Performing Regression Analysis Using Raster Data

Kevin M. Johnston
Xuguang Wang
Outline

• Linear regression
  - Budworm impact

• Spatial autocorrelation

• Sampling

• Using the coefficients

• Spatial regression

• Logistics regression
  - Species distributions and climate change
Problem 1: Linear regression

- Field data (raster surface) of percent canopy damage caused by spruce bud worm (an insect)

- Assumption: where the insect has caused greater damage there are more favorable features located there

- Know what features the insect is responding to but too complex to quantify the relationship

- Want to predict other areas where the spruce budworm may attack
Regression analysis in GIS

• Establishes the relationship of many features and values

• Presents the relationship in a concise manner

• The analysis output is conducive to the GIS

• Make assumptions from samples and apply to the entire population (every location in the raster)
Character of regression

• Dependent variable
  - Biomass
  - Tree growth
  - Probability of deer

• Independent variable
  - Slope
  - Soils
  - Vegetative type

• Linear regression (methods, stepwise, etc)
  - Continuous data

• Logistic regression
  - Presence or absence
Spatial autocorrelation

• What is it?
• The effects on the output from the regression analysis
• Testing for spatial autocorrelation
  - Spatial correlation indices
• Sample points
  - Correlation (take every 5 cell out of 6 row)
  - Random sampling
• In the statistical algorithm
  - Spatial Regression
Using a statistical package

- Synergy between a statistical package and Spatial Analyst
- Why do we need the statistical package?
- Basic assumption— independent observations
Creating the preference surface

- Run regression with the significant factors
- Obtain the coefficients for each value within each raster
- Use the coefficients in a Map Algebra expression to create a preference surface
- The coefficients identify positive or negative influence and magnitude

\[ Z = a_0 + x_1a_1 + x_2a_2 + x_3a_3 \ldots x_na_n \]
Creating the preference surface

- Output from a regression

<table>
<thead>
<tr>
<th>Coef#</th>
<th>Coef</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.250</td>
</tr>
<tr>
<td>1</td>
<td>-0.029</td>
</tr>
<tr>
<td>2</td>
<td>0.263</td>
</tr>
</tbody>
</table>

- Creating the prediction surface with Map Algebra

\[ \text{Outgrid} = 1.25 + (-0.029 \times \text{elevation}) + (0.263 \times \text{distancetoroads}) \]
Spatial Regression

- Still must determine significant variables
- Spatial regression uses spatial autocorrelation
- Use the results to create a probability surface
- Where the regression capability exist:
  - Classical statistical packages
    - SAS, SPSS, R
  - ArcGIS Spatial Statistics toolbox
    - Ordinary Least Squares
    - Geographically Weighted Regression
Demo 1: Regression analysis

Linear Spatial autocorrelation
Problem 2: Logistics regression – True absence

- Want to examine the potential affects of climate change on the distribution of animal species

- Know the current distributions of the species (the dependent variable)

- We have a series of independent variables:
  - Vegetation type (as dummy variables)
  - Elevation, slope, and aspect
  - Distance from roads and cities
  - Etc.
Logistics regression

- Presence/absence model
- Create a probability surface

\[ Z = \frac{1}{1 + \exp(-\sum \alpha_i x_i)} \]
Logistics regression

- Steps
  - Determine sample points
  - Run the logistic regression in R (4 scenarios)
    - Alter input independent variables (All – average temp) (min/max)
    - Greater than 90% and 95% confidence
  - Identify top two models based on statistics
  - Create probability and distribution surfaces
  - Spatial evaluation
  - Identify best model
  - Run model on
    - 2029, 2059, and 2099
Sample Points
Logistics regression: Dependent variables

<table>
<thead>
<tr>
<th>American marten</th>
<th>Pygmy rabbit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big horn sheep</td>
<td>Red tree vole</td>
</tr>
<tr>
<td>Desert rat</td>
<td>Rocky mountain elk</td>
</tr>
<tr>
<td>Dusky wood rat</td>
<td>Sage grouse</td>
</tr>
<tr>
<td>Fisher</td>
<td>Silver bat</td>
</tr>
<tr>
<td>Flying squirrel</td>
<td>Spotted owl</td>
</tr>
<tr>
<td>Gray wolf</td>
<td>Townsend bat</td>
</tr>
<tr>
<td>Prong horn sheep</td>
<td>White vole</td>
</tr>
</tbody>
</table>
Logistics regression: Independent variables

- Aspect
- Elevation
- Slope
- Euclidean distance from population
- Density of streams
- Density of roads
- Average yearly temperature
- Average maximum temperature
- Average minimum temperature
- Average precipitation
Logistics regression: Independent variables (cont.)

