Exploring local variability in statistical relationships with

GEOGRAPHICALLY WEIGHTED REGRESSION

Betsy Breyer
Department of Geography
Portland State University
breyer@pdx.edu
GWR
Geographically Weighted Regression

OLS
Ordinary Least Squares (regression)
OVERVIEW

1. ORIGINS OF GWR
2. CRITIQUES OF GWR
3. USING GWR IN ARCGIS
   a) Exploring data
   b) Developing an OLS model
   c) Running GWR, interpreting results
   d) Model comparison
4. REFERENCES & RESOURCES
5. DISCUSSION
Section 1

ORIGINS OF GWR
Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity

Spatial nonstationarity is a condition in which a simple “global” model cannot explain the relationships between some sets of variables. The nature of the model must alter over space to reflect the structure within the data. In this paper, a technique is developed, termed geographically weighted regression, which attempts to capture this variation by calibrating a multiple regression model which allows different relationships to exist at different points in space. This technique is loosely based on kernel regression. The method itself is introduced and related issues such as the choice of a spatial weighting function are discussed. Following this, a series of related statistical tests are considered which can be described generally as tests for spatial nonstationarity. Using Monte Carlo methods, techniques are proposed for investigating the null hypothesis that the data may be described by a global model rather than a non-stationary one and also for testing whether individual regression coefficients are stable over geographic space. These techniques are demonstrated on a data set from the 1991 U.K. census relating car ownership rates to social class and male unemployment. The paper concludes by discussing ways in which the technique can be extended.

Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity

Spatial nonstationarity is a condition in which a simple “global” model cannot explain the relationships between some sets of variables. The nature of the model must alter over space to reflect the structure within the data. In this paper, a technique is developed, termed geographically weighted regression, which attempts to capture this variation by calibrating a multiple regression model which allows different relationships to exist at different points in space. This technique is loosely based on kernel regression. The method itself is introduced and related issues such as the choice of a spatial weighting function are discussed. Following this, a series of related statistical tests are considered which can be described generally as tests for spatial nonstationarity. Using Monte Carlo methods, techniques are proposed for investigating the null hypothesis that the data may be described by a global model rather than a non-stationary one and also for testing whether individual regression coefficients are stable over geographic space. These techniques are demonstrated on a data set from the 1991 U.K. census relating car ownership rates to social class and male unemployment. The paper concludes by discussing ways in which the technique can be extended.
• GWR estimates an OLS-like regression for each feature
  – Analogous to other local/global statistics
    • Morans I vs Local Morans
    • Getis-Ord General G vs Getis-Ord Gi*
• Designed to explore spatial nonstationarity in parameters
  – Nonstationary model, heterogeneous pattern
  – Spatial heterogeneity vs. spatial dependence
• Key to the analysis:
  – Model development
  – Define the neighborhood
The nature of local variation in statistical relationships

1. Primacy of place
   a) Nonstationarity arises from intrinsic local differences
   b) Global statements of spatial behavior are not possible
   c) Raises additional questions: what defines a neighborhood?

2. Model misspecification
   a) *Apparent* nonstationarity simply results from omitted variables
   b) When all sources of variability are included, the effect disappears
   c) Sampling bias?

See also

Section 2

CRITIQUES OF GWR
Simulation results

- GWR has not consistently differentiated between stationary and nonstationary data-generating processes
- Multicollinearity in estimated coefficients may bias results
- Unclear what tests can reliably diagnose model problems

Simulation studies:


"notoriously unreliable"


There is no "correct" here, as ?pchisq shows that if lower.tail= is not set it is taken by default to be TRUE. You will need to check this yourself. Note that GWR is a notoriously unreliable technique, and simulation studies indicate that it finds pattern in coefficients even when there is none. So any tests are doubtful anyway - it should only be used for exploring the data for possible missing variables or inappropriate functional forms.

