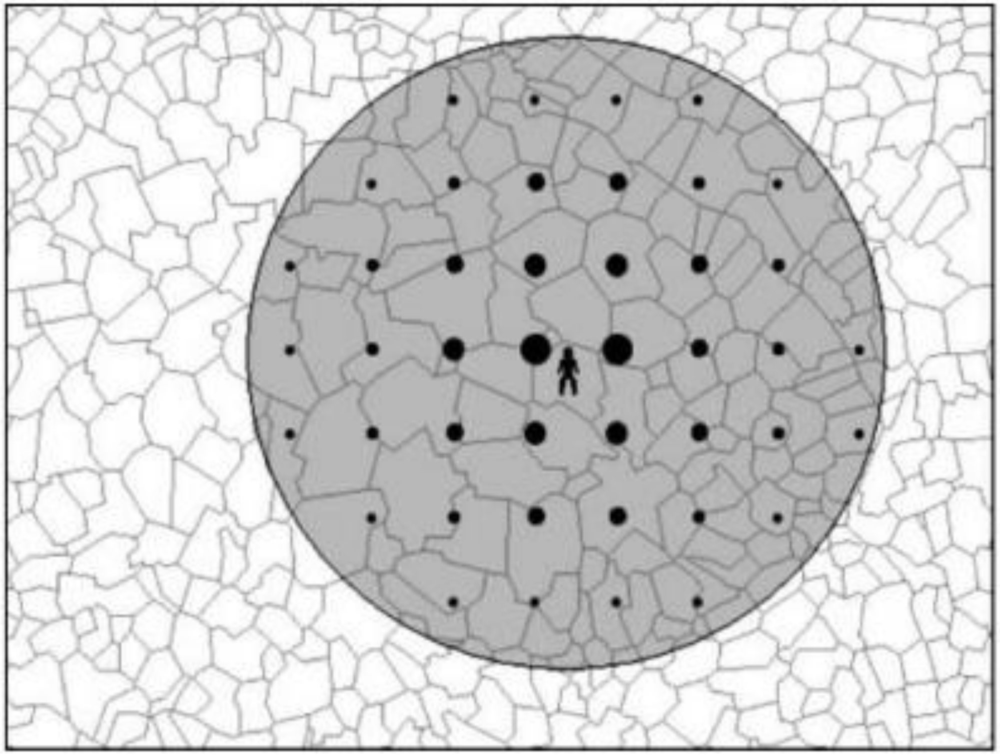


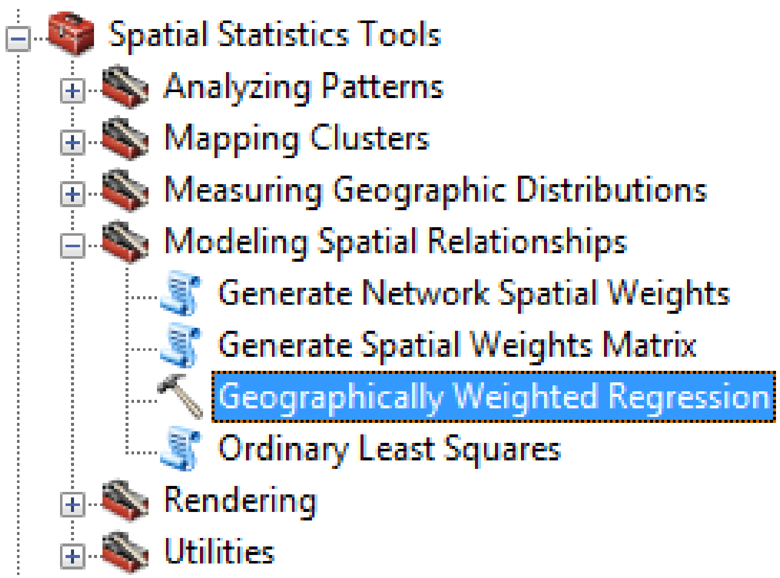
Exploring local variability in statistical relationships with

# GEOGRAPHICALLY WEIGHTED REGRESSION

Betsy Breyer
Department of Geography
Portland State University
breyer@pdx.edu

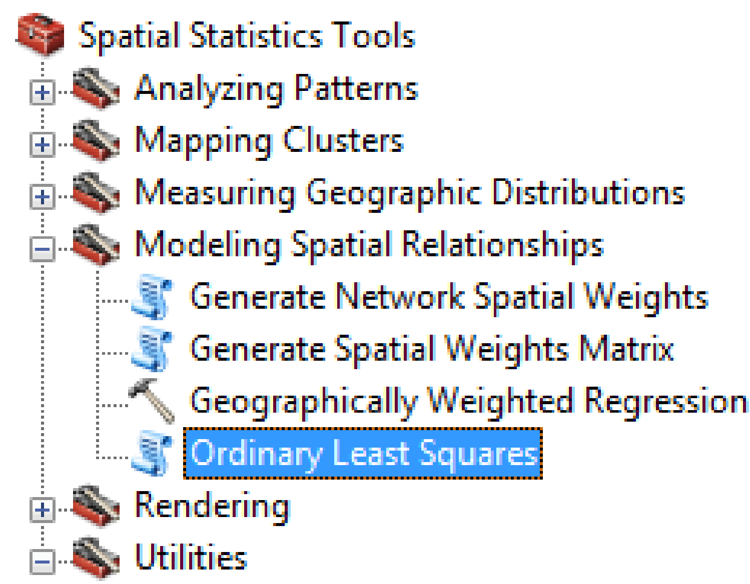
# GWR

## 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

*Chris Brunsdon, A. Stewart Fotheringham*
*and Martin E. Charlton*

# *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.*

Brunsdon, C., A.S. Fotheringham, and M.E. Charlton. 1996. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis* 28 (4): 281-298.

**Chris Brunsdon, A. Stewart Fotheringham and Martin E. Charlton**

# 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.*


Brunsdon, C., A.S. Fotheringham, and M.E. Charlton. 1996. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis* 28 (4): 281-298.

# GWR: A local regression model

- 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) A*pparent* nonstationarity simply results from omitted variables
   b) When all sources of variability are included, the effect disappears
   c) Sampling bias?

See also


Fotheringham, A.S. and C. Brunsdon. 1999. Local forms of spatial analysis. *Geographical Analysis* 31 (4): 340-358.

Fotheringham, A.S., C. Brunsdon, and M.E. Charlton. 2000. *Quantitative geography: Perspectives on spatial data analysis*. London: SAGE Publications

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:

Wheeler D, Tiefelsdorf M, 2005, "Multicollinearity and correlation among local regression coefficients in geographically weighted regression" Journal of Geographical Systems 7 (2) 161 - 187.

Páez, A., S. Farber, and D. Wheeler. 2011. A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. *Environment and Planning A* 43 (12): 2992 – 3010.

# “notoriously unreliable”

http://r-sig-geo.2731867.n2.nabble.com/A-question-about-gwr-morantest-pvalue-td7292670.html

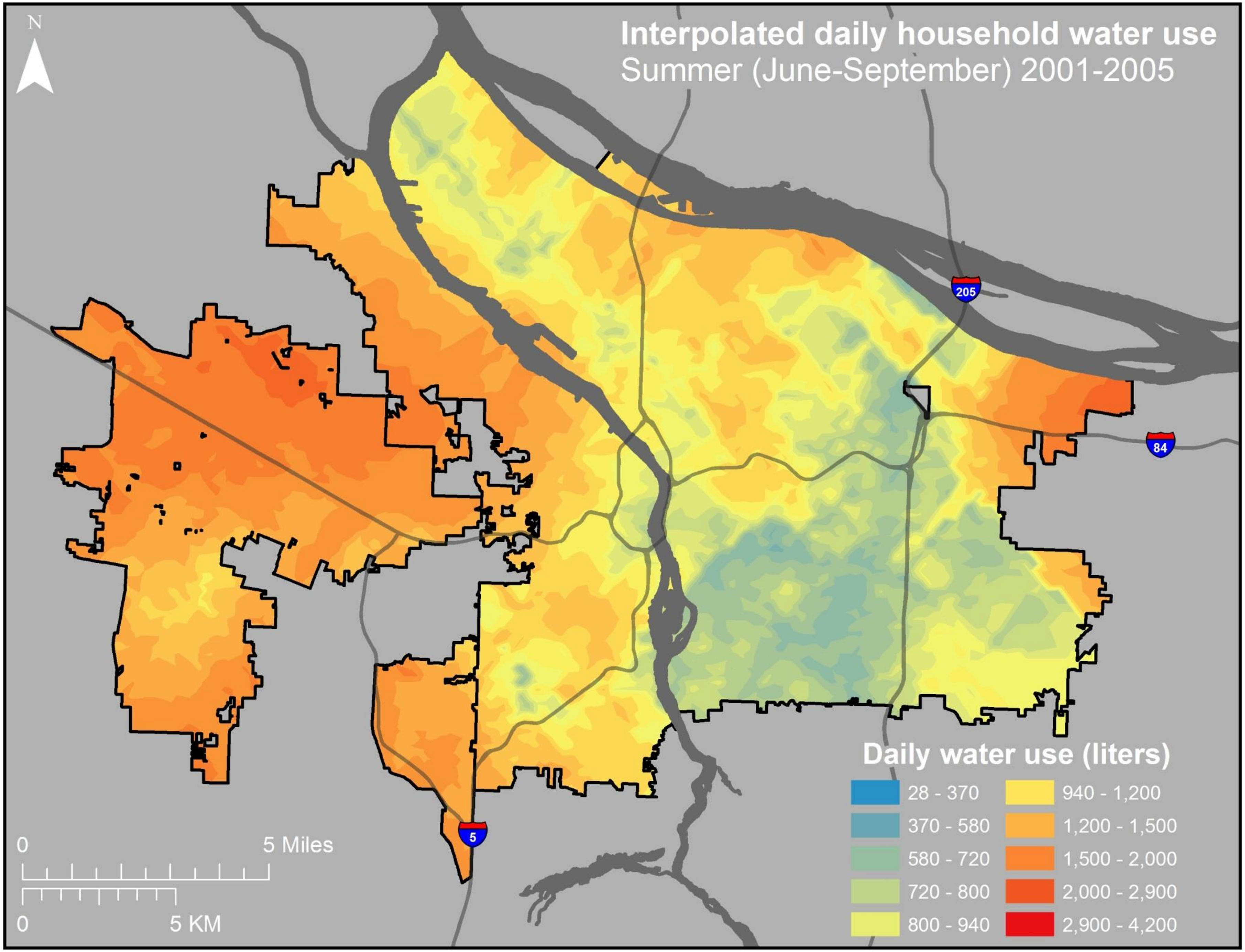
> There is no "correct" here, as ?pchisq shows that if lower.tail= is not set it is taken by default to be TRUE. You will need to check this yourself. Note that GWR is a notoriously unreliable technique, and simulation studies indicate that it finds pattern in coefficients even when there is none. So 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

Wrote GWR package (spgwr) in R

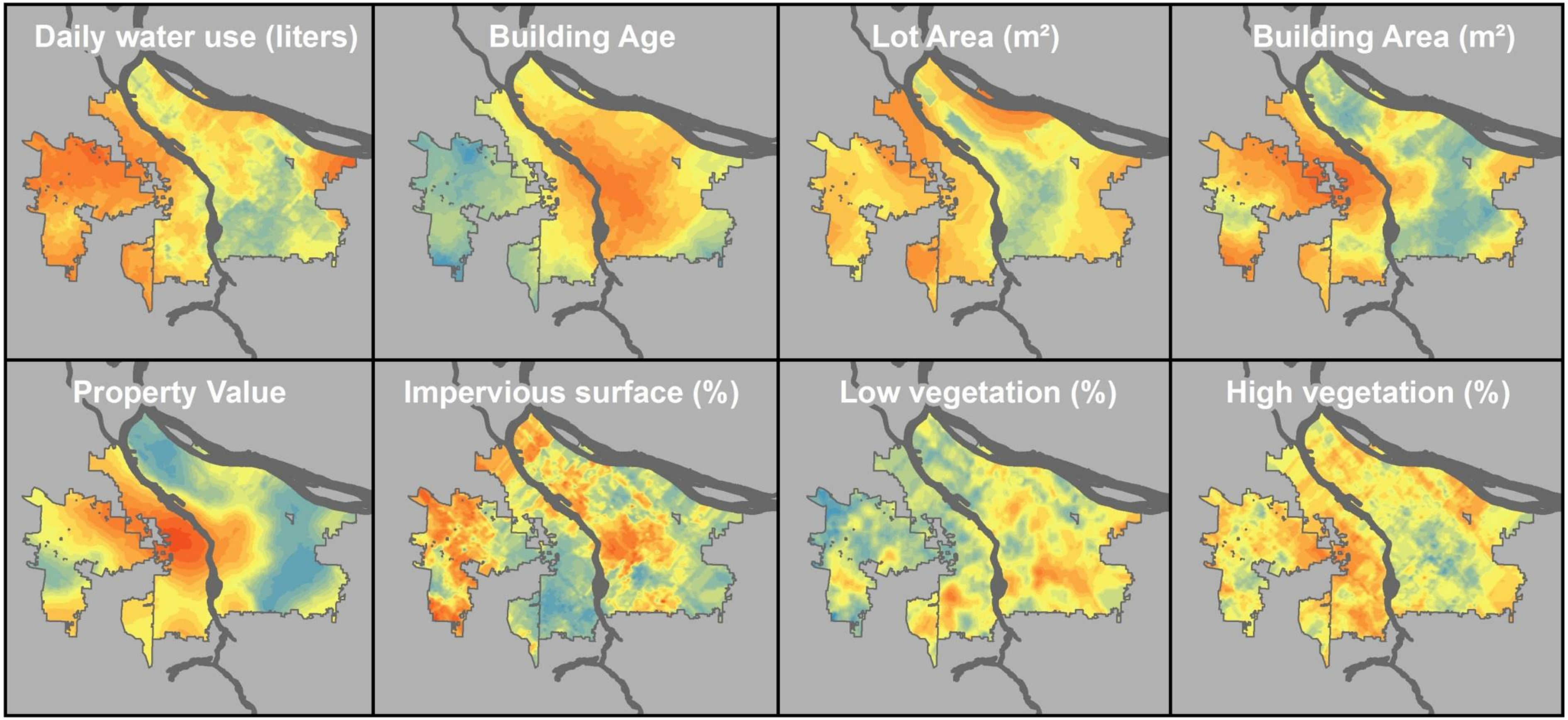
R package spgwr: No standard errors and t-statistcs
ArcGIS: includes standard errors, does not calculate t-statistics
GWR4: includes both standard errors and t-statistics

Section 3: An example in ArcGIS 10

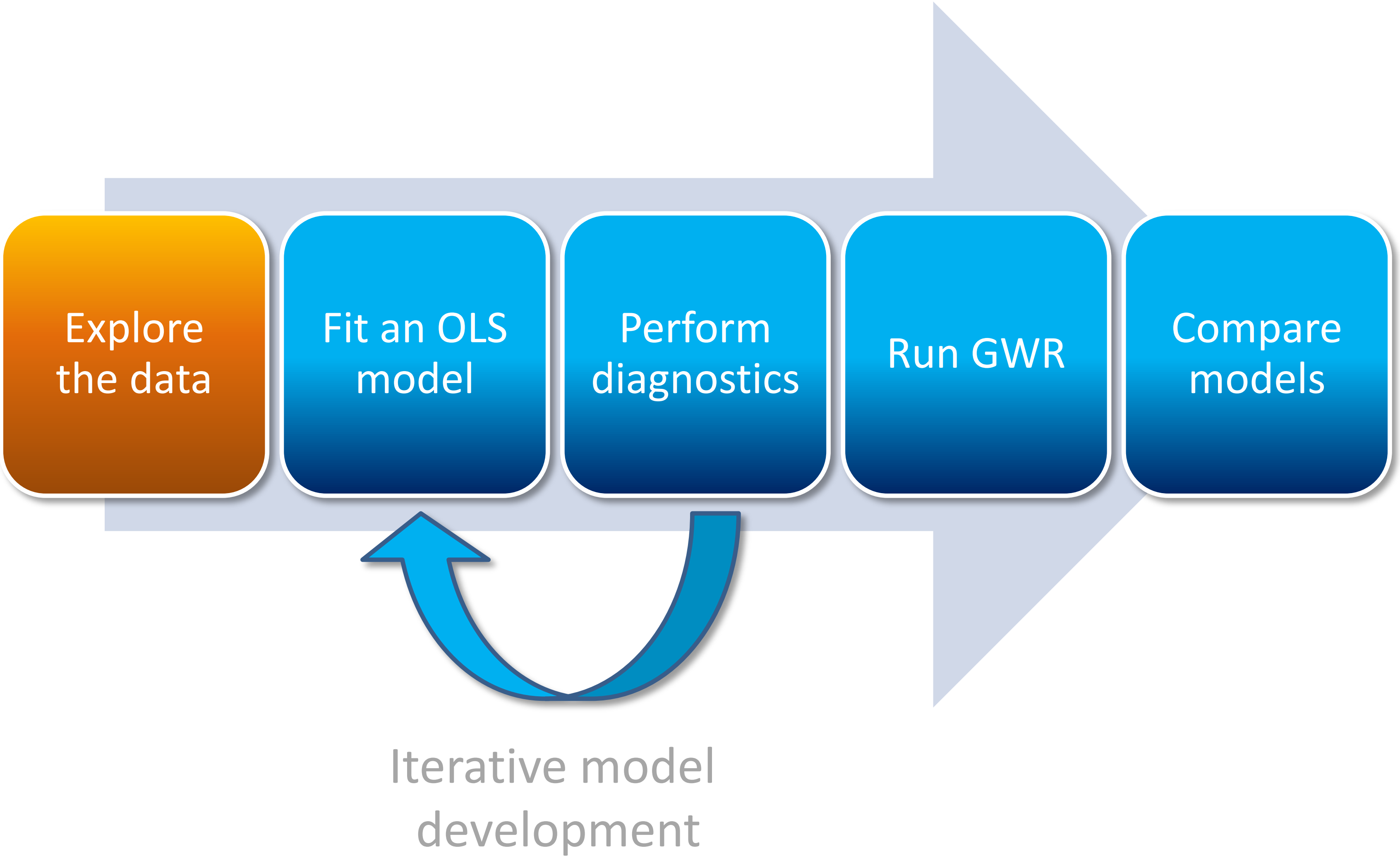
# RESIDENTIAL LAND USE AND DAILY HOUSEHOLD WATER USE

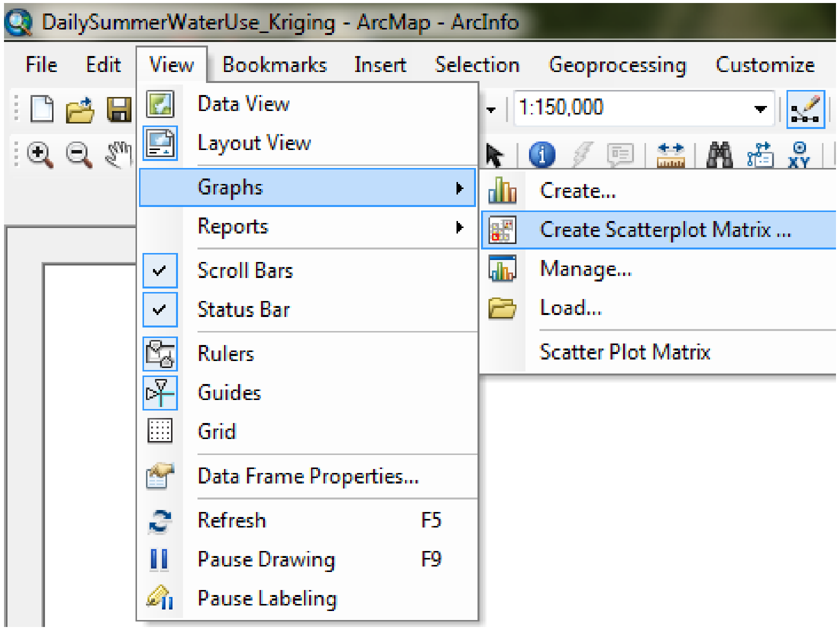

N
Interpolated daily household water use
Summer (June-September) 2001-2005
205
84
5
Daily water use (liters)
28 - 370
370 - 580
580 - 720
720 - 800
800 - 940
940 - 1,200
1,200 - 1,500
1,500 - 2,000
2,000 - 2,900
2,900 - 4,200
0
5 Miles
0
5 KM

