statistic [6]: 4.207198 Prob(>chi-sq), (2) degrees of freedom: 0.122017

```
                                   OLS Diagnostics
Input Features:             DakotaData.shp   Dependent Variable:                          AVEAGEDTH
Number of Observations:                119   Akaike's Information Criterion (AICc) [2]:  527.985211
Multiple R-Squared [2]:           0.870551   Adjusted R-Squared [2]:                       0.864823
Joint F-Statistic [3]:          151.985705   Prob(>F), (5,113) degrees of freedom:         0.000000*
Joint Wald Statistic [4]:       496.057428   Prob(>chi-squared), (5) degrees of freedom:   0.000000*
Koenker (BP) Statistic [5]:      21.590491   Prob(>chi-squared), (5) degrees of freedom:   0.000626*
Jarque-Bera Statistic [6]:        4.207198   Prob(>chi-squared), (2) degrees of freedom:   0.122017
```

5. Spatial Autocorrelation

WARNING 000851: Use the Spatial Autocorrelation (Moran's I) Tool to ensure residuals are not spatially autocorrelated.

Error code:	000851: Use the Spatial Autocorrelation (Moran's I) Tool to ensure residuals are not spatially autocorrelated.
Description:	Results from regression analysis are only trustworthy when the model and data meet the assumptions/limitations of that method. Statistically significant spatial autocorrelation in the regression residuals indicates misspecification (a key missing explanatory variable). Results are invalid when a model is misspecified.
Solution:	Run the Spatial Autocorrelation (Moran's I) tool on the regression residuals in the output feature class. If the Z score indicates spatial autocorrelation is statistically significant, map the residuals and perhaps run hot spot analysis on the residuals to see if the spatial pattern of over and under predictions provides clues about missing key variables from the model. If you cannot identify the key missing variables, results of the regression are invalid and you should consider using a spatial regression method designed to deal with spatial autocorrelation in the error term. When spatial autocorrelation in OLS residuals is due to non-stationary spatial processes, use Geographically Weighted Regression instead of OLS.



Statistically significant clustering of under and over predictions.

Random spatial pattern of under and over predictions.

6. Model performance

- Compare models by looking for the lowest AIC value.
  - As long as the dependent variable remains fixed, the AIC value for different OLS/GWR models are comparable
- Look for a model with a high Adjusted R-Squared value.

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

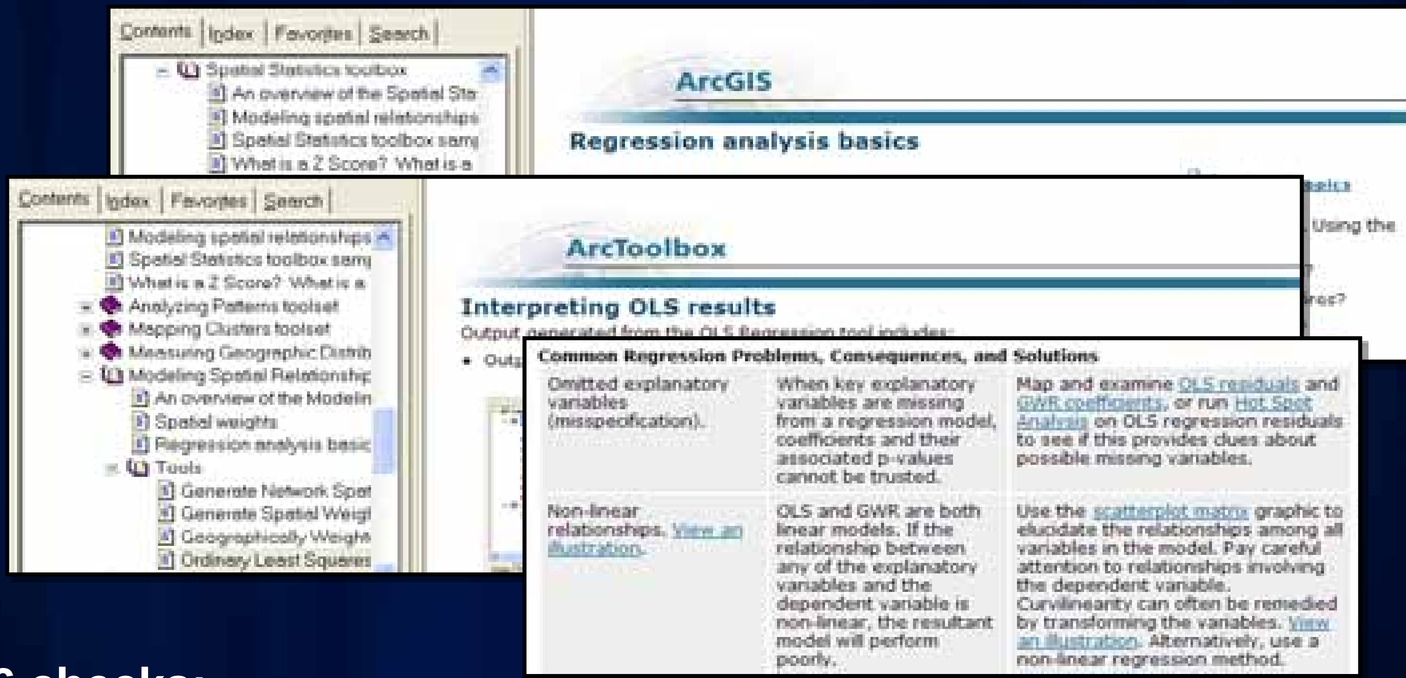
[2] Measure of model fit/performance.