Hope this clarifies,

Roger

Roger Bivand
Department of Economics, NHH Norwegian School of Economics,
Helleveien 30, N-5045 Bergen, Norway.
voice: +47 55 95 93 55; fax +47 55 95 95 43
e-mail: Roger.Bivand at nhh.no

R package spgwr: No standard errors and t-statistics
ArcGIS: includes standard errors, does not calculate t-statistics
GWR4: includes both standard errors and t-statistics
Section 3: An example in ArcGIS 10

RESIDENTIAL LAND USE AND DAILY HOUSEHOLD WATER USE
Candidate explanatory variables

- Daily water use (liters)
- Building Age
- Lot Area (m²)
- Building Area (m²)
- Property Value
- Impervious surface (%)
- Low vegetation (%)
- High vegetation (%)

Maps illustrating spatial distribution of various explanatory variables.
Model development

1. Explore the data
2. Fit an OLS model
3. Perform diagnostics
4. Run GWR
5. Compare models

Iterative model development
Scatterplot matrix: USEFUL

Geostatistical Analyst Toolbar: AWESOME
Model development

- Explore the data
- Fit an OLS model
- Perform diagnostics
- Run GWR
- Compare models

Iterative model development
Fit an OLS model

Ordinary Least Squares

Performs global Ordinary Least Squares (OLS) linear regression to generate predictions or to model a dependent variable in terms of its relationships to a set of explanatory variables. Results are accessible from the Results window.

Ordinary Least Squares

Input Feature Class
Sample_Taxlots_Summer_xycoord

Unique ID Field
ID

Output Feature Class
I:\Students\Instructors\changh\Betsy\Thesis\GWR\OLS\OLS_bldg.shp

Dependent Variable
SortLiners

Explanatory Variables
BLDG\_AGE
LANDVAL
BLDG\_VAL
TOTAL\_VAL
LOTAREA\_M
BLDG\_AREA\_M
IMP\_PER
HIGH\_PER
LOW\_PER

Output Options
Coefficient Output Table (optional)
I:\Students\Instructors\changh\Betsy\Thesis\GWR\OLS\OLS_bldg\_coefficients

Diagnostic Output Table (optional)
I:\Students\Instructors\changh\Betsy\Thesis\GWR\OLS\OLS_bldg\_diagnostics
**Summary of OLS Results**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>StdError</th>
<th>t-Statistic</th>
<th>Probability</th>
<th>Robust_SE</th>
<th>Robust_t</th>
<th>Robust_Pr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>19.41527</td>
<td>1.10009</td>
<td>17.65010</td>
<td>0.00000*</td>
<td>1.485714</td>
<td>13.067976</td>
<td>0.00000*</td>
</tr>
<tr>
<td>BLDGAREA_M</td>
<td>0.078753</td>
<td>0.006507</td>
<td>12.103116</td>
<td>0.000000</td>
<td>0.096600</td>
<td>8.152434</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

**OLS Diagnostics**

- Number of Observations: 450
- Number of Variables: 2
- Degrees of Freedom: 458
- Akaike's Information Criterion (AIC) [2]: 3287.210420
- Adjusted R-Squared [2]: 0.240676
- Joint F-Statistic [3]: 146.485414
- Prob (F, (1, 458)) degrees of freedom: 0.000000*
- Joint Wald Statistic [4]: 66.462176
- Prob (chi-squared, (1)) degrees of freedom: 0.000000*
- Koenker (BP) Statistic [5]: 13.687422
- Prob (chi-squared, (1)) degrees of freedom: 0.000216*
- Jarque-Bera Statistic [6]: 49.307420
- Prob (chi-squared, (2)) degrees of freedom: 0.000000*

**Notes on Interpretation**

- * Statistically significant at the 0.05 level.
- [1] Large VIF (> 7.5, for example) indicates explanatory variable redundancy.
- [5] Significant p-value indicates biased standard errors; use robust estimates.

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

Succeeded at Thu May 23 19:19:42 2013 (Elapsed Time: 6.00 seconds)
A properly specified OLS model

• Coefficients
  – Expected sign
  – Significant, p-value < 0.05
  – No multicollinearity (VIF)

• Goodness of fit
  – Model significance F-statistic
  – Goodness of fit: $R^2$, AICc

• Residuals $\sim N(0, \sigma)$
  – $N$: normally distributed
    • Jarque-Bera p-value > 0.05
  – $\sigma$: constant variance
    • Breusch-Pagen p-value > 0.05

• Residuals independent
  – Moran’s I p-value > 0.05
Residuals independent? No.