# Candidate explanatory variables

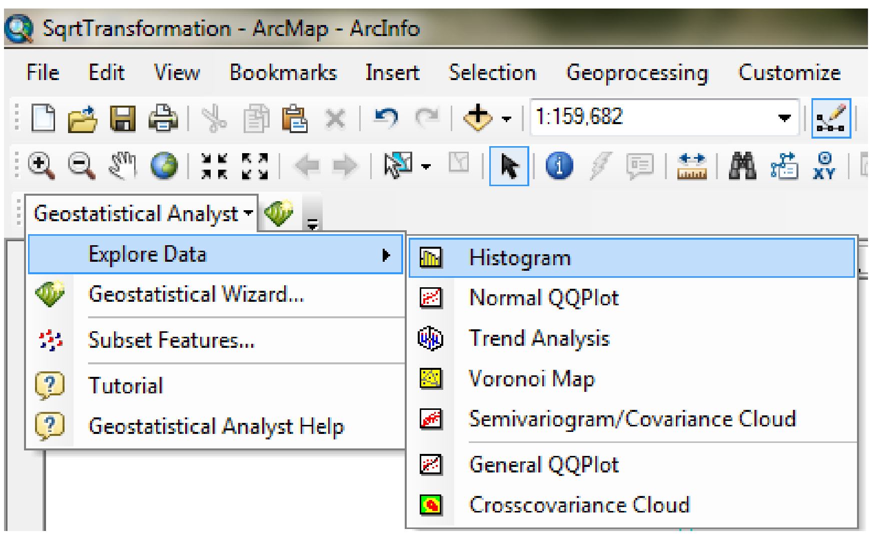


# Model development



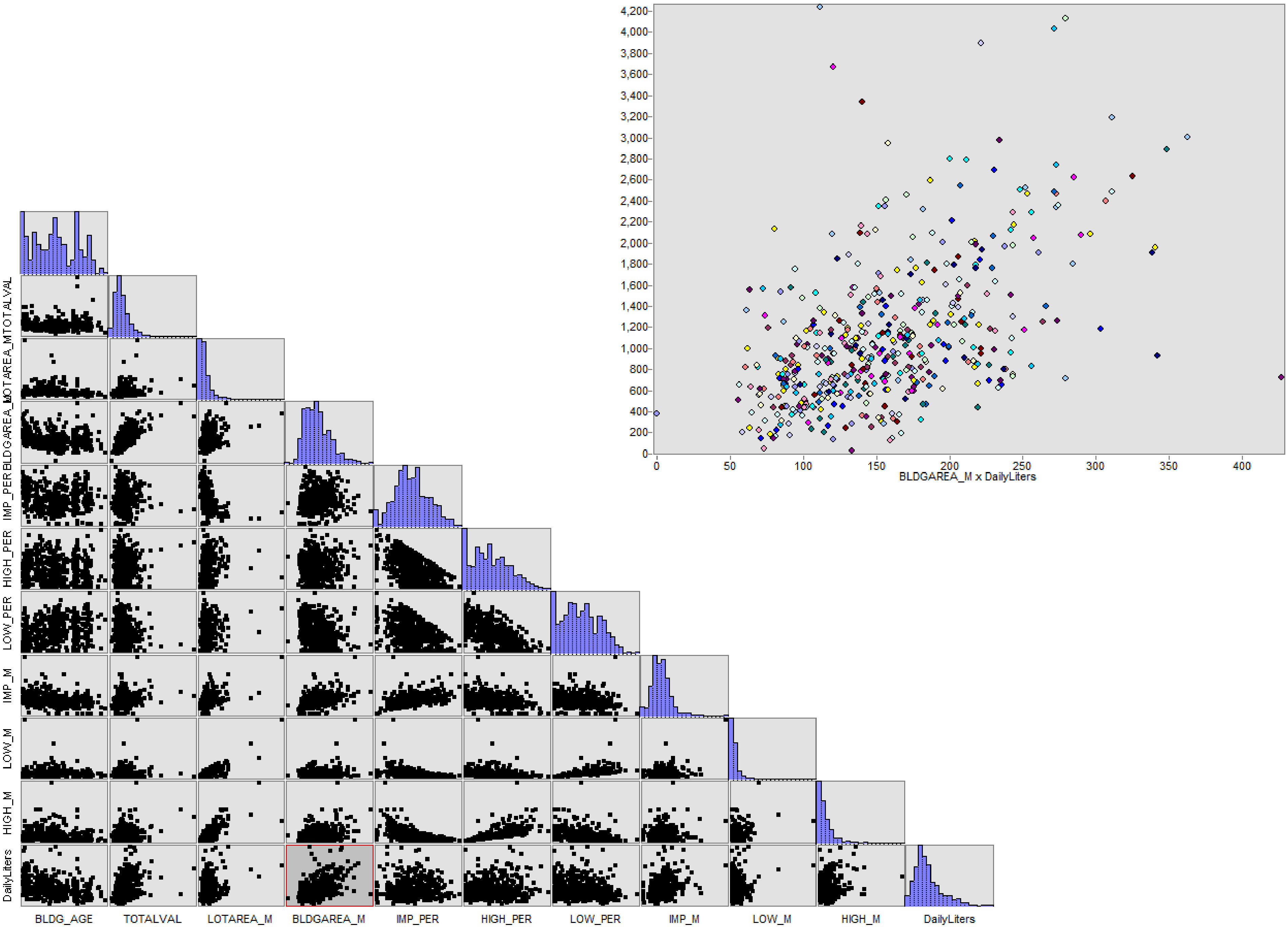



Scatterplot matrix: USEFUL



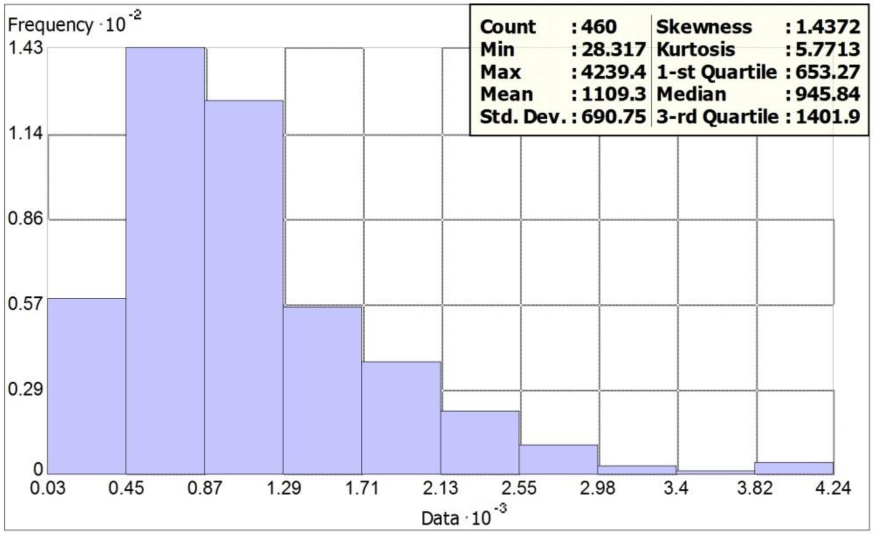

Geostatistical Analyst Toolbar: AWESOME

## Scatter Plot Matrix



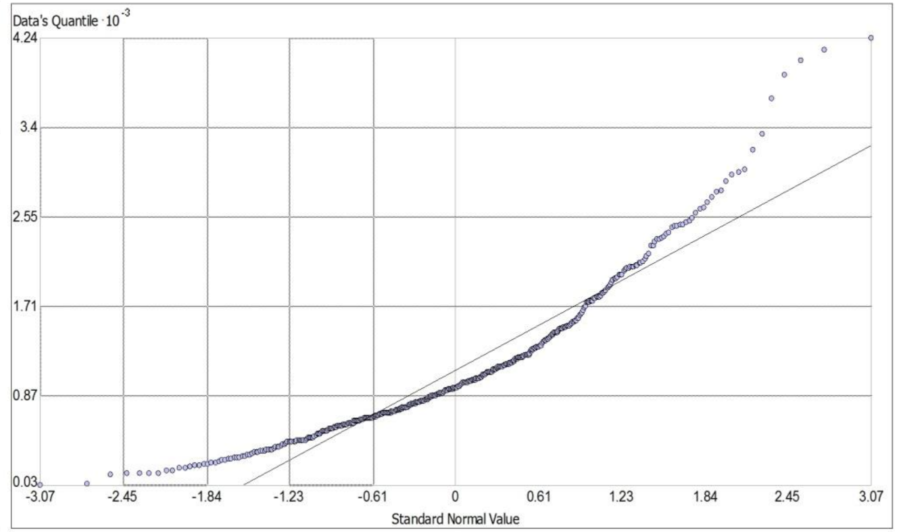

*Water use vs. land use in Portland, Oregon*

Histogram
Transformation: None



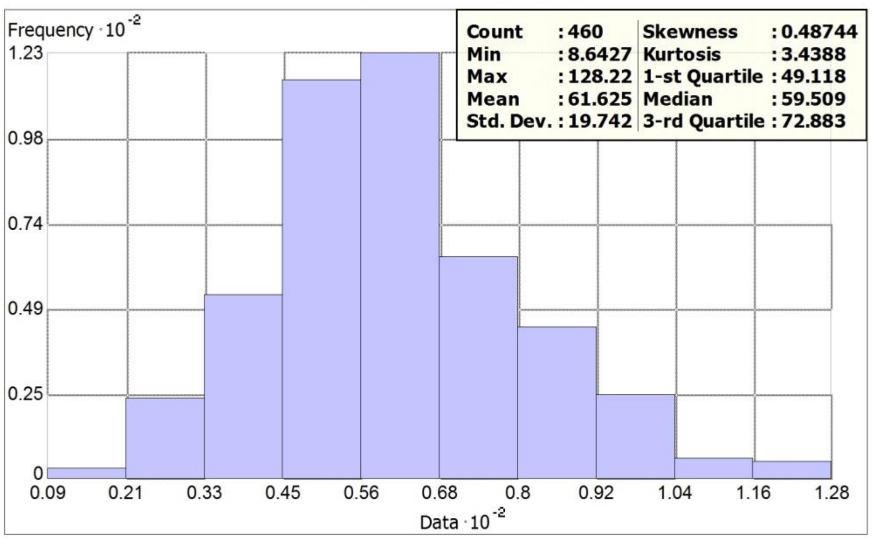

Data Source: Sample_Taxlots_Summer Attribute: DailyLiter

Normal QQPlot
Transformation: None



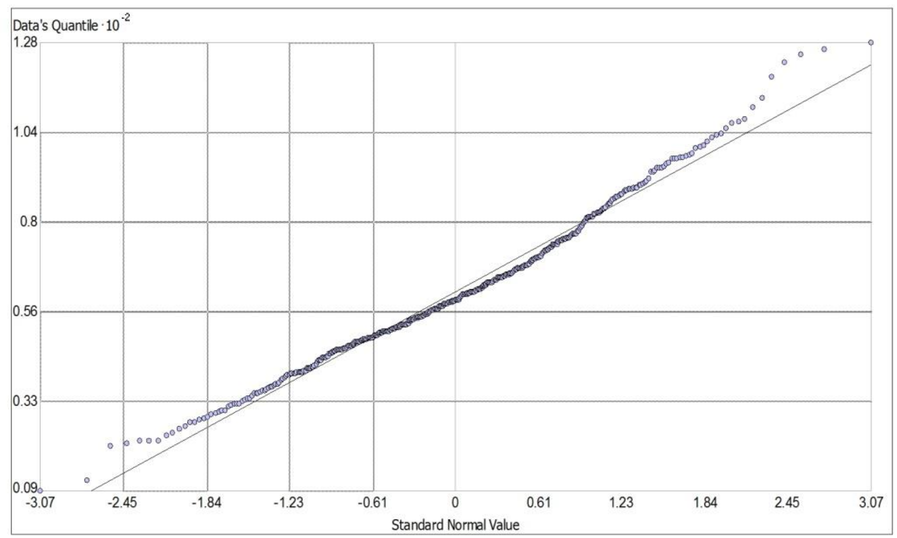

Data Source: Sample_Taxlots_Summer Attribute: DailyLiter

Histogram
Transformation: Box-Cox, Parameter: 0.5



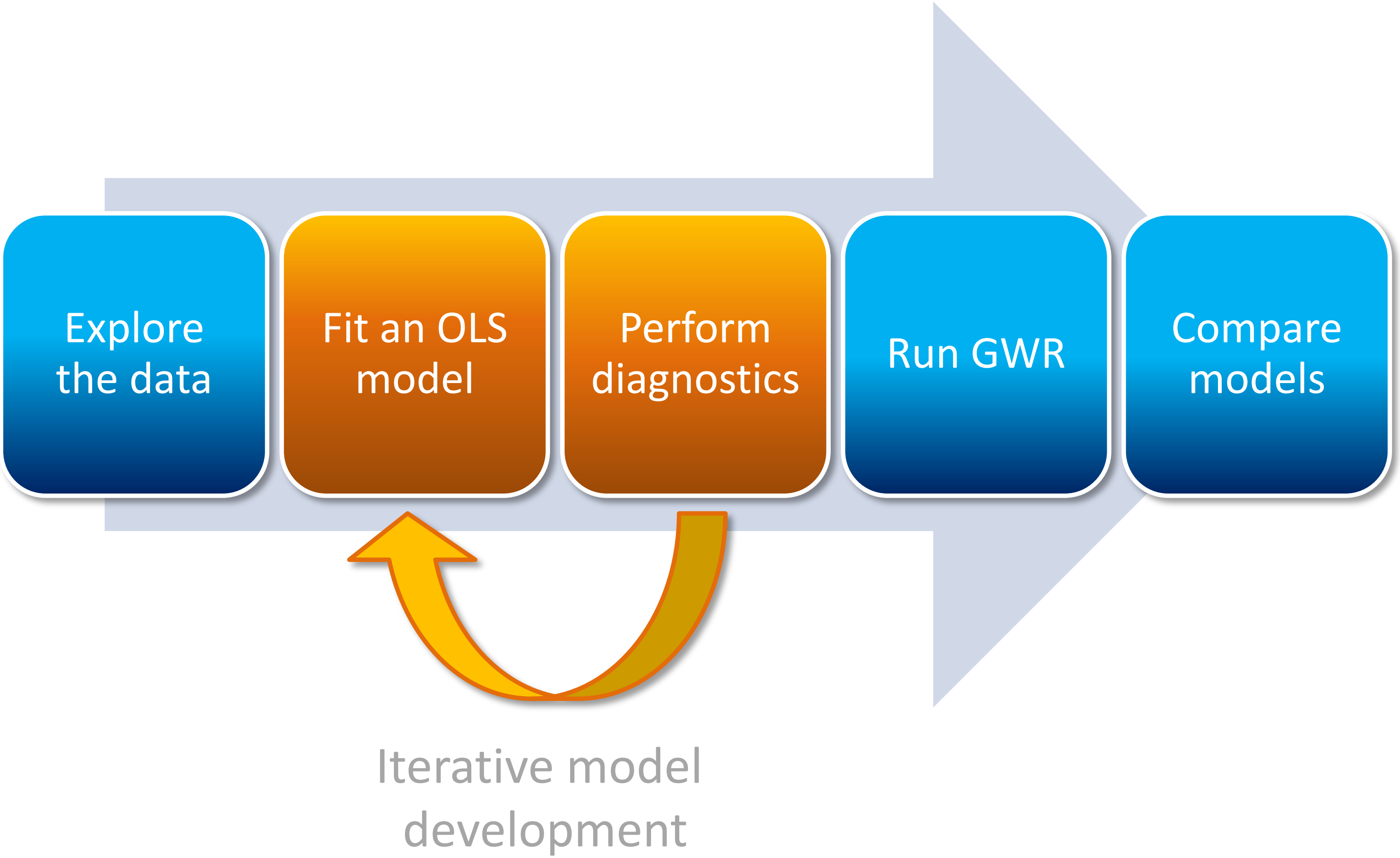

Data Source: Sample_Taxlots_Summer Attribute: DailyLiter

Normal QQPlot
Transformation: Box-Cox, Parameter: 0.5

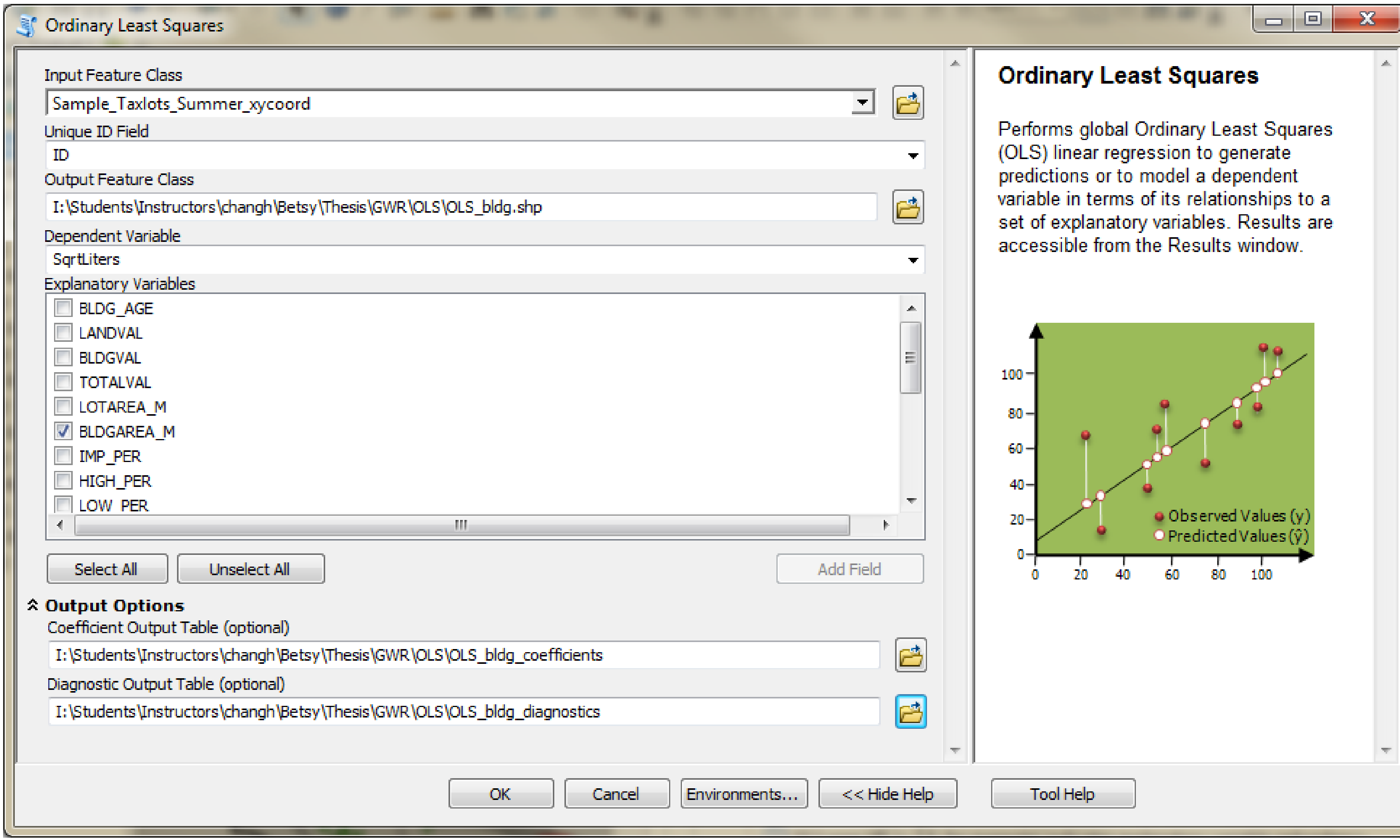



Data Source: Sample_Taxlots_Summer Attribute: DailyLiter

# Model development



# Fit an OLS model



Results

Current Session

Ordinary Least Squares [191942_05232013]