Akaike's Information Criterion (AIC) [2]:	524.9762
Adjusted R-Squared [2]:	0.8648

OLS Diagnostics

Input Features:	DakotaData.shp	Dependent Variable:	AVEAGEDTH
Number of Observations:	119	Akaike's Information Criterion (AICc) [2]:	527.985211
Multiple R-Squared [2]:	0.870551	Adjusted R-Squared [2]:	0.864823
Joint F-Statistic [3]:	151.985705	Prob(>F), (5,113) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	496.057428	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [5]:	21.590491	Prob(>chi-squared), (5) degrees of freedom:	0.000626*
Jarque-Bera Statistic [6]:	4.207198	Prob(>chi-squared), (2) degrees of freedom:	0.122017

Online help is … helpful!



The 6 checks:

ü Coefficients have the expected sign.
ü Coefficients are statistically significant.
ü No redundancy among explanatory variables.
ü Residuals are normally distributed.
ü Residuals are not spatially autocorrelated.
ü Strong Adjusted R-Squared value.

Now are we done?

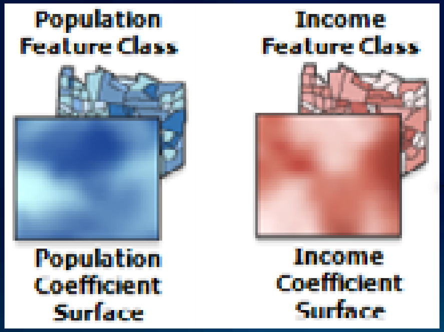
- A statistically significant Koenker OLS diagnostic is often evidence that Geographically Weighted Regresion (GWR) will improve model results
- GWR allows you to explore geographic variation which can help you tailor effective remediation efforts.

OLS Diagnostics

Input Features:	DakotaData.shp	Dependent Variable:	AVEAGEDTH
Number of Observations:	119	Akaike's Information Criterion (AICc) [2]:	527.985211
Multiple R-Squared [2]:	0.870551	Adjusted R-Squared [2]:	0.864823
Joint F-Statistic [3]:	151.985705	Prob(>F), (5,113) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	496.057428	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [5]:	21.590491	Prob(>chi-squared), (5) degrees of freedom:	0.000626*
Jarque-Bera Statistic [6]:	4.207198	Prob(>chi-squared), (2) degrees of freedom:	0.122017