Spatial Autocorrelation Report

Global Moran's I Summary

<table>
<thead>
<tr>
<th>Moran's Index</th>
<th>Expected Index</th>
<th>Variance</th>
<th>z-score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.120487</td>
<td>0.002179</td>
<td>0.000325</td>
<td>6.795325</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

Given the p-value of 0.00, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Dataset Information

<table>
<thead>
<tr>
<th>Input Feature Class</th>
<th>Input Field</th>
<th>Conceptualization</th>
<th>Distance Method</th>
<th>Row Standardization</th>
<th>Distance Threshold</th>
<th>Weights Matrix File</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS_bldg</td>
<td>RESIDUAL</td>
<td>INVERSE_DISTANCE</td>
<td>EUCLIDEAN</td>
<td>False</td>
<td>11925.836998</td>
<td>None</td>
</tr>
</tbody>
</table>
What combination of variables makes a properly specified OLS model?

7! = 5040 possible models
Model development

1. Explore the data
2. Fit an OLS model
3. Perform diagnostics
4. Run GWR
5. Compare models

Iterative model development WITH PYTHON
Exploratory Regression

The Exploratory Regression tool evaluates all of the possible combinations of the input candidate explanatory variables, looking for models that best explain the dependent variable within the context of user-specified criteria. It produces an output report file and optional tables. A full explanation of each output is provided in the Interpreting Exploratory Regression Results document, found in the Documentation folder of the Supplementary Spatial Statistics folder. This tool uses Ordinary Least Squares (OLS) and Spatial Autocorrelation (Global Moran's I), both from the Spatial Statistics toolbox.
Exploratory regression

• Tests all possible combinations of explanatory variables
• All candidate models tested against criteria for a properly specified OLS model
  – Includes Moran’s I on residuals
    • Spatial weights matrix
• Summary statistics on candidate models:
  – Frequency of variable significance
  – Collinear explanatory variables
  – Models with highest goodness of fit
  – Models with normally distributed residuals
  – Models with spatially uncorrelated residuals
Exploratory regression

• Built into ArcGIS 10.1
  – Also new: Incremental Spatial Autocorrelation
• Both tools available for ArcGIS 10.0
  – Supplementary Spatial Statistics toolbox
  – http://esriurl.com/spatialstats
  – Python is available, too.
Using Exploratory Regression in ArcGIS: Possible workflow

1. Calculate Distance Band from Neighbor Count
2. Incremental Spatial Autocorrelation
   – Bandwidth potentially useful for mapping clusters
3. Generate Spatial Weights Matrix
4. Exploratory Regression
   – Ordinary least squares regression
   – Geographically weighted regression

Source: ESRI Spatial Statistics Best Practices
STEP 1

Calculate distance band from Neighbor Count
STEP 1

Calculate distance band from Neighbor Count

Distance where all features have at least 1 neighbor
Incremental Spatial Autocorrelation

Measures spatial autocorrelation at incremental distances and creates a graph of those distances and their corresponding z-scores. The graph can be used to choose an appropriate scale of analysis (distance band) to use for further analysis, for instance in a Hot Spot Analysis (Getis-Ord GI*). Peaks in the output graph indicate distances at which clustering is most pronounced. When more than one peak is present, clustering is pronounced at each of those distances. Select the distance that best corresponds to the scale of analysis you are interested in; often this is the first peak encountered.
STEP 2