Output Feature Class: OLS_bldg.shp

Coefficient Output Table: OLS_bldg_coefficients

Diagnostic Output Table: OLS_bldg_diagnostics

Inputs

Environments

Messages

```
Executing: OrdinaryLeastSquares Sample_Taxlots_Summer_xycoord ID I:\Students\Instructors\changh\Betsy\Thesis\GWR\
Start Time: Thu May 23 19:19:36 2013
Running script OrdinaryLeastSquares...
                    Summary of OLS Results
Variable   Coefficient StdError t-Statistic Probability Robust_SE Robust_t  Robust_Pr
Intercept  19.415272  1.100009 17.650108  0.000000*  1.485714  13.067976 0.000000*
BLDGAREA_M 0.078753   0.006507 12.103116  0.000000*  0.009660  8.152434  0.000000*


                    OLS Diagnostics
Number of Observations:    460       Number of Variables:              2
Degrees of Freedom:        458       Akaike's Information Criterion (AIC) [2]:   3287.210420
Multiple R-Squared [2]:    0.242331  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.
[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.

WARNING 000851: Use the Spatial Autocorrelation (Moran's I) Tool to ensure residuals are not spatially autocorrelated.
Writing Coefficient Output Table....
I:\Students\Instructors\changh\Betsy\Thesis\GWR\OLS\OLS_bldg_coefficients.dbf
Writing Diagnostic Output Table....
I:\Students\Instructors\changh\Betsy\Thesis\GWR\OLS\OLS_bldg_diagnostics.dbf
Completed script OrdinaryLeastSquares...
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)

Summary of OLS Results

| Variable | Coefficient | StdError | t-Statistic | Probability | Robust_SE | Robust_t | Robust_Pr |
|---|---|---|---|---|---|---|---|
| Intercept | 19.415272 | 1.100009 | 17.650108 | 0.000000* | 1.485714 | 13.067976 | 0.000000* |
| BLDGAREA_M | 0.078753 | 0.006507 | 12.103116 | 0.000000* | 0.009660 | 8.152434 | 0.000000* |

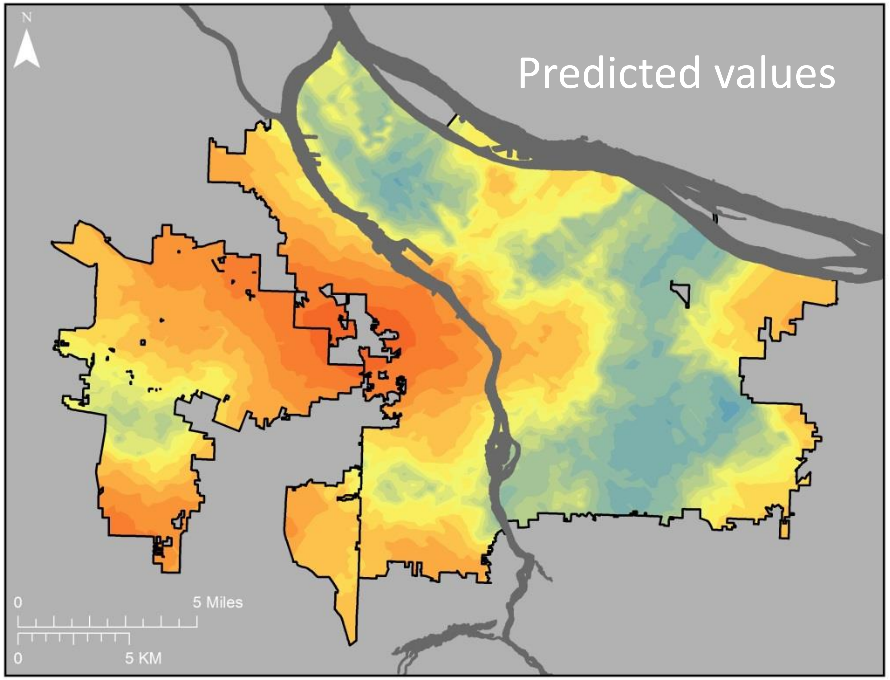
- **Goodness of fit**
  - Model significance F-statistic
  - Goodness of fit: $R^2$, AICc

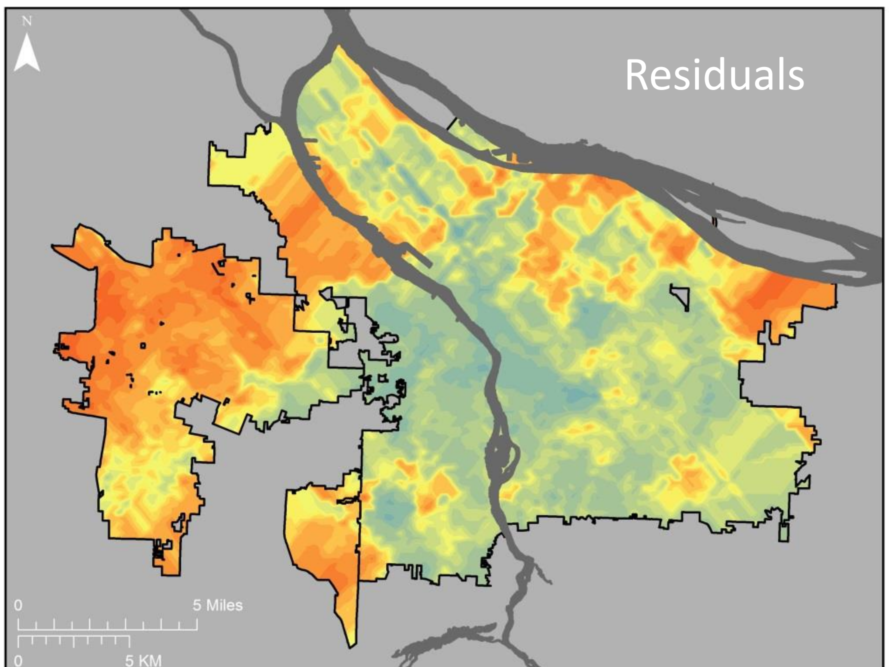
| | | | |
|---|---|---|---|
| Number of Observations: | 460 | Number of Variables: | 2 |
| Degrees of Freedom: | 458 | Akaike's Information Criterion (AIC) [2]: | 3287.210420 |
| Multiple R-Squared [2]: | 0.242331 | Adjusted R-Squared [2]: | 0.240676 |
| Joint F-Statistic [3]: | 146.485414 | Prob(>F), (1,458) degrees of freedom: | 0.000000* |
| Joint Wald Statistic [4]: | 66.462176 | Prob(>chi-squared), (1) degrees of freedom: | 0.000000* |

- **Residuals ~ N (0, σ)**
  - N: normally distributed
    - Jarque-Bera p-value > 0.05
  - 0: mean = zero
  - σ: constant variance
    - Breusch-Pagen p-value > 0.05

| | | | |
|---|---|---|---|
| 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* |