- Vegetation – Subtropical mixed forest
- Vegetation – Temperate grassland
- Vegetation – Temperate shrub land
- Vegetation – Temperate evergreen needle-leaf woodland
- Vegetation – Temperate warm mixed forest
- Vegetation – Temperate evergreen needle-leaf forest
- Vegetation – Maritime evergreen needle-leaf forest
- Vegetation – Subalpine forest
- Vegetation - Tundra
The climate data

From Ron Nielson’s group at Oregon State University/ US Forest Service

- We have two climate change models
  - Hadley (from the UK)
  - MIROC 3.2 (from Japan)

- Each model has two scenarios
  - The moderate, mid-level “A1B” carbon scenario
  - The higher, more extreme “A2” carbon scenario

- There are three time periods
  - “e”: Early-century, or 2020-2024 averaged
  - “m”: Mid-century, or 2050-2054 averaged
  - “l”: Late-century, or 2095-2099 averaged
Spatial Regression

- Still must determine significant variables
- Spatial regression uses spatial autocorrelation
- Use the results to create a probability surface

Where the regression capability exist:
- Classical statistical packages
  - SAS, SPSS, R
- ArcGIS Spatial Statistics toolbox
  - Ordinary Least Squares
  - Geographically Weighted Regression
Logistics regression

• Steps
  - Determine sample points
  - Run the logistic regression in R (4 scenarios)
    - Alter input independent variables (All – average temp) (min/max)
    - Greater than 90% and 95% confidence
  - Identify top two models based on statistics
  - Create probability and distribution surfaces
  - Spatial evaluation
  - Identify best model
  - Run model on
    - 2029, 2059, and 2099
### Logistics Diagnostics Table

<table>
<thead>
<tr>
<th>Diag_Name</th>
<th>Diag_Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>75290</td>
</tr>
<tr>
<td>Max Deriv</td>
<td>0.000048</td>
</tr>
<tr>
<td>Model L.R.</td>
<td>30899.737182</td>
</tr>
<tr>
<td>d.f</td>
<td>18</td>
</tr>
<tr>
<td>P</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0.83951</td>
</tr>
<tr>
<td>Dxy</td>
<td>0.679021</td>
</tr>
<tr>
<td>Gamma</td>
<td>0.679896</td>
</tr>
<tr>
<td>Tau-a</td>
<td>0.337119</td>
</tr>
<tr>
<td>R2</td>
<td>0.449892</td>
</tr>
<tr>
<td>Brier</td>
<td>0.162851</td>
</tr>
<tr>
<td>Penalty</td>
<td>0</td>
</tr>
<tr>
<td>AIC</td>
<td>30863.737182</td>
</tr>
</tbody>
</table>

(0 out of 13 Selected)
Logistic Statistics Table

<table>
<thead>
<tr>
<th>OID</th>
<th>Variable</th>
<th>Coef</th>
<th>StdError</th>
<th>Wald</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Intercept</td>
<td>-9.104343</td>
<td>0.182421</td>
<td>-4.908446</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>aspectprj</td>
<td>0.000242</td>
<td>0.000087</td>
<td>2.790799</td>
<td>0.005258</td>
</tr>
<tr>
<td>2</td>
<td>clipdemsnp</td>
<td>-0.000233</td>
<td>0.00004</td>
<td>-5.852734</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>slopeprj</td>
<td>0.024962</td>
<td>0.000932</td>
<td>26.773582</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>ecupoparea</td>
<td>-0.000018</td>
<td>0.000001</td>
<td>-24.719516</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>streamsden</td>
<td>-0.025707</td>
<td>0.005805</td>
<td>-4.428518</td>
<td>0.0000009</td>
</tr>
<tr>
<td>6</td>
<td>histpt970</td>
<td>0.008199</td>
<td>0.000213</td>
<td>38.495107</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>histv22</td>
<td>15.771799</td>
<td>0.384474</td>
<td>41.021719</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>histv17</td>
<td>8.353126</td>
<td>0.284933</td>
<td>29.361434</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>histv16</td>
<td>9.793813</td>
<td>0.235166</td>
<td>41.646298</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>histv12</td>
<td>10.78428</td>
<td>0.23449</td>
<td>45.990416</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>histv11</td>
<td>15.936118</td>
<td>0.344772</td>
<td>46.222242</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>histv6</td>
<td>11.205744</td>
<td>0.231479</td>
<td>48.409268</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>histv7</td>
<td>11.968694</td>
<td>0.242448</td>
<td>49.36596</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>histv6</td>
<td>12.628216</td>
<td>0.279726</td>
<td>45.144935</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>histv2</td>
<td>10.5174</td>
<td>0.551709</td>
<td>19.063296</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>roadsden</td>
<td>0.01792</td>
<td>0.005321</td>
<td>3.367785</td>
<td>0.000758</td>
</tr>
<tr>
<td>17</td>
<td>histtmppmax</td>
<td>-0.117777</td>
<td>0.006206</td>
<td>-18.978045</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>histtmppmin</td>
<td>-0.5010137</td>
<td>0.01032</td>
<td>-49.43125</td>
<td>0</td>
</tr>
</tbody>
</table>
Logistics regression