Incremental Spatial Autocorrelation

Beginning distance is output from Step 1 ➞
**STEP 2**

Incremental Spatial Autocorrelation

Beginning distance is output from Step 1 ➤

First peak in z-scores at ~38,000 ft ➤
**STEP 3**

Generate spatial weights matrix

<table>
<thead>
<tr>
<th>Input Feature Class</th>
<th>I:\Students\Instructors\changh\Betsy\Thesis\GWR\Sample_Taxlots_Summer_xcoord.shp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique ID Field</td>
<td>ID</td>
</tr>
<tr>
<td>Output Spatial Weights Matrix File</td>
<td>I:\Students\Instructors\changh\Betsy\Thesis\GWR\SWM\Inverse_Distance.swm</td>
</tr>
<tr>
<td>Conceptualization of Spatial Relationships</td>
<td>INVERSE_DISTANCE</td>
</tr>
<tr>
<td>Distance Method (optional)</td>
<td>EUCLIDEAN</td>
</tr>
<tr>
<td>Exponent (optional)</td>
<td>1</td>
</tr>
<tr>
<td>Threshold Distance (optional)</td>
<td>38000</td>
</tr>
<tr>
<td>Number of Neighbors (optional)</td>
<td></td>
</tr>
<tr>
<td>Row Standardization (optional)</td>
<td></td>
</tr>
<tr>
<td>Input Table (optional)</td>
<td></td>
</tr>
</tbody>
</table>

**Conceptualization of Spatial Relationships**

Specifies how spatial relationships among features are conceptualized.

- **INVERSE_DISTANCE**—The impact of one feature on another feature decreases with distance.
- **FIXED_DISTANCE**—Everything within a specified critical distance of each feature is included in the analysis; everything outside the critical distance is excluded.
- **K NEAREST NEIGHBORS**—The closest *k* features are included in the analysis; *k* is a specified numeric parameter.
- **CONTIGUITY_EDGES_ONLY**—Polygon features that share a boundary are neighbors.
- **CONTIGUITY_EDGES_Corners**—Polygon features that share a boundary and/or share a node are neighbors.
- **DELAUNAY_TRIANGULATION**—A mesh of nonoverlapping triangles is created from feature centroids; features associated with triangle nodes that share edges are neighbors.
- **CONVERT_TABLE**—Spatial relationships are defined in a table.

Note: Polygon Contiguity methods are only available with an Arcinfo license.
**STEP 3**

Generate spatial weights matrix

Input is the bandwidth from Step 2
STEP 3
Generate spatial weights matrix

Output is a .swm file

Input is the bandwidth from Step 2

Conceptualization of Spatial Relationships

Specifies how spatial relationships among features are conceptualized:

- **INVERSE_DISTANCE**—The impact of one feature on another feature decreases with distance.
- **FIXED_DISTANCE**—Everything within a specified critical distance of each feature is included in the analysis; everything outside the critical distance is excluded.
- **K NEAREST NEIGHBORS**—The closest "K" features are included in the analysis; k is a specified numeric parameter.
- **CONTIGUITY_EDGES_ONLY**—Polygon features that share a boundary are neighbors.
- **CONTIGUITY_EDGES_CORNERS**—Polygon features that share a boundary and/or share a node are neighbors.
- **DELAUNAY_TRIANGULATION**—A mesh of nonoverlapping triangles is created from feature centroids; features associated with triangle nodes that share edges are neighbors.
- **CONVERT_TABLE**—Spatial relationships are defined in a table.

Note: Polygon Contiguity methods are only available with an ArcInfo license.
Conceptualization of spatial relationships

- Inverse distance
- Fixed distance
- Zone of indifference

Conceptualization of spatial relationships

See also: Space-time clustering, Grouping Analysis tool
STEP 4

Exploratory regression

The Exploratory Regression tool evaluates all of the possible combinations of the input candidate explanatory variables, looking for models that best explain the dependent variable within the context of user-specified criteria. It produces an output report file and optional tables. A full explanation of each output is provided in the Interpreting Exploratory Regression Results document, found in the Documentation folder of the Supplementary Spatial Statistics folder. This tool uses Ordinary Least Squares (OLS) and Spatial Autocorrelation (Global Moran’s I), both from the Spatial Statistics toolbox.
STEP 4

Exploratory regression

The Exploratory Regression tool evaluates all of the possible combinations of the input candidate explanatory variables, looking for models that best explain the dependent variable within the context of user-specified criteria. It produces an output report file and optional tables. A full explanation of each output is provided in the Interpreting Exploratory Regression Results document, found in the Documentation folder of the Supplementary Spatial Statistics folder. This tool uses Ordinary Least Squares (OLS) and Spatial Autocorrelation (Global Moran’s I), both from the Spatial Statistics toolbox.
STEP 4