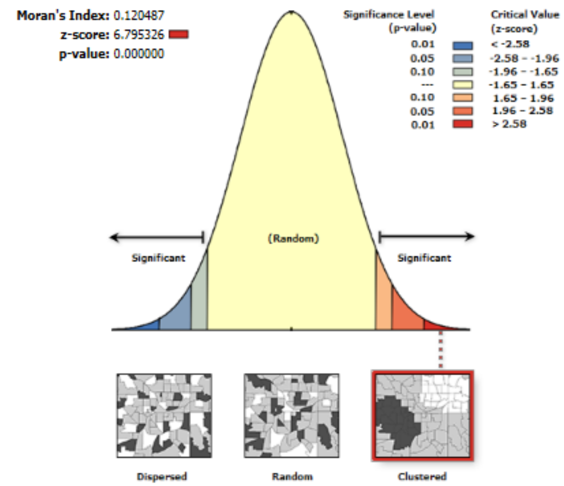
- **Residuals independent**
  - Moran's I p-value > 0.05




# Residuals independent?
# No.

## Spatial Autocorrelation Report




Given the z-score of 6.80, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

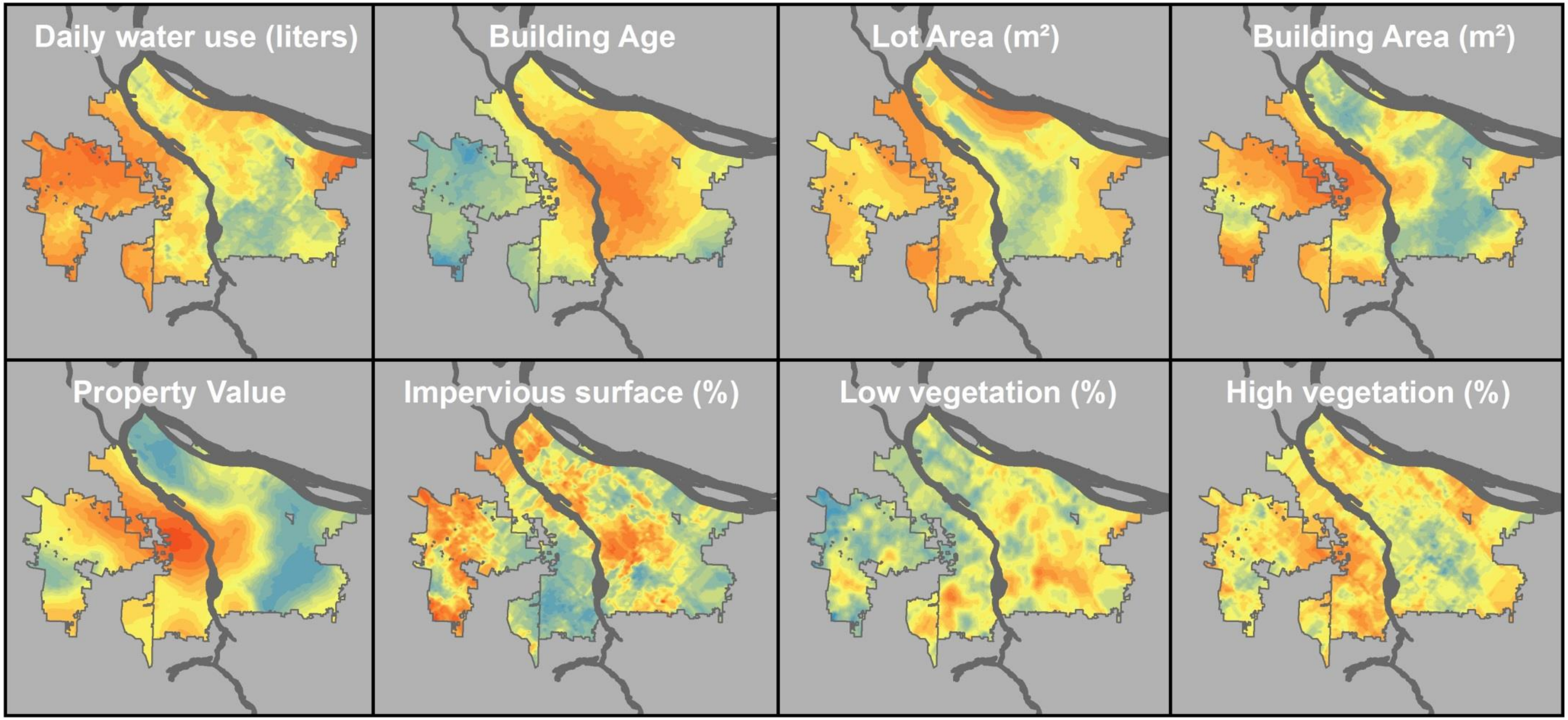
## Global Moran's I Summary

| | |
|---|---|
| **Moran's Index:** | 0.120487 |
| **Expected Index:** | -0.002179 |
| **Variance:** | 0.000326 |
| **z-score:** | 6.795326 |
| **p-value:** | 0.000000 |

## Dataset Information

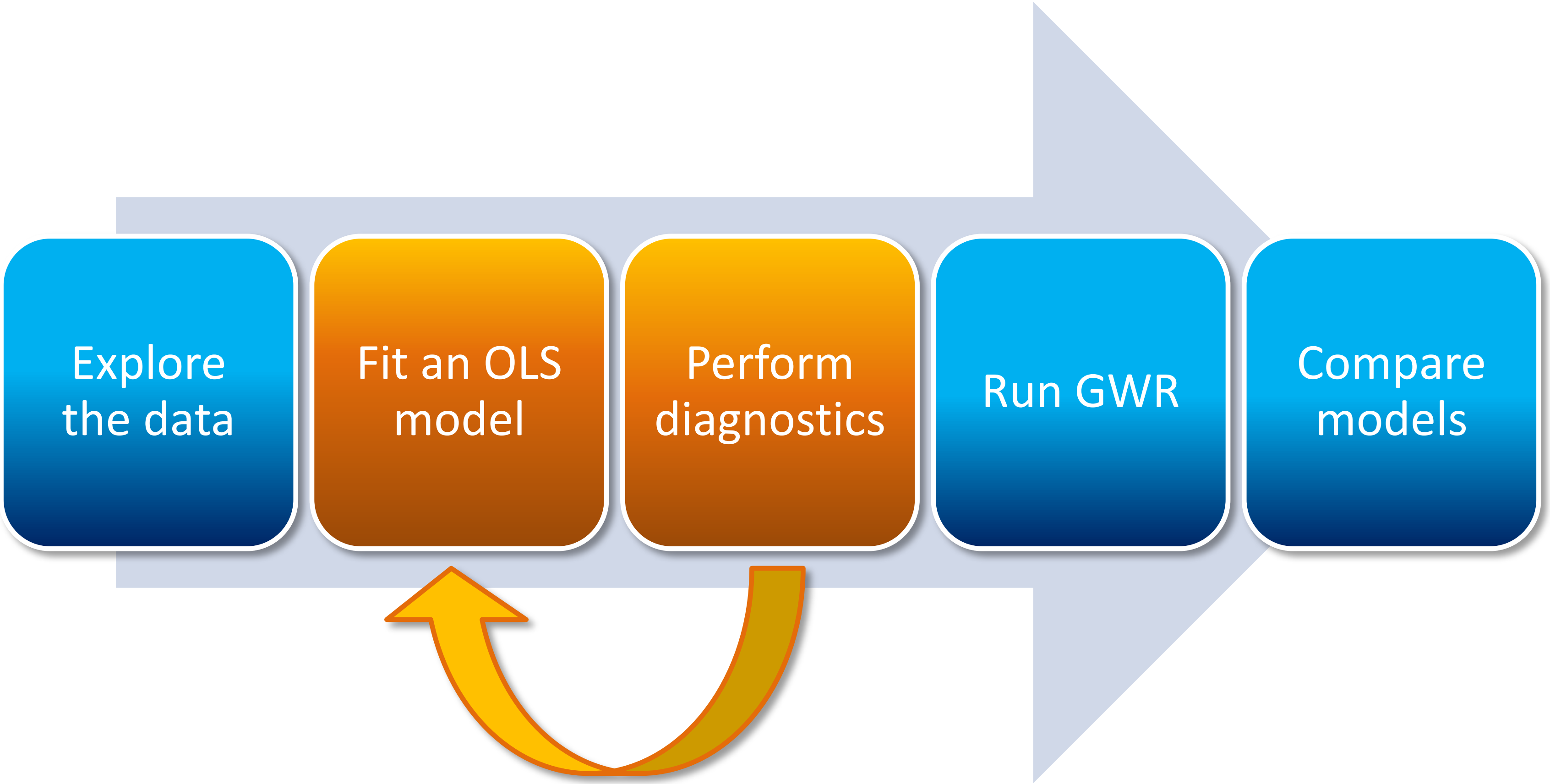
| | |
|---|---|
| **Input Feature Class:** | OLS_bldg |
| **Input Field:** | RESIDUAL |
| **Conceptualization:** | INVERSE_DISTANCE |
| **Distance Method:** | EUCLIDEAN |
| **Row Standardization:** | False |
| **Distance Threshold.:** | 11948.6849698 |
| **Weights Matrix File:** | None |

# What combination of variables makes a properly specified OLS model?



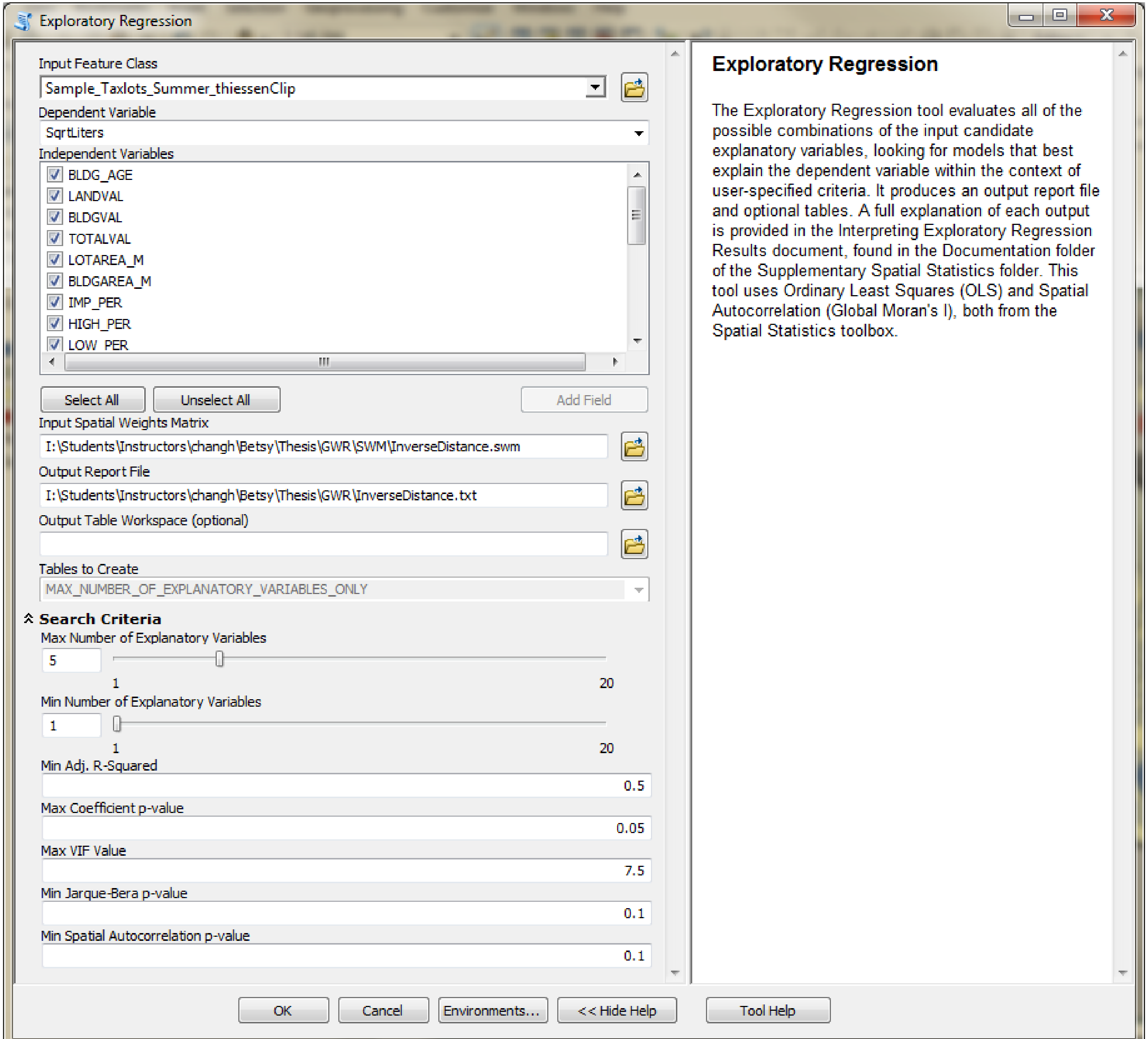

7! = 5040 possible models

# Model development




Iterative model development

WITH PYTHON


Exploratory Regression
Input Feature Class
Sample_Taxlots_Summer_thiessenClip
Dependent Variable
SqrtLiters
Independent Variables
BLDG_AGE
LANDVAL
BLDGVAL
TOTALVAL
LOTAREA_M
BLDGAREA_M
IMP_PER
HIGH_PER
LOW_PER
Select All
Unselect All
Add Field
Input Spatial Weights Matrix
I:\Students\Instructors\changh\Betsy\Thesis\GWR\SWM\InverseDistance.swm
Output Report File
I:\Students\Instructors\changh\Betsy\Thesis\GWR\InverseDistance.txt
Output Table Workspace (optional)
Tables to Create
MAX_NUMBER_OF_EXPLANATORY_VARIABLES_ONLY
Search Criteria
Max Number of Explanatory Variables
5
1
20
Min Number of Explanatory Variables
1
1
20
Min Adj. R-Squared
0.5
Max Coefficient p-value
0.05
Max VIF Value
7.5
Min Jarque-Bera p-value
0.1
Min Spatial Autocorrelation p-value
0.1
OK
Cancel
Environments...
<< Hide Help
Tool Help
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.

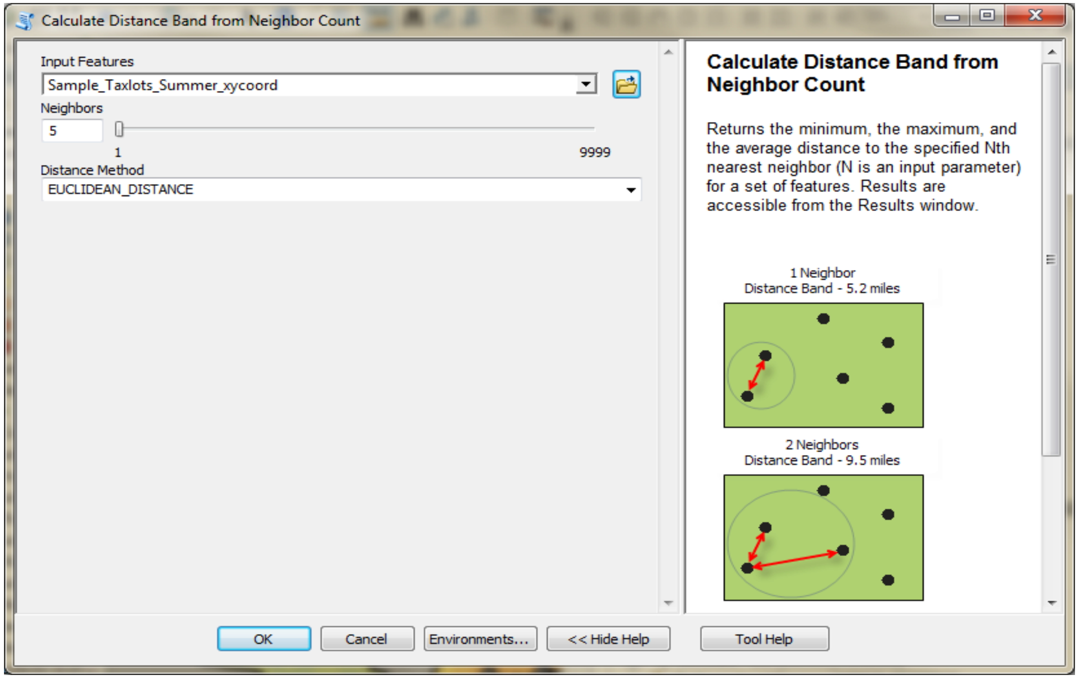
# Using Exploratory Regression in ArcGIS: Possible workflow

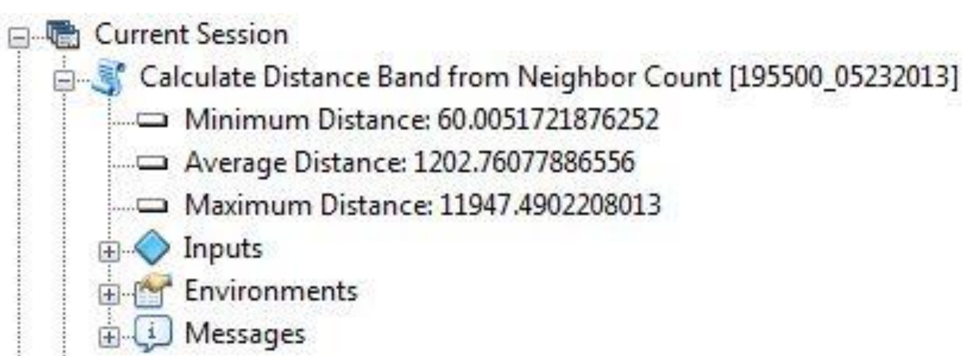
1. Calculate Distance Band from Neighbor Count
2. Incremental Spatial Autocorrelation
   – Bandwidth potentially useful for mapping clusters
3. Generate Spatial Weights Matrix
4. Exploratory Regression
   – Ordinary least squares regression
   – Geographically weighted regression

Source: ESRI Spatial Statistics Best Practices

# STEP 1

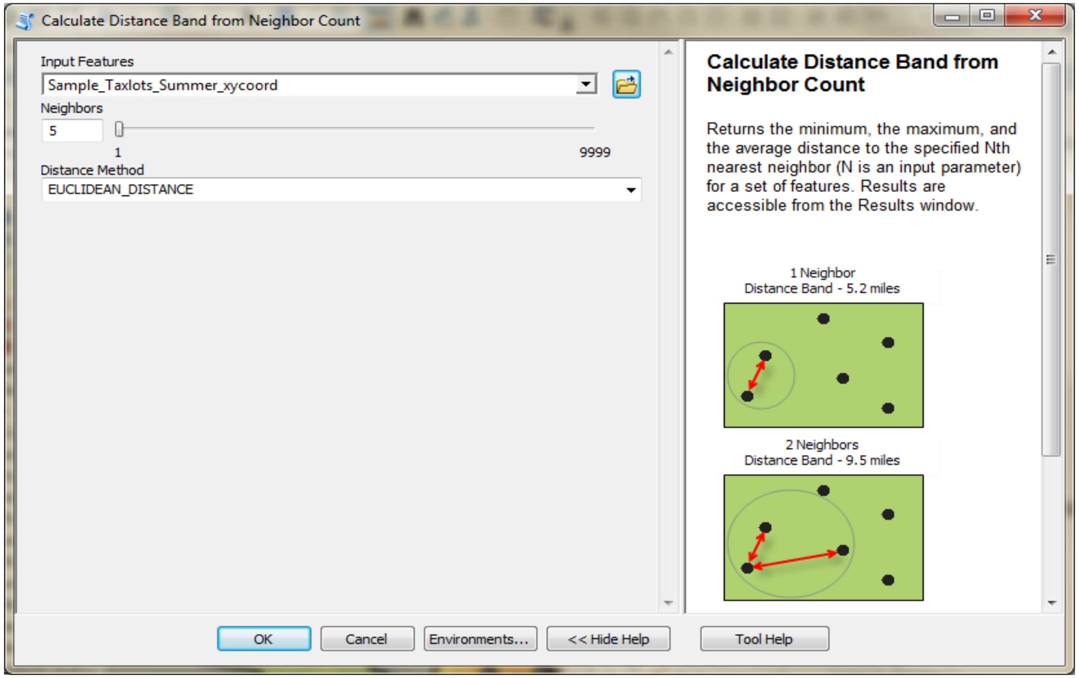
# Calculate distance band from Neighbor Count

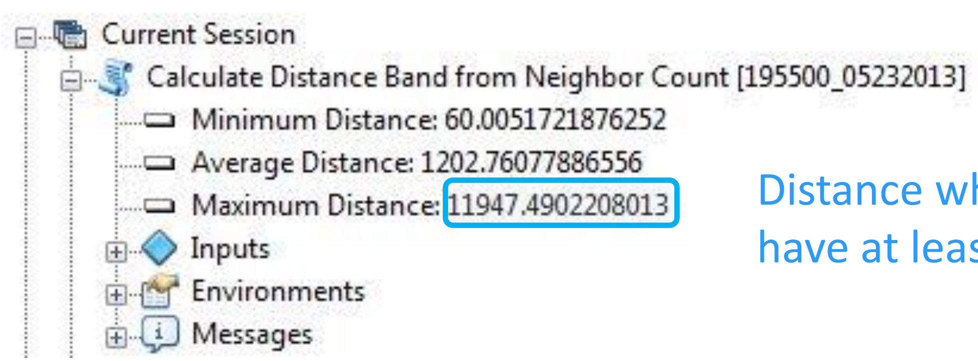


## STEP 1

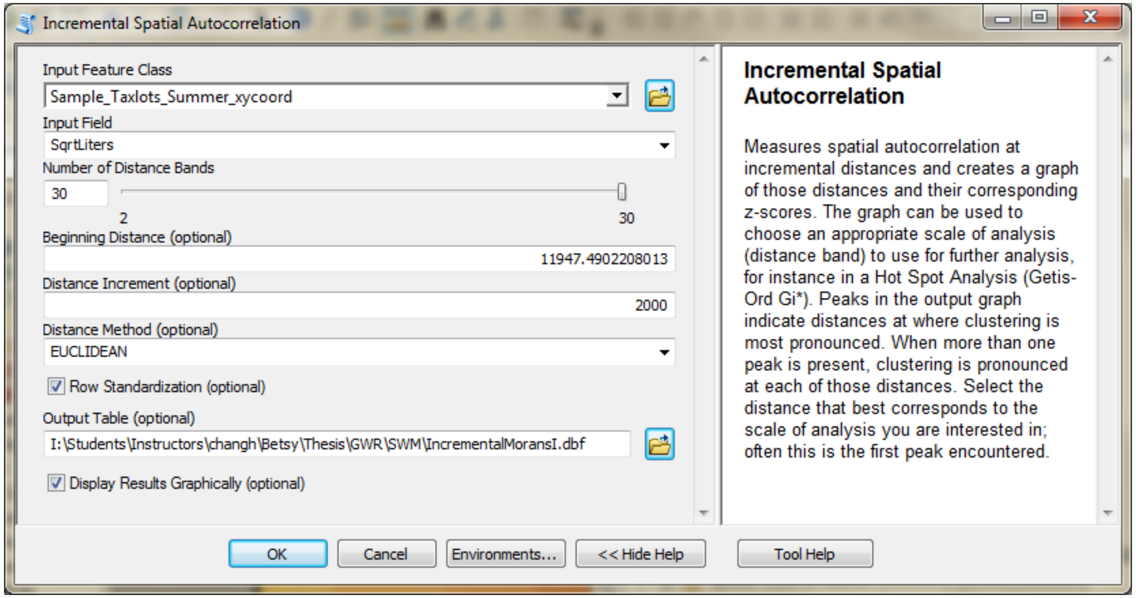
# Calculate distance band from Neighbor Count

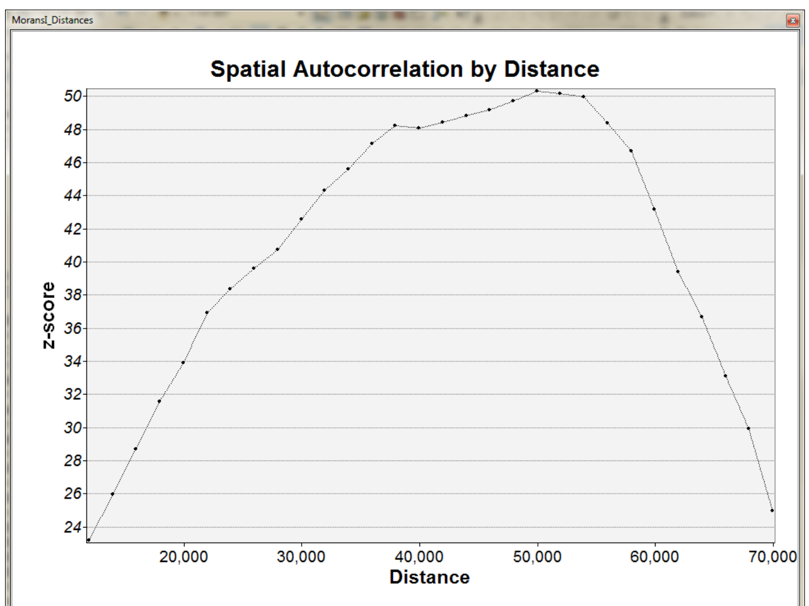