Global vs. local regression models

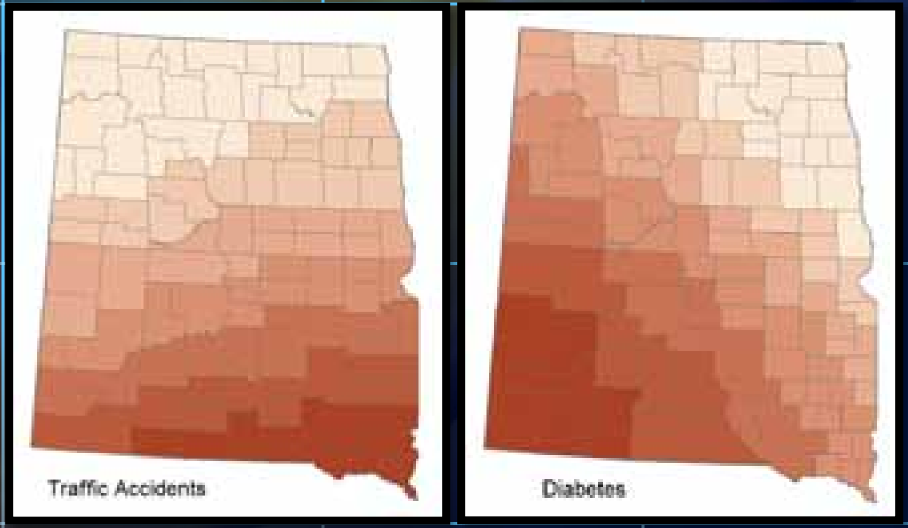
- OLS
  - Global regression model
  - One equation, calibrated using data from all features
  - Relationships are fixed
- GWR
  - Local regression model
  - One equation for every feature, calibrated using data from nearby features
  - Relationships are allowed to vary across the study area



For each explanatory variable, GWR creates a coefficient surface showing you where relationships are strongest.

GWR

Exploring regional variation



Running GWR

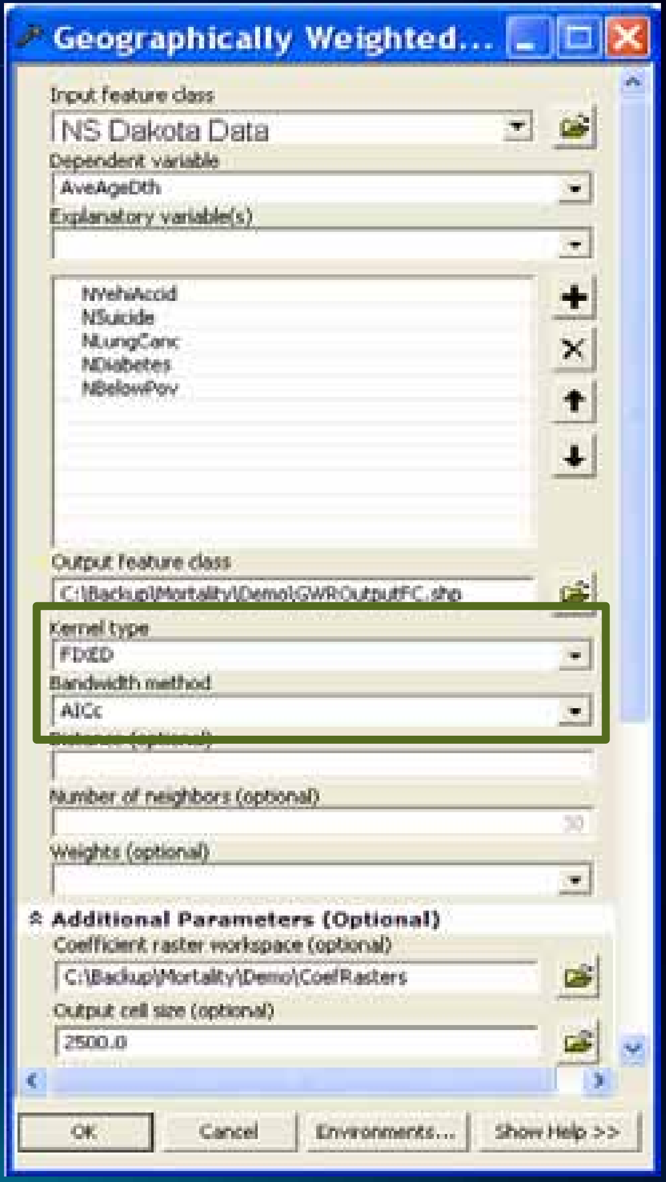
- GWR is a local spatial regression model
  - Modeled relationships are allowed to vary
- GWR variables are the same as OLS, except:
  - Do not include spatial regime (dummy) variables
  - Do not include variables with little value variation

Defining *local*

- GWR constructs an equation for each feature
- Coefficients are estimated using nearby feature values
- GWR requires a definition for *nearby*

Defining *local*

- GWR requires a definition for *nearby*
  - Kernel type
    - Fixed: *Nearby* is determined by a fixed distance band
    - Adaptive: *Nearby* is determined by a fixed number of neighbors
  - Bandwidth method
    - AIC or Cross Validation (CV): GWR will find the optimal distance or optimal number of neighbors
    - Bandwidth parameter: User-provided distance or user-provided number of neighbors



Interpreting GWR results

- Compare GWR R2 and AICc values to OLS R2 and AICc values
  - The better model has a lower AIC and a high R2.