Exploratory regression

Candidate explanatory variables

.swm file ➔
**STEP 4**

Exploratory regression

Candidate explanatory variables

.swm file ➔

Min/max number of variables per model

Significance test p-values

The Exploratory Regression tool evaluates all of the possible combinations of the input candidate explanatory variables, looking for models that best explain the dependent variable within the context of user-specified criteria. It produces an output report file and optional tables. A full explanation of each output is provided in the Interpreting Exploratory Regression Results document, found in the Documentation folder of the Supplementary Spatial Statistics folder. This tool uses Ordinary Least Squares (OLS) and Spatial Autocorrelation (Global Moran's I), both from the Spatial Statistics toolbox.
STEP 4

Exploratory regression

Running script ExploratoryRegression...

Choose 1 of 12 Summary
Highest Adjusted R-Squared Results
R2 AICc JB BP VIF MI Model
0.24 3289.26 0.00 0.00 1.00 0.00 + BLDGAREA_M***
0.17 3332.10 0.00 0.00 1.00 0.00 + LANDVAL***
0.16 3336.06 0.00 0.14 1.00 0.00 - BLDG_AGE***

Passing Models
R2 AICc JB BP VIF MI Model

Choose 2 of 12 Summary
Highest Adjusted R-Squared Results
R2 AICc JB BP VIF MI Model
0.31 3248.04 0.00 0.05 1.10 0.00 - BLDG_AGE*** + BLDGAREA_M****
0.28 3264.27 0.00 0.00 1.01 0.00 - BLDG_AGE*** + TOTALVAL***
0.28 3268.99 0.00 0.00 1.22 0.00 + BLDGAREA_M*** + IMP_M***

Passing Models
R2 AICc JB BP VIF MI Model

Choose 3 of 12 Summary
Highest Adjusted R-Squared Results
R2 AICc JB BP VIF MI Model
0.33 3236.40 0.00 0.00 1.50 0.00 - BLDG_AGE*** + LANDVAL** + BLDGAREA_M****
0.32 3240.67 0.00 0.09 2.15 0.00 - BLDG_AGE*** + TOTALVAL** + BLDGAREA_M****
0.32 3241.65 0.00 0.05 1.35 0.00 - BLDG_AGE*** + BLDGAREA_M*** + IMP_M***

Passing Models
R2 AICc JB BP VIF MI Model

Choose 4 of 12 Summary
Highest Adjusted R-Squared Results
R2 AICc JB BP VIF MI Model
0.33 3231.86 0.00 0.00 1.60 0.00 - BLDG_AGE*** + LANDVAL** + BLDGAREA_M**** + IMP_M***
0.33 3233.74 0.00 0.01 1.54 0.00 - BLDG_AGE*** + LANDVAL** + LCTAREA_M** + BLDGAREA_M***
0.33 3234.62 0.00 0.01 6.72 0.00 - BLDG_AGE*** - BLDGVAL** + TOTALVAL** + BLDGAREA_M***

Passing Models
R2 AICc JB BP VIF MI Model

Passing models would be here
### Step 4

**Exploratory regression**

Which diagnostic tests are not being passed?

<table>
<thead>
<tr>
<th>Search Criterion Cutoff</th>
<th>Trials</th>
<th># Passed</th>
<th>% Passed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min Adjusted R-Squared  &gt; 0.50</td>
<td>1585</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Max Coefficient p-value &lt; 0.05</td>
<td>1585</td>
<td>283</td>
<td>17.85</td>
</tr>
<tr>
<td>Max VIF Value &lt; 7.50</td>
<td>1585</td>
<td>1351</td>
<td>85.24</td>
</tr>
<tr>
<td>Min Jarque-Bera p-value &gt; 0.10</td>
<td>1585</td>
<td>3</td>
<td>0.19</td>
</tr>
<tr>
<td>Min Moran's I p-value &gt; 0.10</td>
<td>18</td>
<td>6</td>
<td>33.33</td>
</tr>
</tbody>
</table>

Which variables are consistently significant?