Distance where all features have at least 1 neighbor

## STEP 2
# Incremental Spatial Autocorrelation

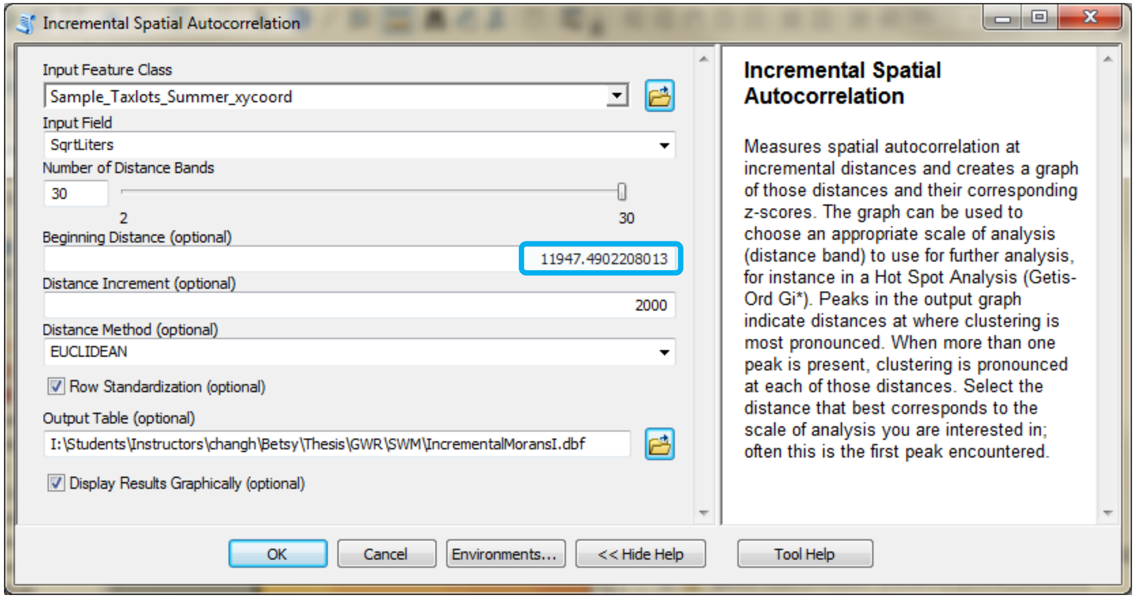


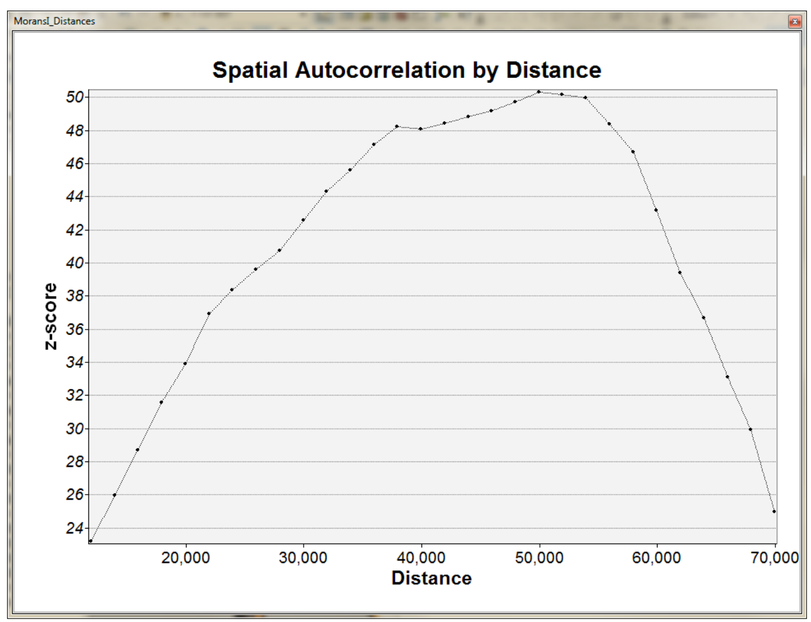



## STEP 2

# Incremental Spatial Autocorrelation

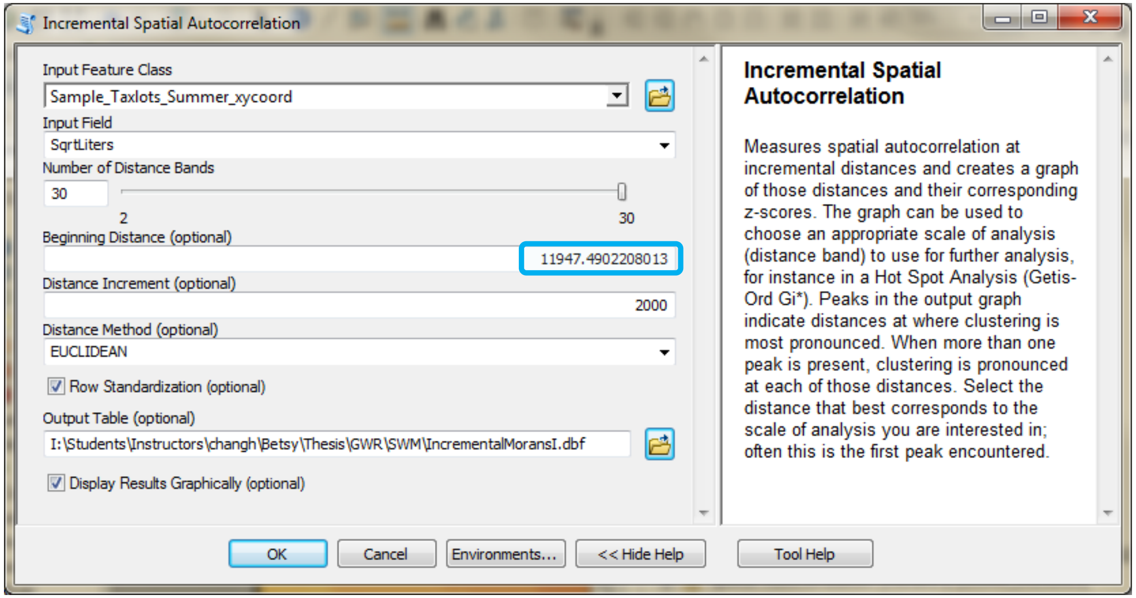
Beginning distance is output from Step 1 ➔

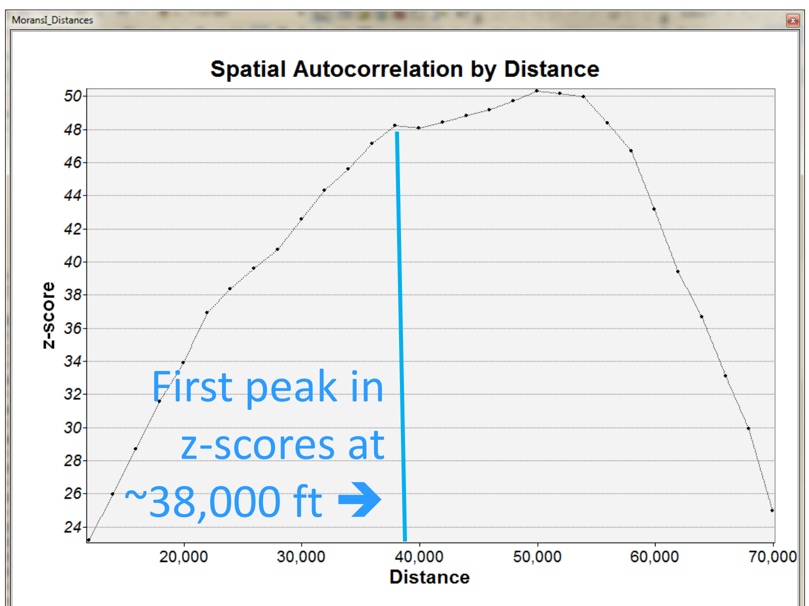


# STEP 2
# Incremental Spatial Autocorrelation

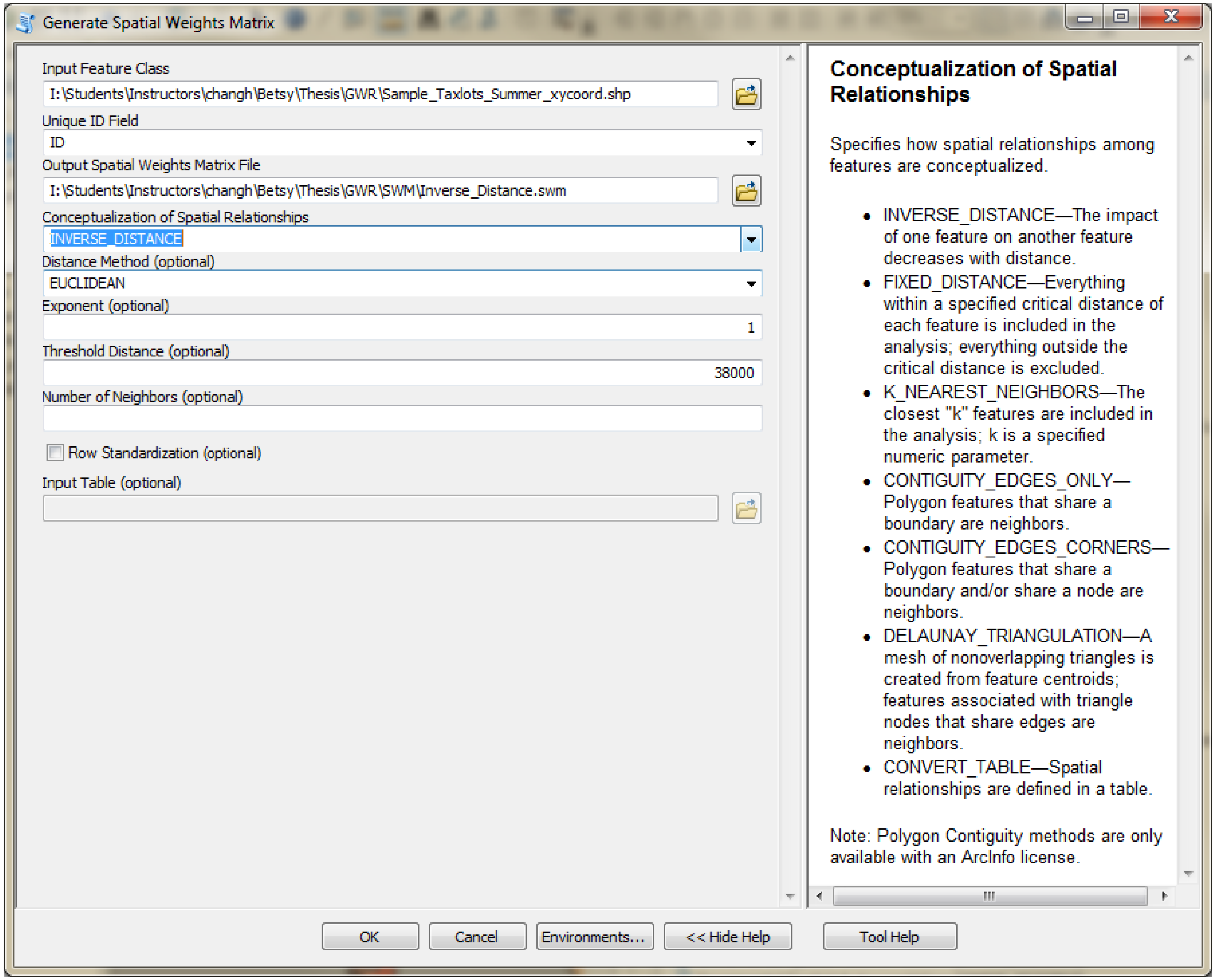
Beginning distance is output from Step 1 →



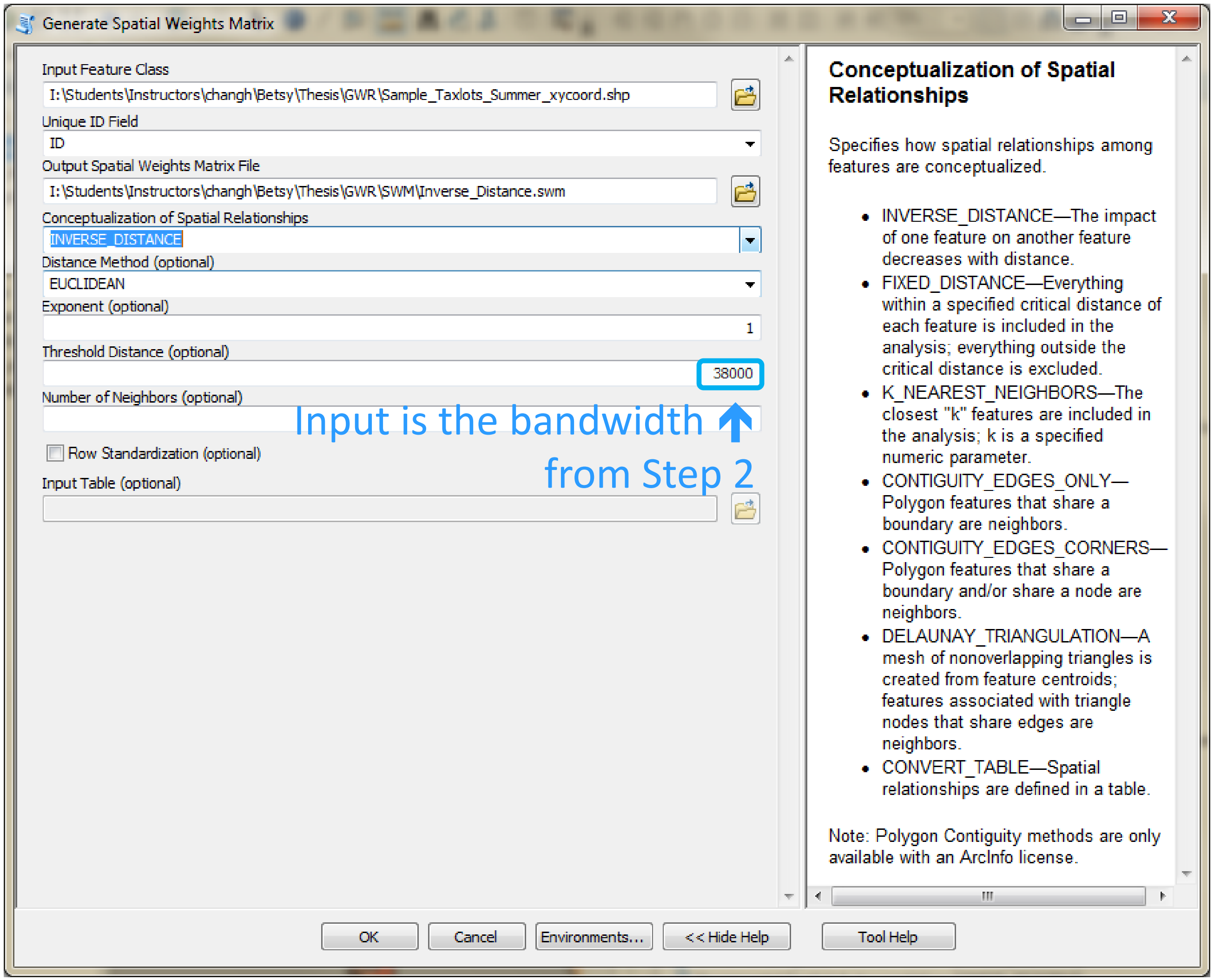



STEP 3

# Generate spatial weights matrix



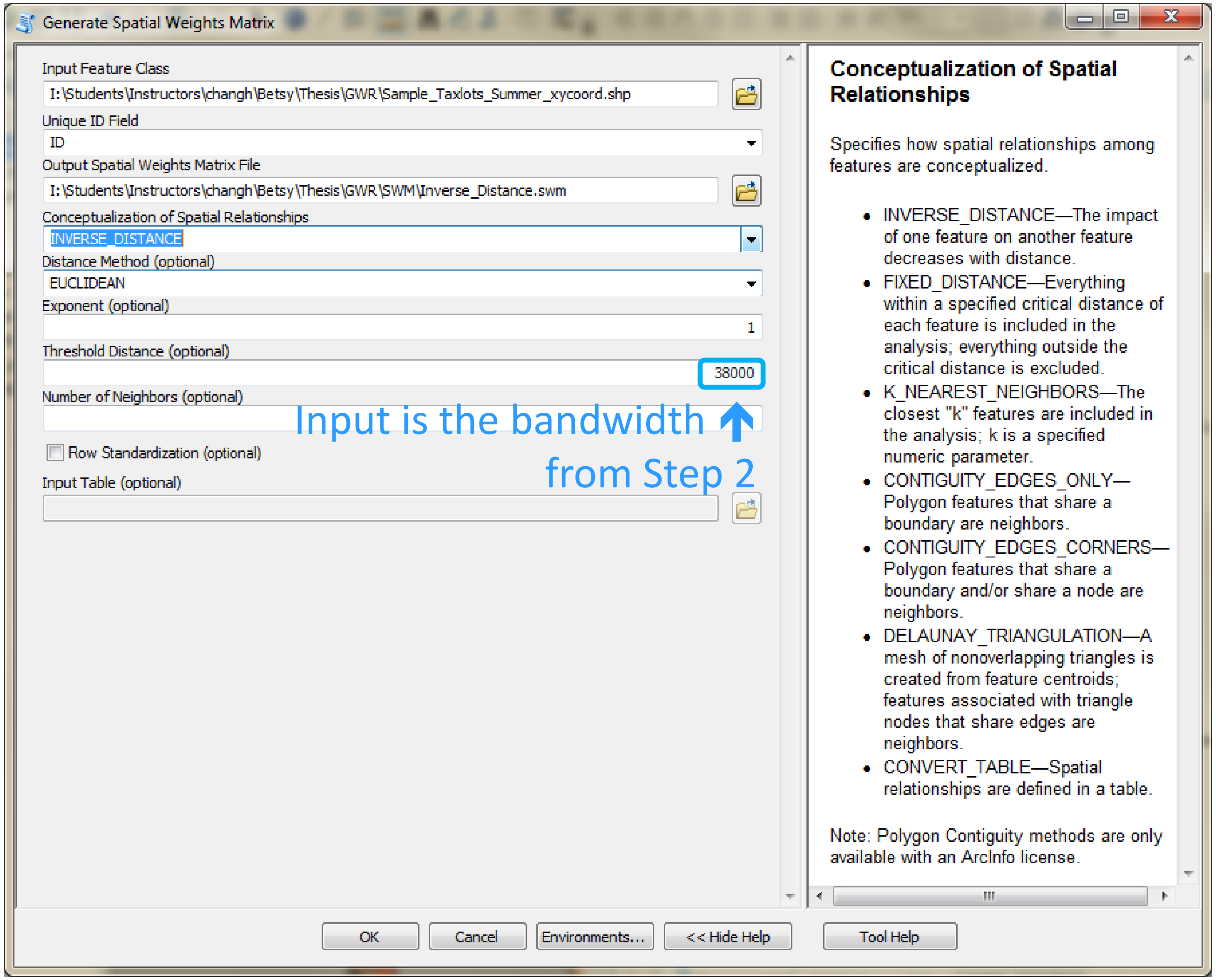
# STEP 3
# Generate spatial weights matrix



## STEP 3

# Generate spatial weights matrix

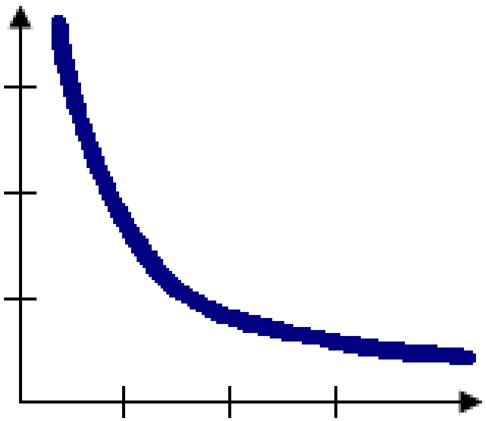
Output is a .swm file



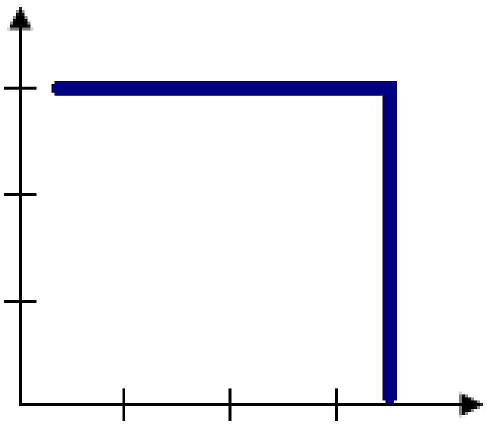

Input is the bandwidth from Step 2

STEP 3

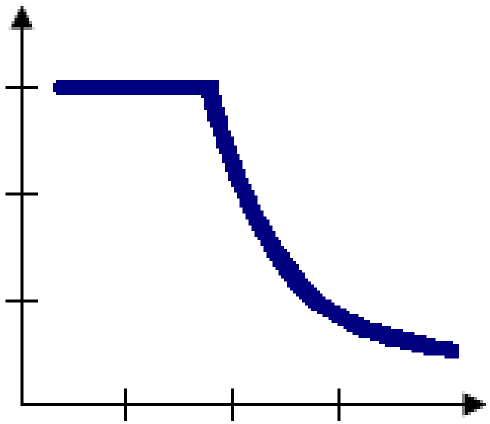
# Conceptualization of spatial relationships



Inverse distance



Fixed distance

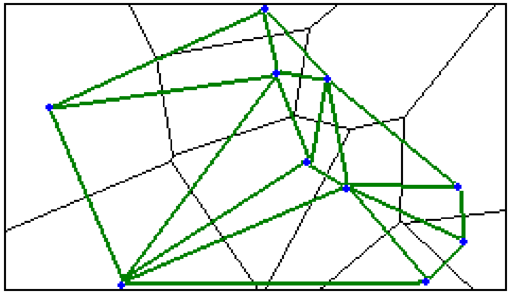


Zone of indifference

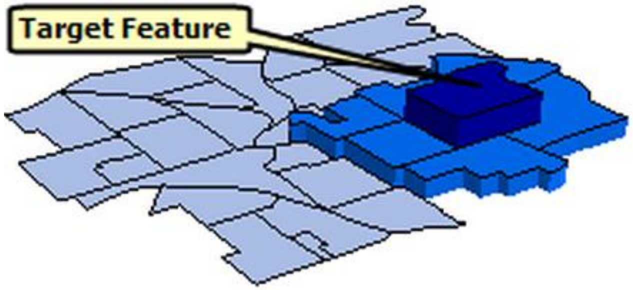
http://resources.arcgis.com/en/help/main/10.1/index.html#//005p00000005000000

STEP 3

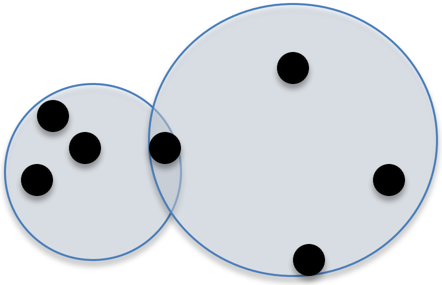
# Conceptualization of spatial relationships



Delaunay Triangulation



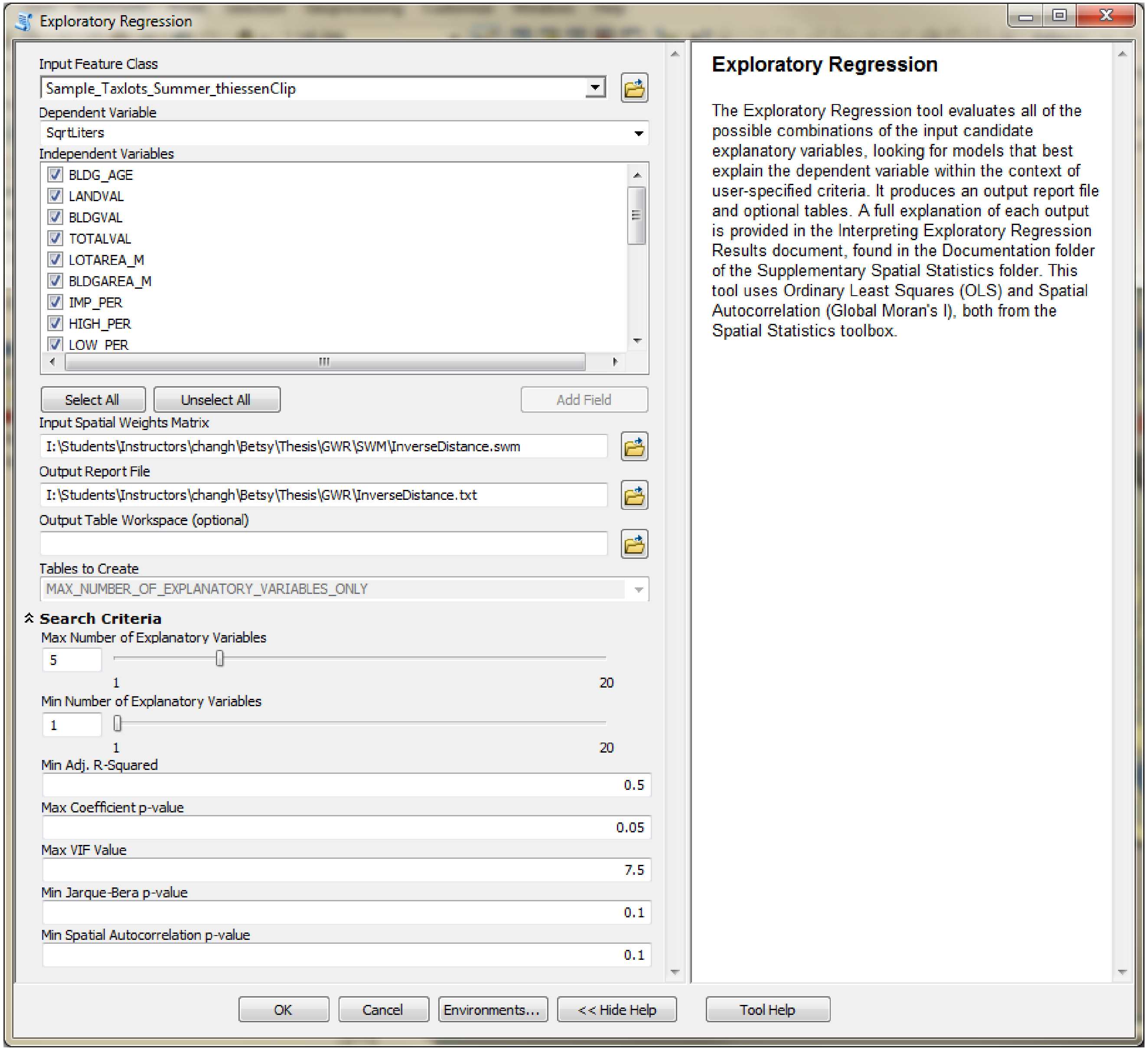

Polygon Contiguity



K nearest neighbors

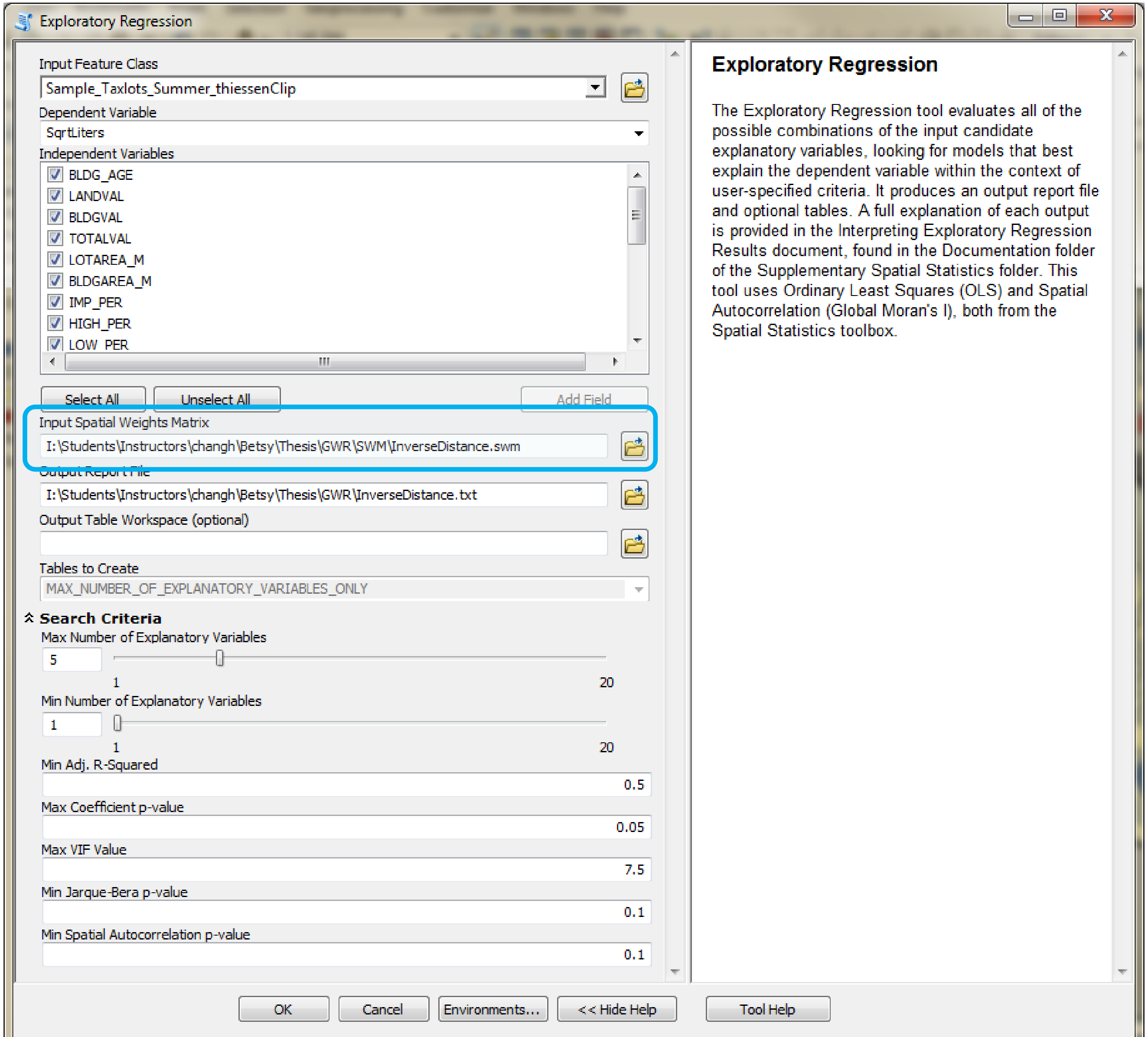
See also: Space-time clustering, Grouping Analysis tool

# STEP 4
# Exploratory regression



# STEP 4
# Exploratory regression

.swm file →

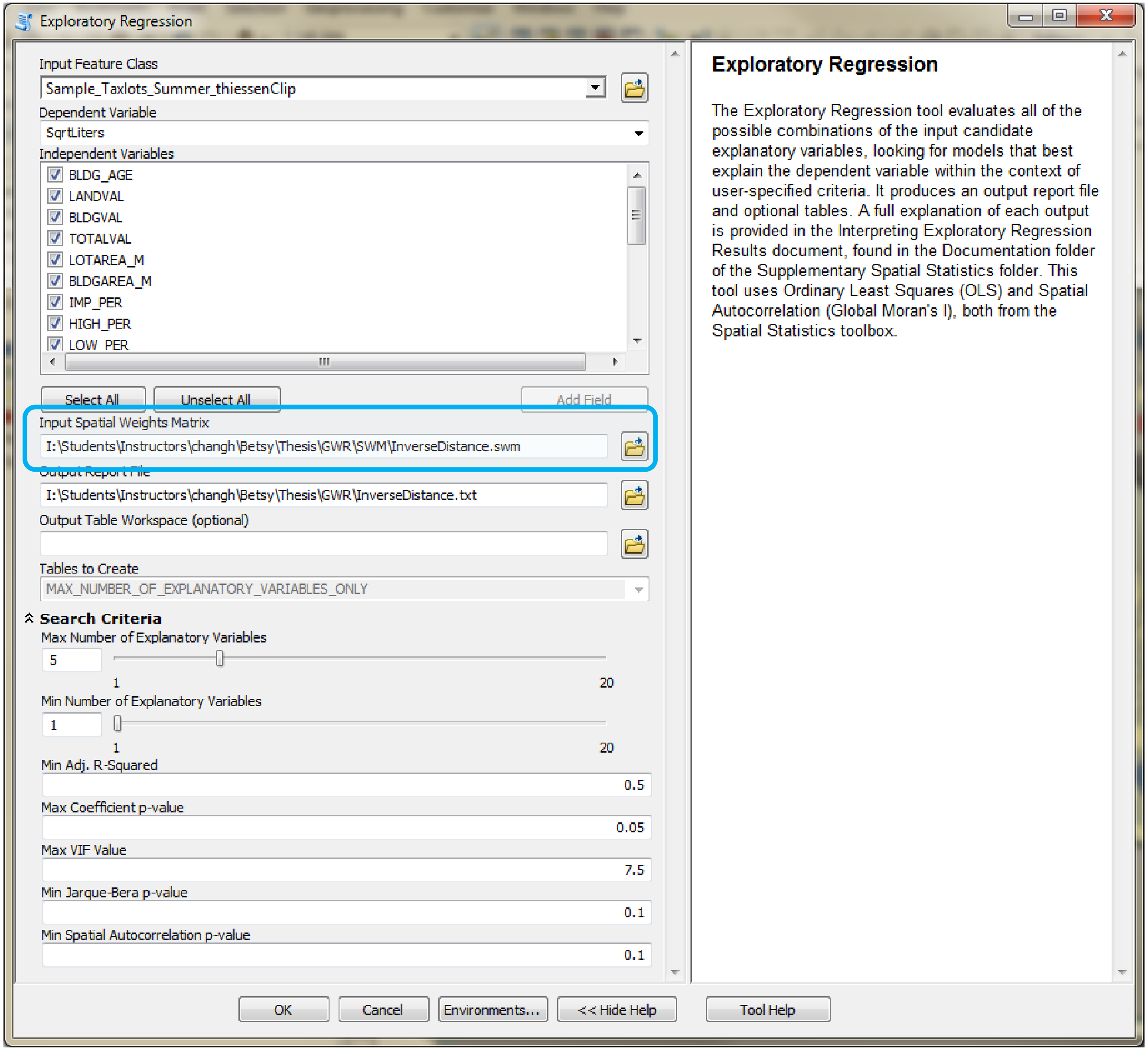


# STEP 4

# Exploratory regression

Candidate explanatory variables

.swm file →



STEP 4

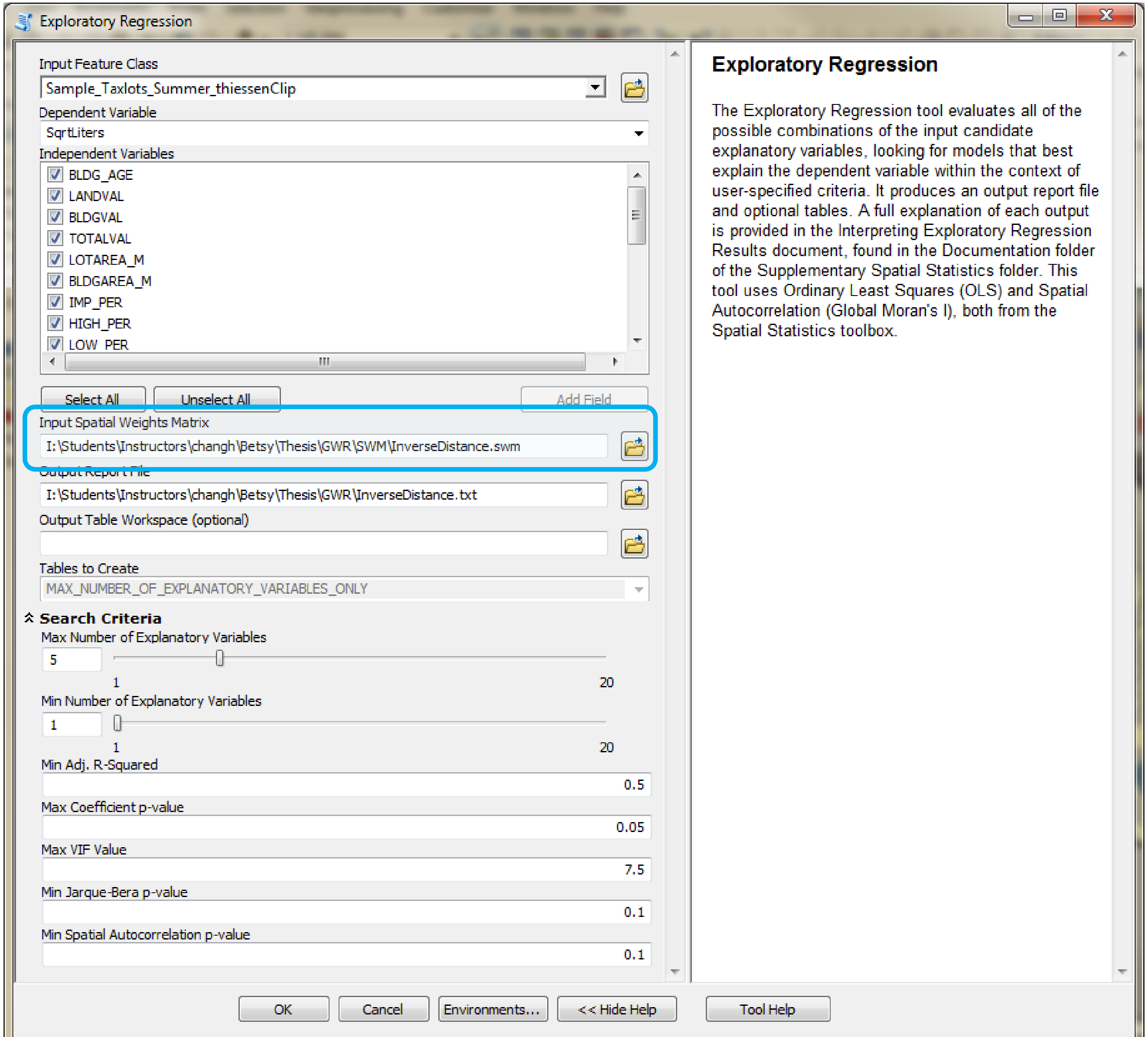
# Exploratory regression

Candidate explanatory variables

.swm file →

Min/max number of variables per model

Significance test p-values



**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 3338.06 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 1.36 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 3231.96 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**  +LOTAREA_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?

************* Exploratory Regression Global Summary (SQRTLITERS) *************

Percentage of Search Criteria Passed

| Search Criterion | Cutoff | Trials | # Passed | % Passed |
|---|---|---|---|---|
| Min Adjusted R-Squared | > 0.50 | 1585 | 0 | 0.00 |
| Max Coefficient p-value | < 0.05 | 1585 | 283 | 17.85 |
| Max VIF Value | < 7.50 | 1585 | 1351 | 85.24 |
| Min Jarque-Bera p-value | > 0.10 | 1585 | 3 | 0.19 |