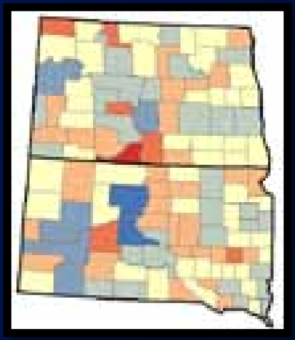
Bandwidth	:	2e+005
ResidualSquares	:	327.57434924067235
EffectiveNumber	:	30.68145239098456
Sigma	:	1.9258789406364027
AICc	:	518.280903017286
R2	:	0.9183016622131718
R2Adjusted	:	0.8908450815844072

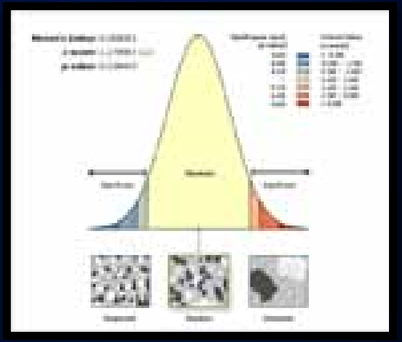
- Model predictions, residuals, standard errors, coefficients, and condition numbers are written to the output feature class.

Observed	Cond	LocalR2	Predicted	Intercept	C1_HVehiAc
70.419998	12.613701	0.881075	70.510341	85.562480	-79.900532
76.5	14.048718	0.834124	77.920484	85.36851	-78.018965
68.200009	12.25015	0.847111	68.064384	85.920030	-80.417780
73.190002	12.515339	0.857244	72.182168	84.312503	-65.634913
78.290001	11.758007	0.836626	77.979938	85.876829	-79.211314
70.230003	12.612641	0.874040	79.139514	84.765217	-77.439124

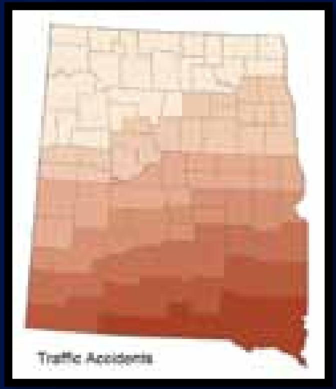
Mapped Output

- Residual maps show model under- and over-predictions. They shouldn't be clustered.





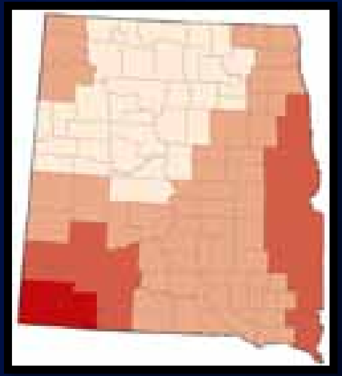
- Coefficient maps show how modeled relationships vary across the study area.



Mapped Output

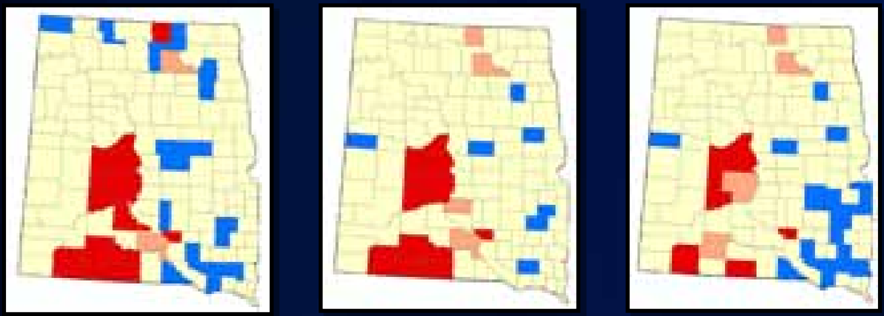
- Maps of Local R^2 values show where the model is performing best