<table>
<thead>
<tr>
<th>Variable</th>
<th>% Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLDG_AGE</td>
<td>100.00</td>
</tr>
<tr>
<td>LANDVAL</td>
<td>80.60</td>
</tr>
<tr>
<td>BLDGVAL</td>
<td>65.12</td>
</tr>
<tr>
<td>TOTALVAL</td>
<td>77.76</td>
</tr>
<tr>
<td>LOTAREA_M</td>
<td>72.95</td>
</tr>
<tr>
<td>BLDGAREA_M</td>
<td>100.00</td>
</tr>
<tr>
<td>IMP_PER</td>
<td>42.70</td>
</tr>
<tr>
<td>HIGH_PER</td>
<td>30.60</td>
</tr>
<tr>
<td>LOW_PER</td>
<td>33.63</td>
</tr>
<tr>
<td>IMP_M</td>
<td>97.51</td>
</tr>
<tr>
<td>LOW_M</td>
<td>48.22</td>
</tr>
<tr>
<td>HIGH_M</td>
<td>43.24</td>
</tr>
</tbody>
</table>
**STEP 4**
Exploratory regression

### Which models have normal residuals?

<table>
<thead>
<tr>
<th>JB</th>
<th>R2</th>
<th>AICc</th>
<th>BP</th>
<th>VIF</th>
<th>MI</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.124366</td>
<td>0.173228</td>
<td>3331.519326</td>
<td>0.000003</td>
<td>7.909751</td>
<td>0.000000</td>
<td>+BLDGVAL*** +LOTAREA_M*** -LOW_PER -LOW_M</td>
</tr>
<tr>
<td>0.110102</td>
<td>0.172788</td>
<td>3330.720833</td>
<td>0.000000</td>
<td>2.976807</td>
<td>0.000000</td>
<td>+BLDGVAL*** +LOTAREA_M*** -LOW_M***</td>
</tr>
<tr>
<td>0.102163</td>
<td>0.171989</td>
<td>3331.164556</td>
<td>0.000023</td>
<td>1.046833</td>
<td>0.000000</td>
<td>+BLDGVAL*** +LOTAREA_M*** -LOW_PER***</td>
</tr>
</tbody>
</table>

### Which models have spatially uncorrelated residuals?

<table>
<thead>
<tr>
<th>MI</th>
<th>R2</th>
<th>AICc</th>
<th>JB</th>
<th>BP</th>
<th>VIF</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.223620</td>
<td>0.330270</td>
<td>3234.618079</td>
<td>0.000000</td>
<td>0.014348</td>
<td>6.724090</td>
<td>-BLDG_AGE*** -BLDGVAL** +TOTALVAL** +BLDGAREA_M***</td>
</tr>
<tr>
<td>0.320787</td>
<td>0.334135</td>
<td>3231.956051</td>
<td>0.000000</td>
<td>0.001452</td>
<td>1.596901</td>
<td>-BLDG_AGE*** +LANDVAL** +BLDGAREA_M*** +IMP_M**</td>
</tr>
<tr>
<td>0.213209</td>
<td>0.341931</td>
<td>3227.588542</td>
<td>0.000000</td>
<td>0.006685</td>
<td>283.707489</td>
<td>-BLDG_AGE*** -LANDVAL*** -BLDGVAL*** +TOTALVAL***</td>
</tr>
</tbody>
</table>
OLS regression: Estimated coefficients

\[ Y = 22.34 + 0.00025(BLDGVAL) + 0.00985(LOTAREA_M) - 0.008885(LOW_M) + \text{error} \]
OLS regression: Residual normality

OLS Diagnostics

Number of Observations: 460  Number of Variables: 4
Degrees of Freedom: 456  Akaike's Information Criterion (AIC) [2]: 3328.588674
Multiple R-Squared [2]: 0.178194  Adjusted R-Squared [2]: 0.172788
Joint F-Statistic [3]: 32.958560  Prob(>F), (3, 456) degrees of freedom: 0.000000*
Joint Wald Statistic [4]: 49.558717  Prob(>chi-squared), (3) degrees of freedom: 0.000000*
Koenker (BP) Statistic [5]: 34.735759  Prob(>chi-squared), (3) degrees of freedom: 0.000000*
Jarque-Bera Statistic [6]: 4.412699  Prob(>chi-squared), (2) degrees of freedom: 0.110102