| Min Moran's I p-value | > 0.10 | 18 | 6 | 33.33 |

### Which variables are consistently significant?

Summary of Variable Significance

| Variable | % Significant |
|---|---|
| BLDG_AGE | 100.00 |
| LANDVAL | 80.60 |
| BLDGVAL | 65.12 |
| TOTALVAL | 77.76 |
| LOTAREA_M | 72.95 |
| BLDGAREA_M | 100.00 |
| IMP_PER | 42.70 |
| HIGH_PER | 30.60 |
| LOW_PER | 33.63 |
| IMP_M | 97.51 |
| LOW_M | 48.22 |
| HIGH_M | 43.24 |

**S T E P 4**

# Exploratory regression

## Which models have normal residuals?

Summary of Residual Normality

| JB | R2 | AICc | BP | VIF | MI | Model |
|---|---|---|---|---|---|---|
| 0.124366 | 0.173228 | 3331.519326 | 0.000003 | 7.909751 | 0.000000 | +BLDGVAL*** +LOTAREA_M*** -LOW_PER -LOW_M |
| 0.110102 | 0.172788 | 3330.720833 | 0.000000 | 2.976807 | 0.000000 | +BLDGVAL*** +LOTAREA_M*** -LOW_M*** |
| 0.102163 | 0.171989 | 3331.164556 | 0.000023 | 1.046833 | 0.000000 | +BLDGVAL*** +LOTAREA_M*** -LOW_PER*** |

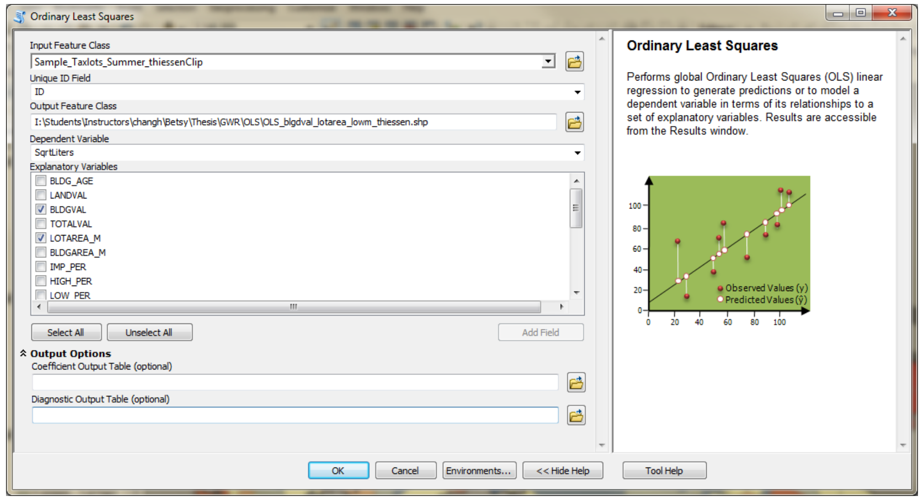
---

## Which models have spatially uncorrelated residuals?

Summary of Residual Autocorrelation

| MI | R2 | AICc | JB | BP | VIF | Model |
|---|---|---|---|---|---|---|
| 0.223620 | 0.330270 | 3234.618079 | 0.000000 | 0.014348 | 6.724090 | -BLDG_AGE*** -BLDGVAL** +TOTALVAL** +BLDGAREA_M*** |
| 0.320787 | 0.334135 | 3231.956051 | 0.000000 | 0.001452 | 1.596901 | -BLDG_AGE*** +LANDVAL** +BLDGAREA_M*** +IMP_M** |
| 0.213209 | 0.341931 | 3227.588542 | 0.000000 | 0.006685 | 283.707489 | -BLDG_AGE*** -LANDVAL*** -BLDGVAL*** +TOTALVAL*** |

# OLS regression: Estimated coefficients




Summary of OLS Results

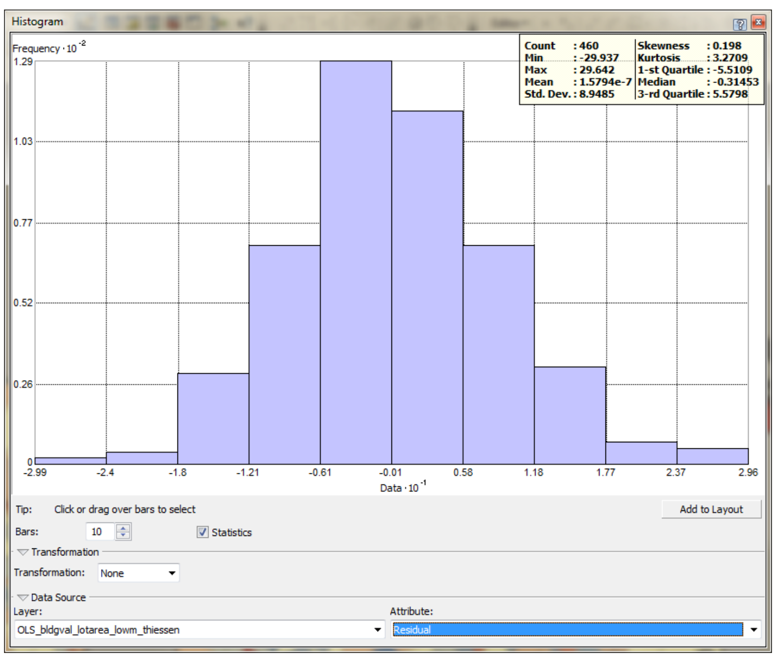
| Variable | Coefficient | StdError | t-Statistic | Probability | Robust_SE | Robust_t | Robust_Pr | VIF [1] |
|---|---|---|---|---|---|---|---|---|
| Intercept | 22.341917 | 1.078627 | 20.713298 | 0.000000* | 1.344188 | 16.621123 | 0.000000* | -------- |
| BLDGVAL | 0.000025 | 0.000005 | 5.087609 | 0.000001* | 0.000006 | 4.471951 | 0.000012* | 1.051748 |
| LOTAREA_M | 0.009847 | 0.001469 | 6.702138 | 0.000000* | 0.001925 | 5.115796 | 0.000001* | 2.976807 |
| LOW_M | -0.008885 | 0.002490 | -3.567676 | 0.000411* | 0.002627 | -3.382256 | 0.000795* | 2.951698 |

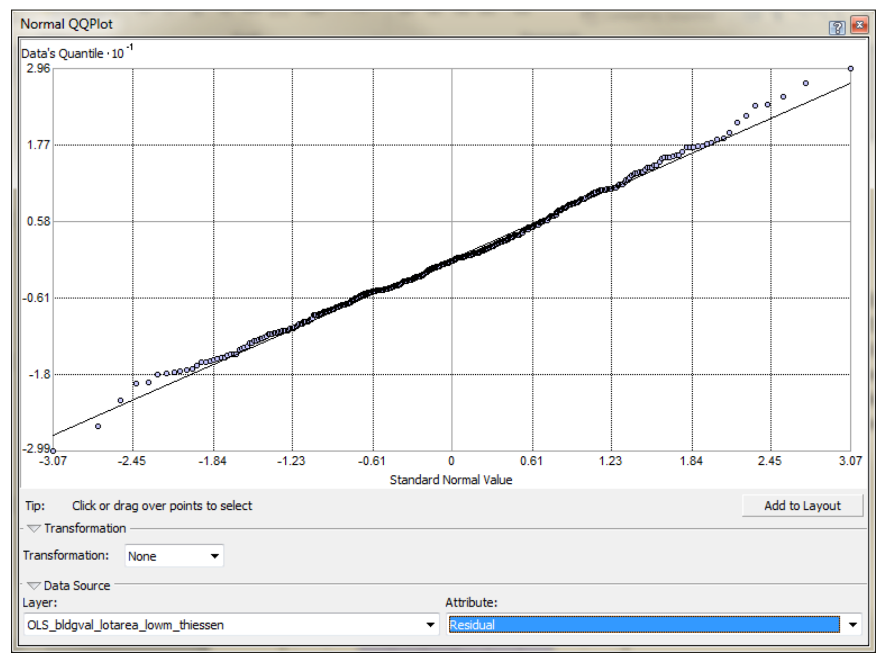
$$Y = 22.34 + 0.00025(BLDGVAL) + 0.00985(LOTAREA\_M) - 0.008885(LOW\_M) + error$$

# OLS regression: Residual normality

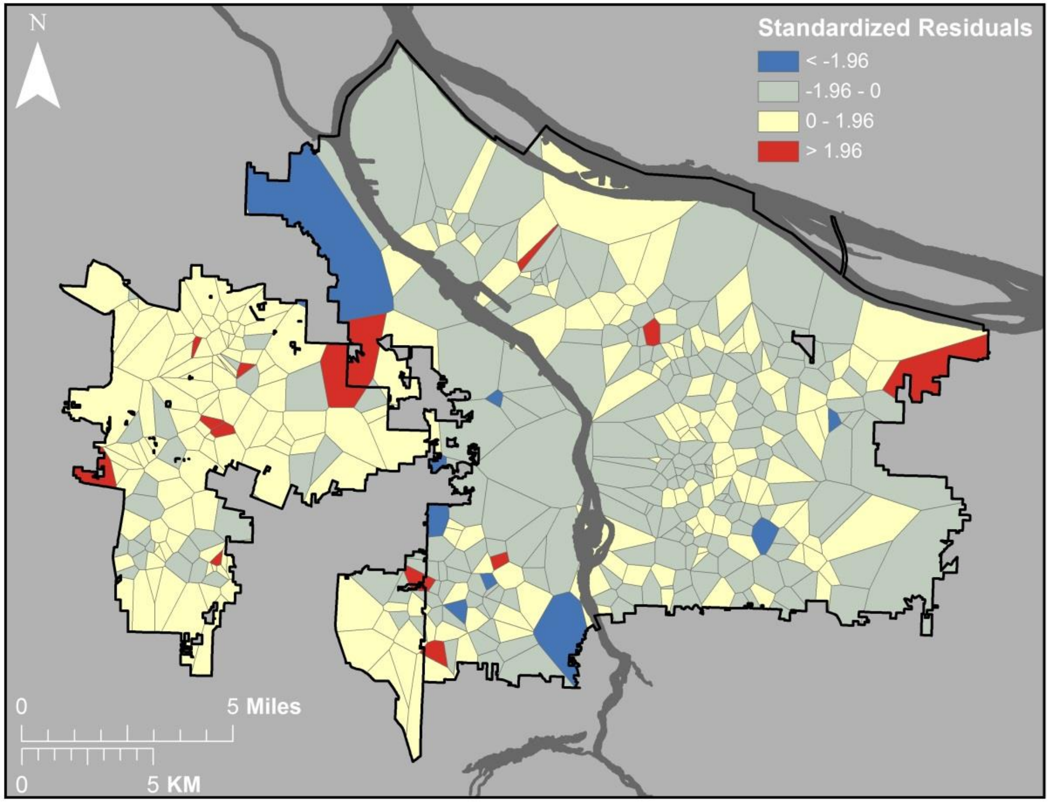
OLS Diagnostics

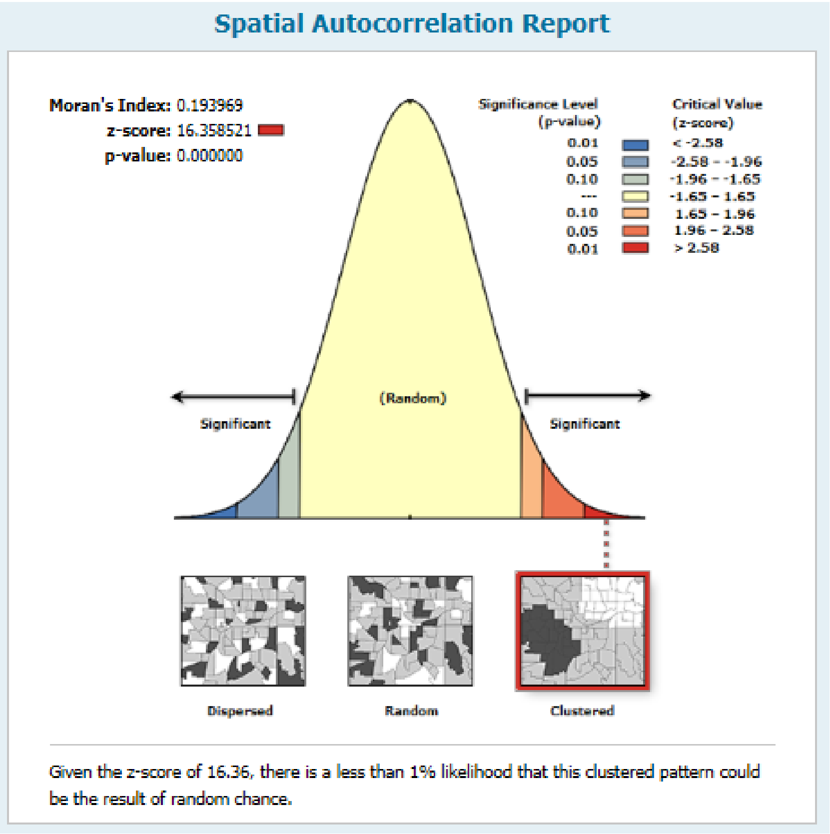
| | | | |
|---|---|---|---|
| Number of Observations: | 460 | Number of Variables: | 4 |
| Degrees of Freedom: | 456 | Akaike's Information Criterion (AIC) [2]: | 3328.588674 |
| Multiple R-Squared [2]: | 0.178194 | Adjusted R-Squared [2]: | 0.172788 |