- To see variation in model stability: apply graduated color rendering to Condition Numbers



GWR prediction

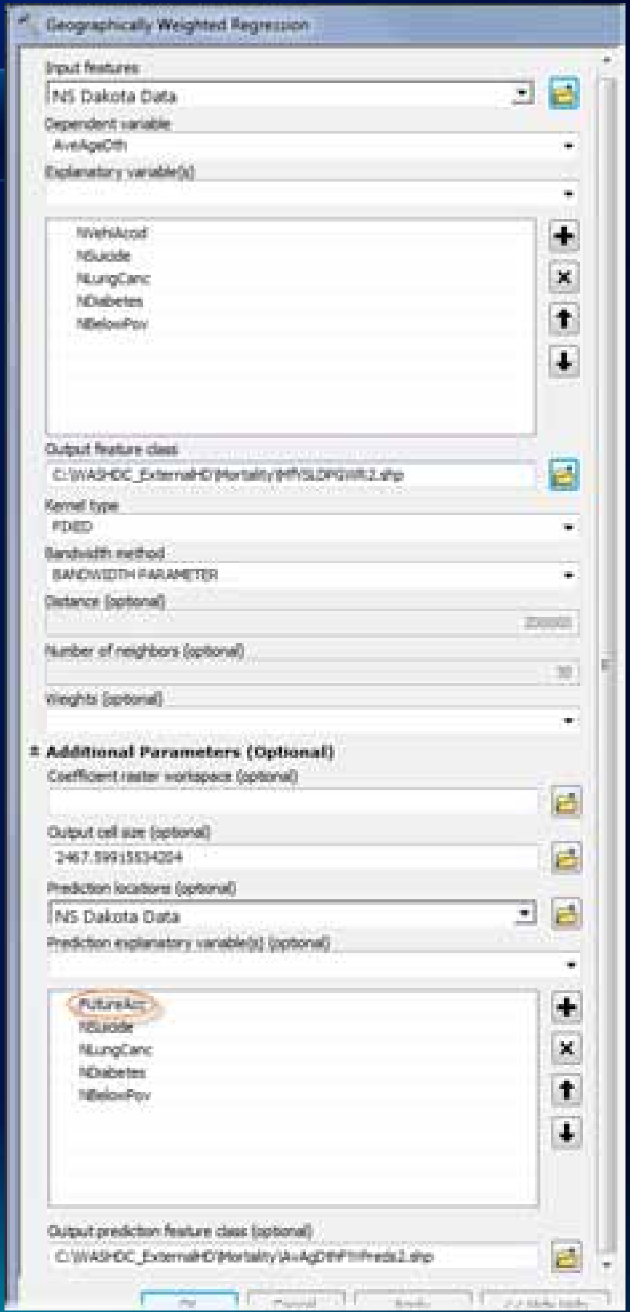
Calibrate the GWR model using known values for the dependent variable and all of the explanatory variables.



Observed Modeled Predicted

Provide a feature class of prediction locations containing values for all of the explanatory variables.

GWR will create an output feature class with the computed predictions.



Resources for learning more…

- **Spatial Pattern Analysis: Mapping Trends and Clusters**
  - Tue 8:30 Rm 2; Wed 3:15 Rm 2
- **Modeling Spatial Relationships using Regression Analysis**
  - Tue 10:15 Rm 2; Thu 1:30 Rm 1A/B
- **Spatial Statistics: Best Practices**
  - Tue 3:15 Rm 2; Thu 3:15 Rm 1A/B
- **Using R in ArcGIS**
  - Wed 12:00 Rm 1A/B
- **Road Ahead: Sharing of Analysis (ArcGIS 10.1)**
  - Wed 11:05 Rm 6B

Resources for learning more…

Kindly complete an evaluation form before you leave

QUESTIONS?

- www.esriurl.com/spatialstats
- Short videos
- Articles and blogs
- Online documentation
- Supplementary model and script tools
- Hot Spot, Regression, and ModelBuilder tutorials

LScott@Esri.com
LRosenshein@Esri.com