Histogram and QQ-Plot showing residual normality.
OLS regression: Residual dependence

Spatial Autocorrelation Report

Moran’s Index: 0.193969
z-score: 16.358521
p-value: 0.000000

Given the z-score of 16.36, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Global Moran's I Summary

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moran's Index</td>
<td>0.193969</td>
</tr>
<tr>
<td>Expected Index</td>
<td>-0.002179</td>
</tr>
<tr>
<td>Variance</td>
<td>0.000144</td>
</tr>
<tr>
<td>z-score</td>
<td>16.358521</td>
</tr>
<tr>
<td>p-value</td>
<td>0.000000</td>
</tr>
</tbody>
</table>
Model development

Explore the data
Fit an OLS model
Perform diagnostics
Run GWR
Compare models

Iterative model development
WITH PYTHON
GWR: Setting kernel and bandwidth

Bandwidth method

Specifies how the extent of the kernel should be determined. When AICc or CV are selected, the tool will find the optimal distance/neighbor parameter for you.

- **AICc**—The extent of the kernel is determined using the Akaike Information Criterion (AICc).
- **CV**—The extent of the kernel is determined using Cross Validation.
- **BANDWIDTH PARAMETER**—The extent of the kernel is determined by a fixed distance or a fixed number of neighbors.
GWR: Output

Full explanation of statistical output:
GWR: Coefficient estimates

Global: 0.000025

Global: 0.009847

Global: -0.008885
GWR: Local multicollinearity
GWR: Standardized residuals

Standardized residuals

-3.13 - -1.85
-1.84 - -0.91
-0.90 - -0.24
-0.23 - 0.40
0.41 - 1.15
1.16 - 2.13
2.14 - 3.72
Model development

- Explore the data
- Fit an OLS model
- Perform diagnostics
- Run GWR
- Compare models

Iterative model development WITH PYTHON
How to compare models?

• Compare with tables
  – Global OLS coefficient against distribution of GWR coefficients
  – Improvement in goodness of fit ($R^2$, AICc)
  – Improvement in residual diagnostics

• ANOVA table to compare models
  – GWR4 outputs this automatically

• Model validation?
  – Divide data into calibration and validation subsets.
  – Use prediction locations (under additional parameters) to compare predicted to observed
Calibration and validation

• Create random points
  – Containing feature = input data
    • Selects a random subset of points from input data
    • Save as your validation dataset
• Join random points back to input data
  – Select points that join, export points as validation dataset with attributes
  – Switch selection, export as calibration dataset
• Run GWR on calibration dataset
  – Include validation dataset as Prediction Locations
  – Add explanatory variables *in same order*
A few tips for GWR

- Use only on a large dataset ( > 160 features)
- Continuous variables with no NULL values
- SEVERE MODEL DESIGN PROBLEMS
  - Local multicollinearity > 30
    - Center variables (subtracting out the mean)
- Residuals still autocorrelated?
  - Try including a distance-based variable
    - E.g. Distance to urban center, major arterial
    - Geostatistical Analyst, Explore data, trend analysis may assist in selection
New directions in GWR research

• Flexible bandwidth
  – Why would we have a global bandwidth in a local model?
  – Allow each explanatory variable to have its own bandwidth

• Nonlinear models

• Non-Euclidean distance metrics
  – Manhattan distance, Network distance
  – Social, cost, aspatial ‘distance’

• Check out GWR4 for the following functionality:
  – Semiparametric model – some coefficients estimated globally, others locally
  – Poisson (for count data) or logistic (binary response)
  – Bisquare in addition to Gaussian kernel
  – Automatically compares OLS with GWR using ANOVA
Section 4

REFERENCES & RESOURCES
Software

GWR4 software (Forthingham, University of St Andrews)
http://www.st-andrews.ac.uk/geoinformatics/gwr/gwr-software/
- GWR4 (free download)
- Link to various R packages

ESRI software
http://esriurl.com/spatialstats
- Spatial statistics toolbox (for < 10.1) with python code
- Also, videos, tutorials, presentations, documentation
ARTICLES


BOOKS


RECOMMENDED VIDEO


Section 5

DISCUSSION