| Joint F-Statistic [3]: | 32.958560 | Prob(>F), (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 |



# OLS regression: Residual dependence

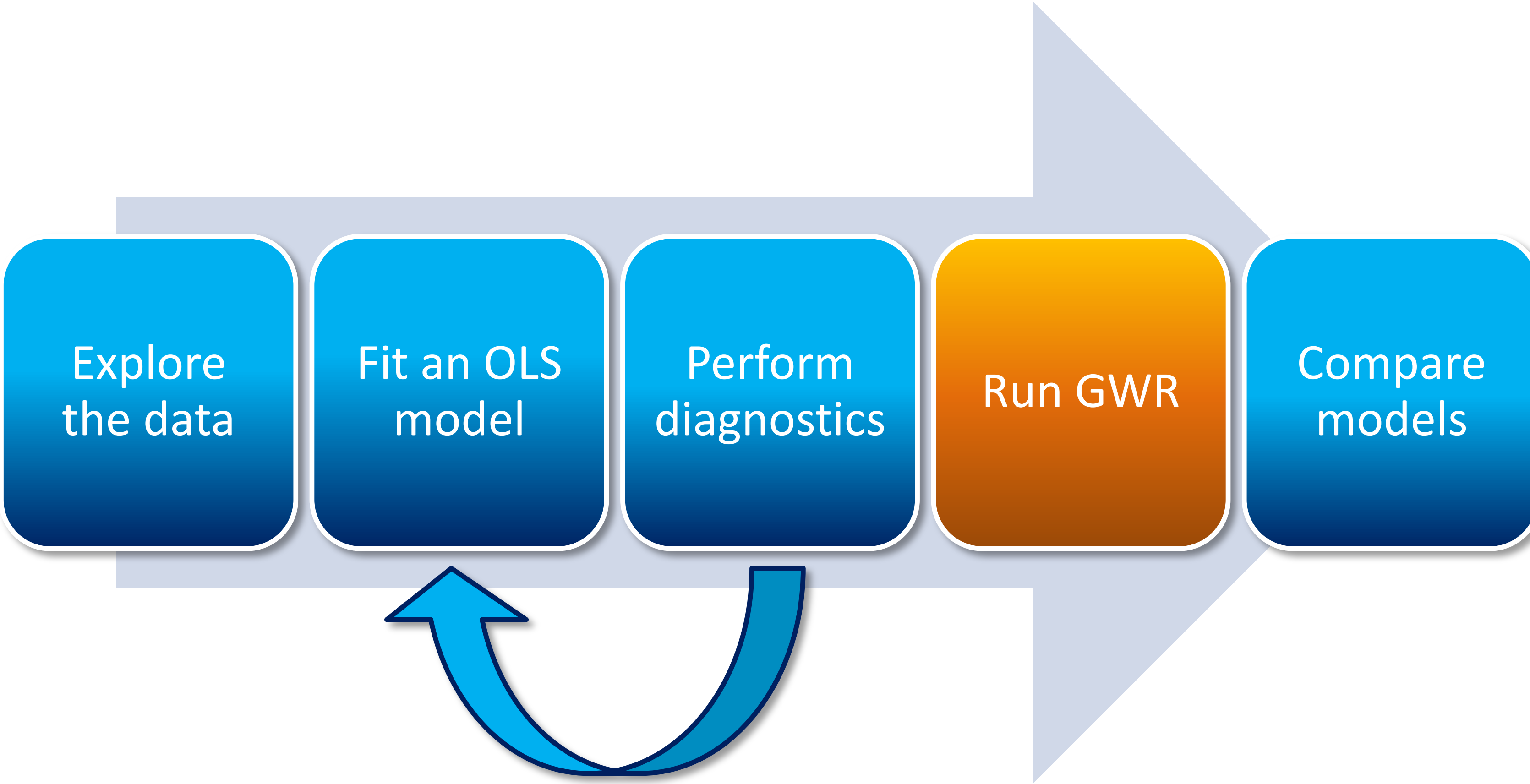



## Global Moran's I Summary

| | |
|---|---|
| Moran's Index: | 0.193969 |
| Expected Index: | -0.002179 |
| Variance: | 0.000144 |
| z-score: | 16.358521 |
| p-value: | 0.000000 |

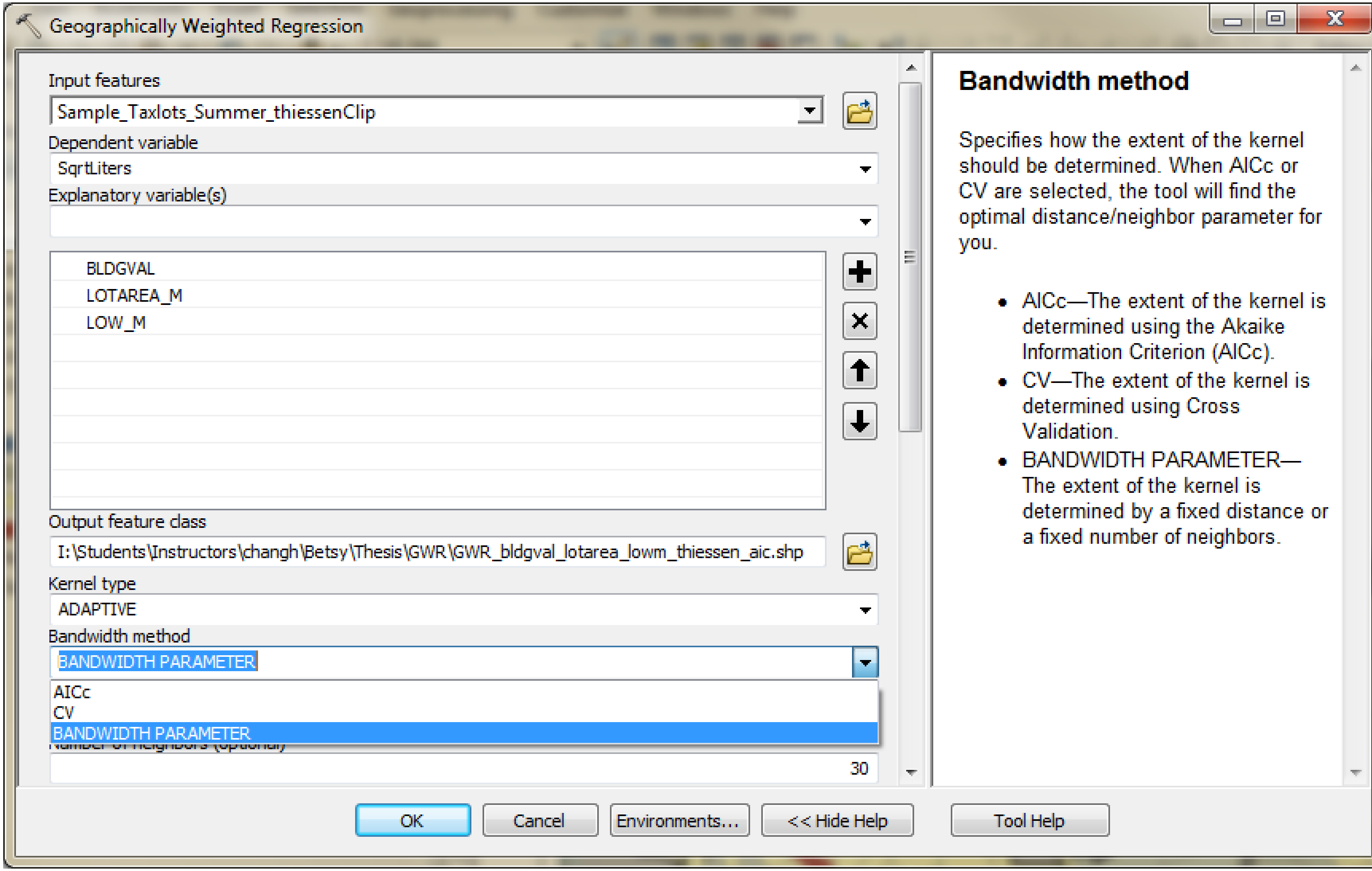
# Model development




Iterative model development

WITH PYTHON

# GWR: Setting kernel and bandwidth



# GWR: Output

Messages
- Executing: GeographicallyWeightedRegression
- Start Time: Tue May 28 16:37:36 2013
- Neighbors : 164
- ResidualSquares : 25970.219444414834
- EffectiveNumber : 33.793550676073622
- Sigma : 7.8059865944561873
- AICc : 3218.5804496988549
- R2 : 0.41932611645259799
- R2Adjusted : 0.37464739688702065

Table

GWR_bldgval_lotarea_lowm_thiessen_aic

| Observed | Cond | LocalR2 | Predicted | Intercept | C1_BLDG | C2_LOTA | C3_LO | Residual | StdError | StdErr_Int | StdErrC1 | StdErrC2 | StdErrC3 | StdResi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 53.739534 | 10.220 | 0.208487 | 53.572759 | 19.87490 | 0.000034 | 0.010397 | -0.0094 | 0.166775 | 2.862581 | 2.337176 | 0.000011 | 0.002891 | 0.00416 | 0.05826 |
| 23.890516 | 7.2591 | 0.215906 | 27.938022 | 16.87395 | 0.000033 | 0.008712 | -0.0011 | -4.04750 | 7.569381 | 2.547647 | 0.000011 | 0.003393 | 0.006247 | -0.5347 |
| 33.052823 | 8.8667 | 0.168227 | 33.010089 | 17.64097 | 0.000054 | 0.003033 | 0.00435 | 0.042734 | 7.127231 | 3.175386 | 0.000016 | 0.003929 | 0.006976 | 0.00599 |
| 63.504146 | 9.7458 | 0.263538 | 41.802754 | 13.37000 | 0.00008 | 0.004307 | 0.00920 | 21.70139 | 6.518928 | 3.464891 | 0.000017 | 0.004406 | 0.007025 | 3.32898 |
| 38.798366 | 8.5593 | 0.111953 | 32.279123 | 22.30576 | 0.000017 | 0.009133 | -0.0087 | 6.519243 | 7.269473 | 2.14511 | 0.000009 | 0.004395 | 0.007334 | 0.89679 |
| 44.229499 | 6.3124 | 0.16125 | 44.521959 | 27.86461 | 0.000022 | 0.00218 | 0.00594 | -0.29246 | 5.747762 | 1.692899 | 0.000007 | 0.002721 | 0.00676 | -0.0508 |
| 20.397856 | 6.9361 | 0.241764 | 29.772157 | 25.08383 | 0.000036 | 0.002785 | 0.00316 | -9.37430 | 7.594773 | 1.78285 | 0.000008 | 0.002709 | 0.005121 | -1.2343 |
| 45.690303 | 9.6165 | 0.229426 | 26.559803 | 14.67353 | 0.000067 | 0.005141 | 0.00815 | 19.1305 | 7.523589 | 3.292605 | 0.000014 | 0.004415 | 0.007004 | 2.54273 |
| 13.583328 | 8.9639 | 0.115431 | 25.634288 | 19.63844 | 0.000033 | 0.006154 | -0.0001 | -12.0509 | 7.558961 | 2.690557 | 0.000009 | 0.004108 | 0.006693 | -1.5942 |
| 24.759926 | 8.9757 | 0.090382 | 26.398506 | 22.65538 | 0.000017 | 0.008834 | -0.0055 | -1.63857 | 7.487637 | 2.061032 | 0.000009 | 0.004441 | 0.006892 | -0.2188 |

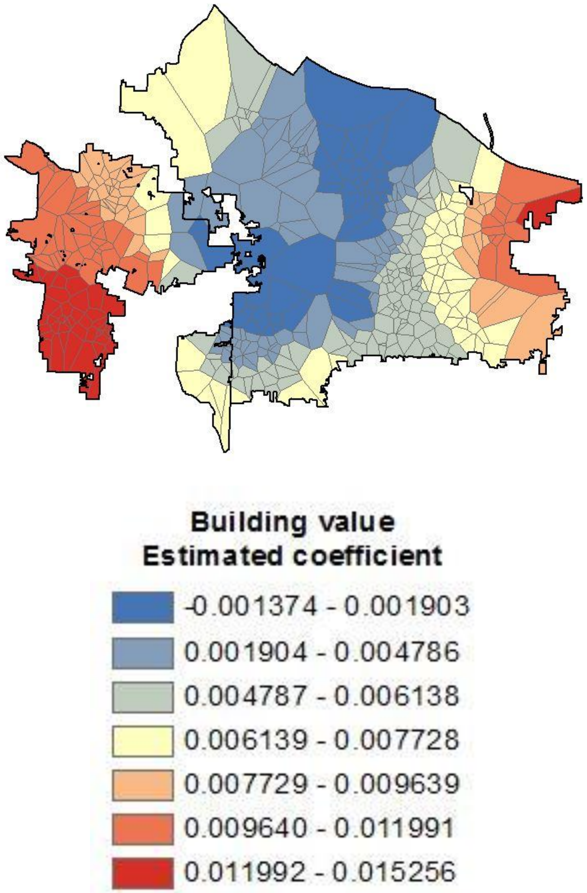
1 (0 out of 460 Selected)

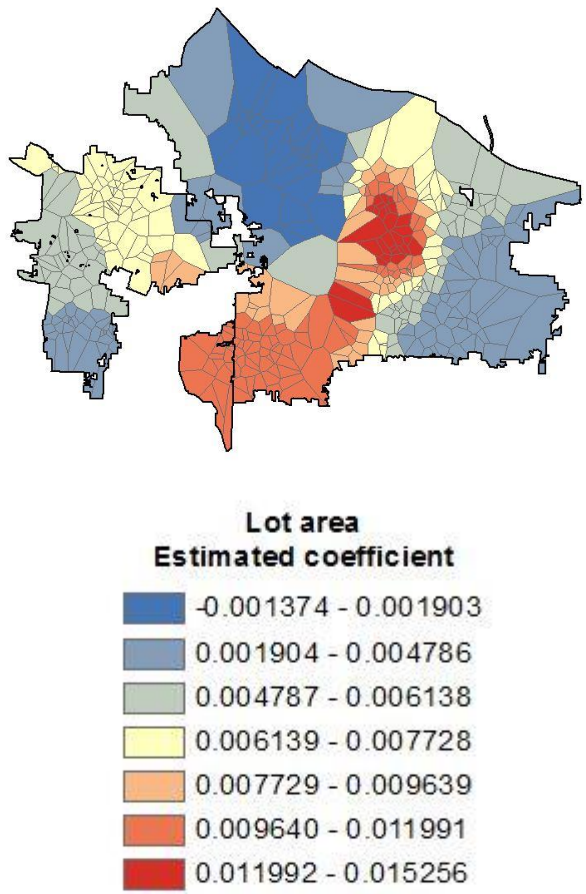
GWR_bldgval_lotarea_lowm_thiessen_aic

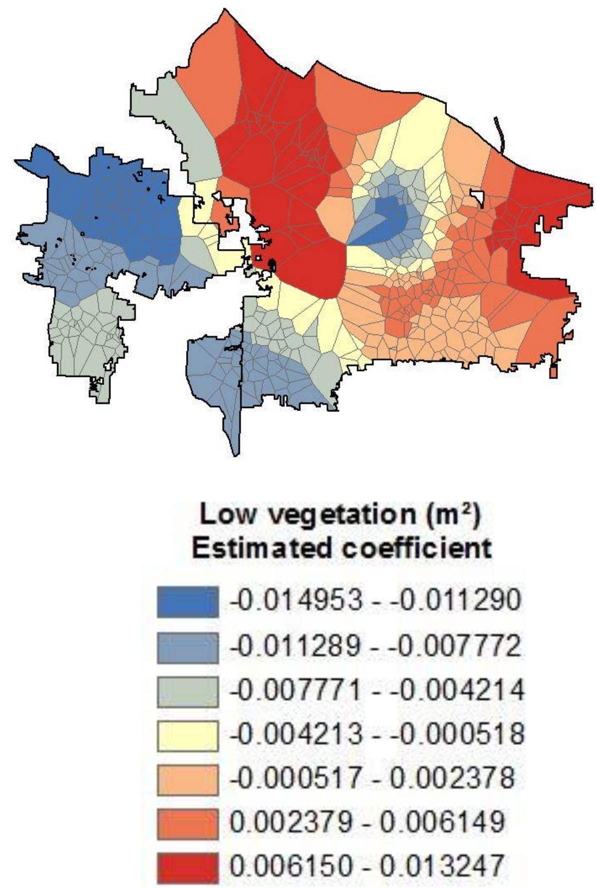
Full explanation of statistical output:

http://resources.esri.com/help/9.3/arcgisdesktop/com/gp_toolref/spatial_statistics_tools/interpreting_gwr_results.htm

# GWR: Coefficient estimates



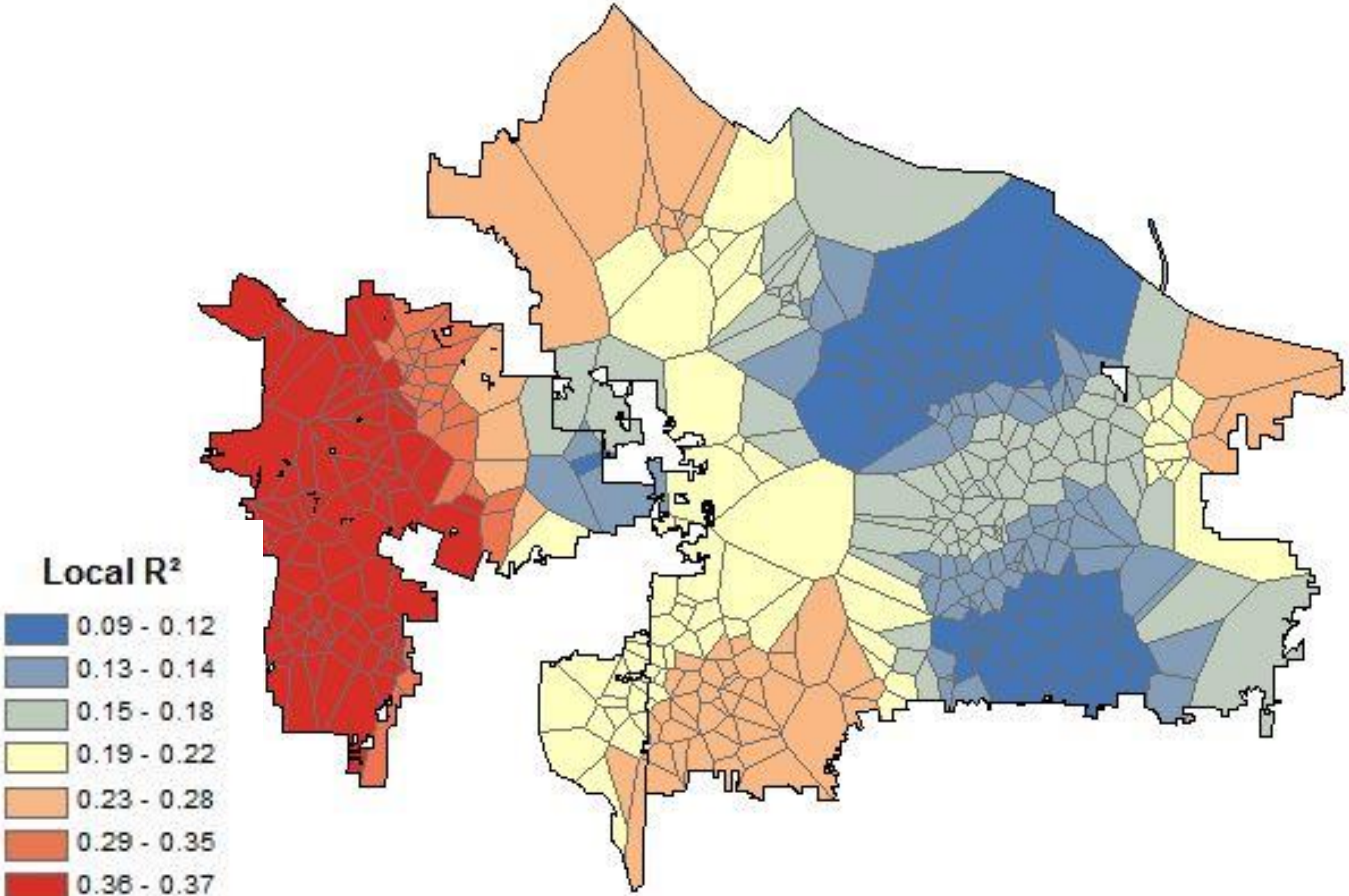




Global: 0.000025

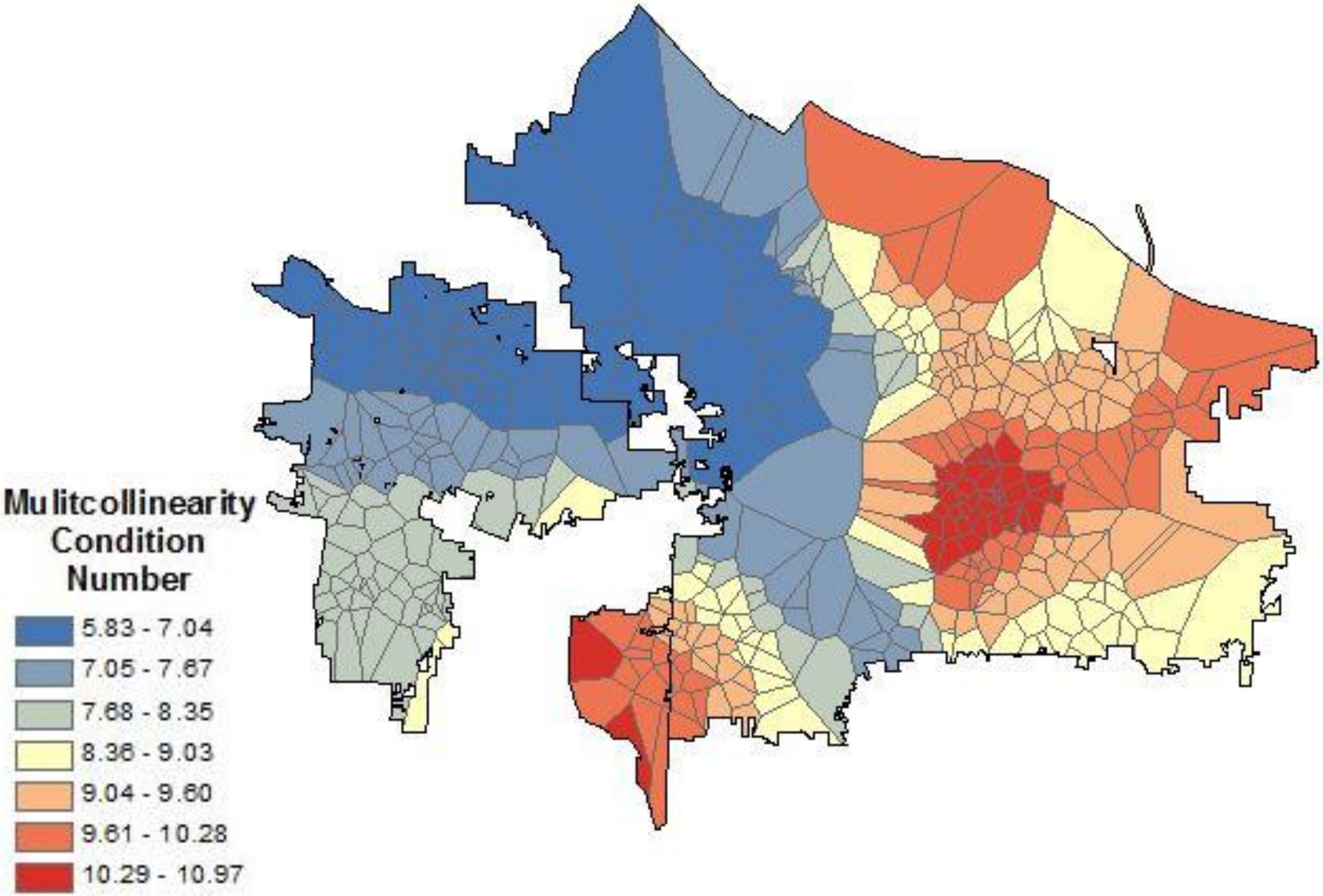
Global: 0.009847

Global: -0.008885

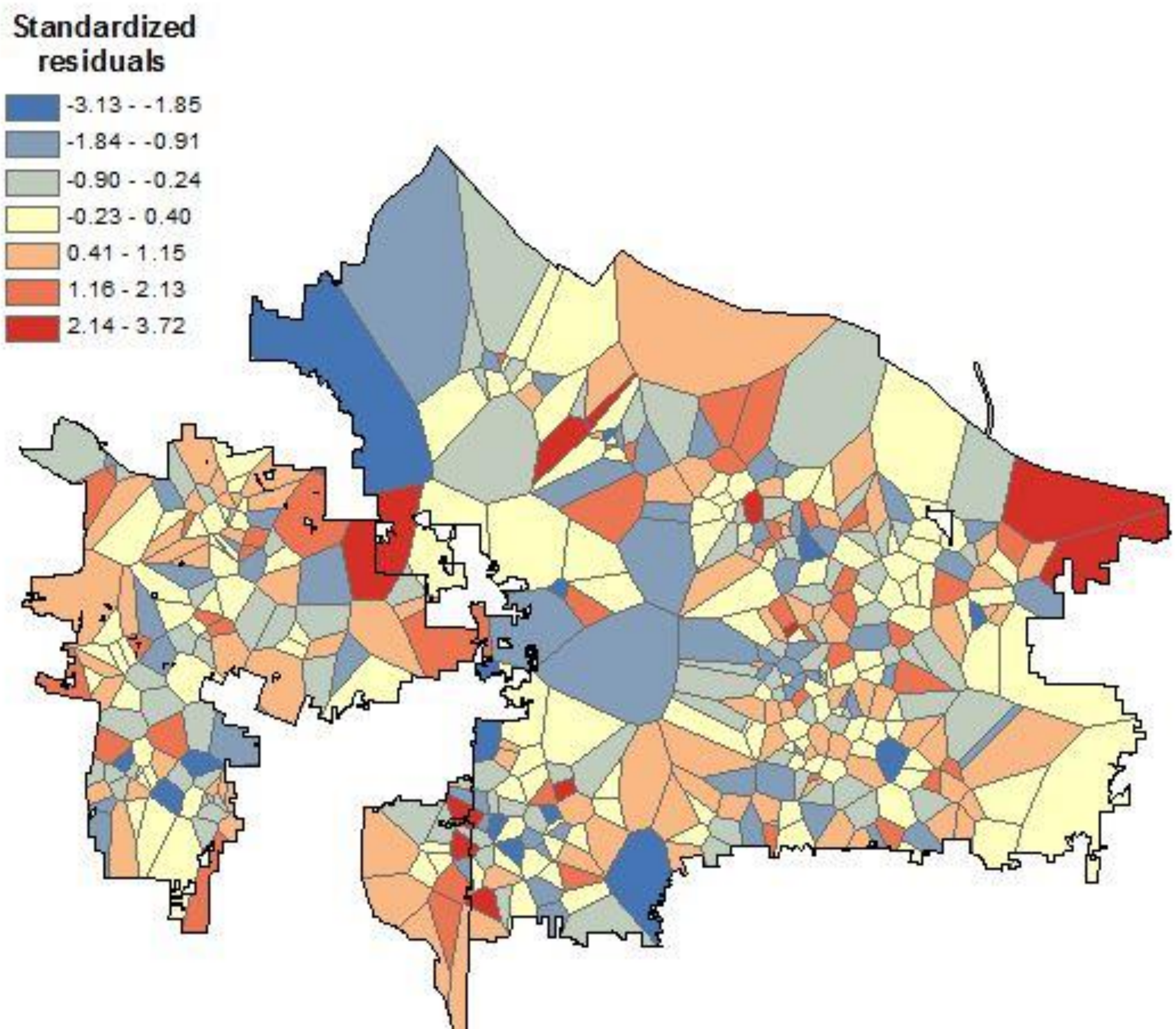
# GWR: Local $R^2$

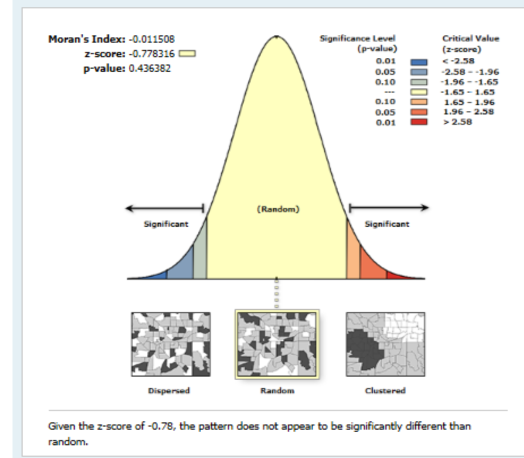


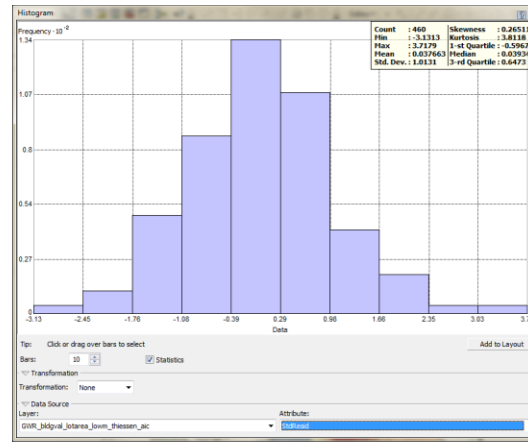
# GWR: Local multicollinearity

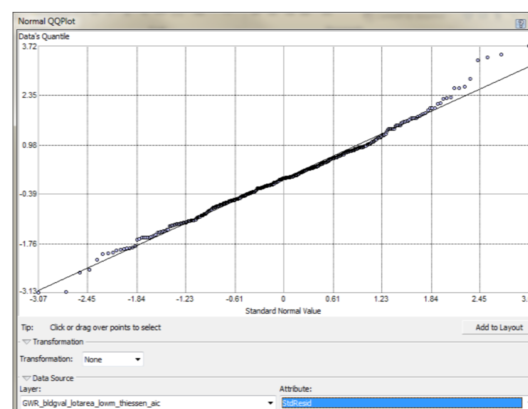


# GWR: Standardized residuals

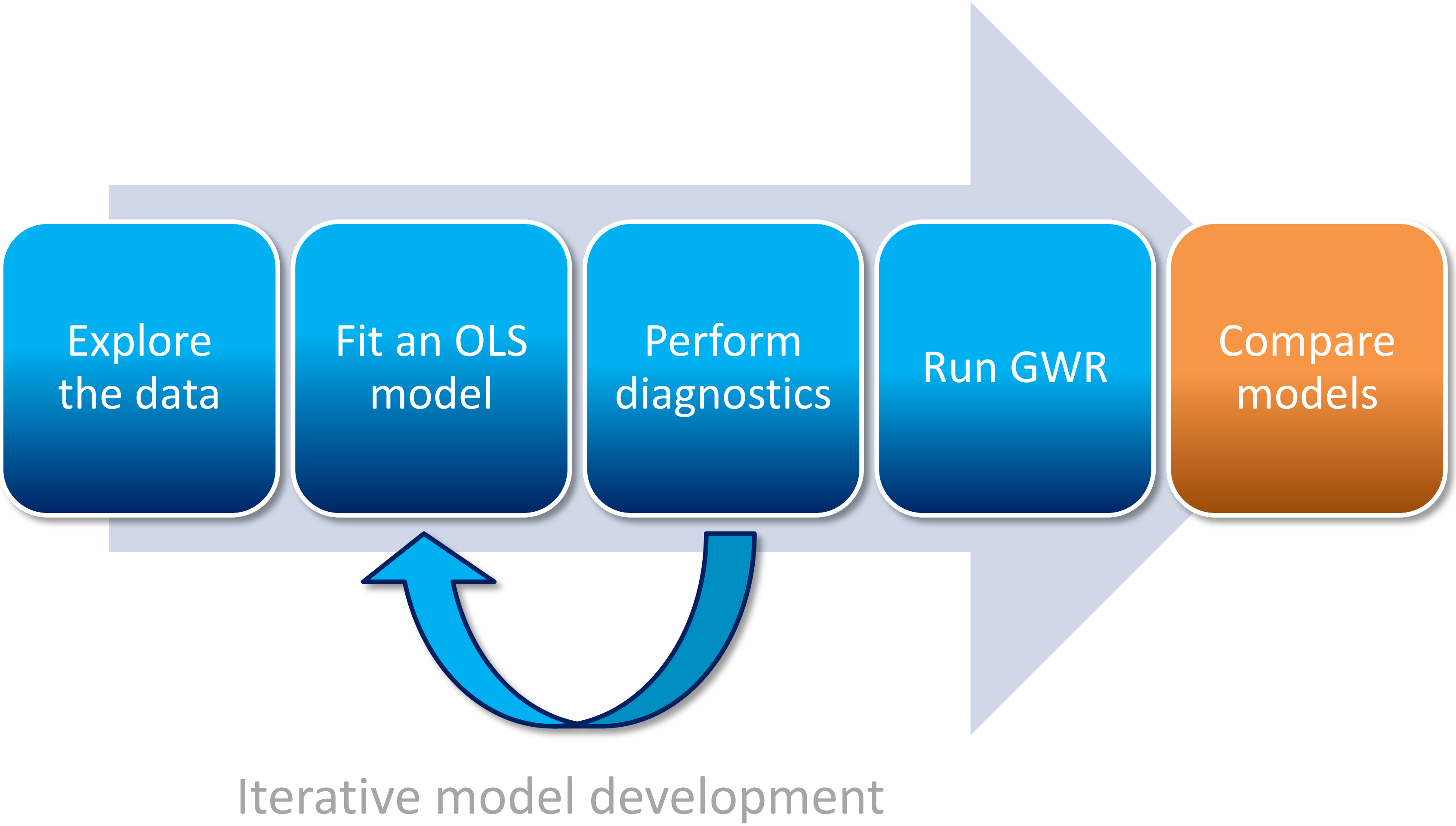


# Model development




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
  - Contraining 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

Brunsdon, C., A.S. Fotheringham, and M.E. Charlton. 1996. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis* 28 (4): 281-298.

Chaix, B., J. Merlo, and P. Chauvin. 2005. Comparison of a spatial approach with a multilevel approach for investigating place effects on health: The example of healthcare utilization in France. *Journal of Epidemiology and Community Health* 59 (6): 517-526.

Fotheringham, A.S. and C. Brunsdon. 1999. Local forms of spatial analysis. *Geographical Analysis* 31 (4): 340-358.

Páez, A., S. Farber, and D. Wheeler. 2011. A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. *Environment and Planning A* 43 (12): 2992 – 3010.

Wheeler D and M. Tiefelsdorf. 2005, Multicollinearity and correlation among local regression coefficients in geographically weighted regression *Journal of Geographical Systems* 7 (2): 161 - 187.

## BOOKS

Fotheringham, A.S., C. Brunsdon, and M.E. Charlton. 2000. *Quantitative geography: Perspectives on spatial data analysis*. London: SAGE Publications.

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

Mitchell, Andy. *The ESRI Guide to GIS Analysis, Volume 2.* Redlands: ESRI Press, 2005.

## RECOMMENDED VIDEO

ESRI Users Conference 2011. Spatial statistics: Best practices.

http://video.esri.com/watch/903/spatial-statistics-best-practices

Charlton, M. 2011. Geographically weighted regression: Modeling spatial heterogeneity. Lecture at Hamilton Institute.

http://www.podcasts.com/hamilton_institute_seminars_hd_large/episode
geographically_weighted_regression_modelling_spatial_heterogeneity

Section 5

# DISCUSSION