• Steps
  - Determine sample points
  - Run the logistic regression in R (4 scenarios)
    - Alter input independent variables (All – average temp) (min/max)
    - Greater than 90% and 95% confidence
  - Identify top two models based on statistics
  - Create probability and distribution surfaces
  - Spatial evaluation
  - Identify best model
  - Run model on
    - 2029, 2059, and 2099
Map Algebra Statement – Probability Surface

\[
\text{outraster} = \frac{1}{1 + \exp(-1 \times (-9.10434320922079 + (0.00024204847633 \times \text{Raster(aspectdata)}) + (-0.00023299740223 \times \text{Raster(demdata)}) + (0.02496203077777 \times \text{Raster(slopedata)}) + (-0.00001843603003 \times \text{Raster(eucpopareasdata)}) + (0.01792043302346 \times \text{Raster(roaddendata)}) + (-0.02570716609559 \times \text{Raster(streamsdendata)}) + (-0.1177663211510 \times \text{Raster(tempmaxdata)}) + (-0.51013740745761 \times \text{Raster(tempmindata)}) + (0.00819861349118 \times \text{Raster(precpptdata)}) + (15.77179850521890 \times \text{Raster(veg22data)}) + (8.35312593006918 \times \text{Raster(veg17data)}) + (9.79381294103217 \times \text{Raster(veg16data)}) + (10.78427995755950 \times \text{Raster(veg12data)}) + (15.93611773165130 \times \text{Raster(veg11data)}) + (11.20574409634300 \times \text{Raster(veg8data)}) + (11.96869401781200 \times \text{Raster(veg7data)}) + (12.62821578657110 \times \text{Raster(veg6data)}) + (10.51739997038930 \times \text{Raster(veg2data)})))
\]
Probability Surface
Actual Distribution
Prediction Surface Current .5
Logistics regression

• Steps
  - Determine sample points
  - Run the logistic regression in R (4 scenarios)
    - Alter input independent variables (All – average temp) (min/max)
    - Greater than 90% and 95% confidence
  - Identify top two models based on statistics
  - Create probability and distribution surfaces
  - **Spatial evaluation**
  - Identify best model
  - Run model on
    - 2029, 2059, and 2099
Spatial evaluation

• Evaluation
  - If species is there and predict it is there (100)
  - If the species is not there and predict it is not there (100)

• Adjustments false positives and false negatives
  - Adjust based on how far the misprediction is
  - Adjust if Maxent also predicted the species is there or not
  - Adjust if historically the species was present
outPow = Int(((Power((Abs(Raster(distance1) - 132537) + 0), 4) - minValue1) / (maxValue1 - minValue1)) * ((25 - 1) + 1))
Logistics regression

• Steps
  - Determine sample points
  - Run the logistic regression in R (4 scenarios)
    - Alter input independent variables (All – average temp) (min/max)
    - Greater than 90% and 95% confidence
  - Identify top two models based on statistics
  - Create probability and distribution surfaces
  - Spatial evaluation
  - Identify best model
  - Run model on
    - 2029, 2059, and 2099
Demo 2: Regression analysis

Logistics regression
Climate change analysis
Summary

- **Linear regression**
  - Magnitude

- **Logistics regression**
  - Presence/absence

- **Spatial regression**

- **Sample, calculate coefficients, and create surface**

- **Statistical capability**
  - Spatial Statistics Toolbox
  - ArcGIS to R; SAS Bridge
Thank you...

- Please fill out the session survey:

  First Offering ID: 340/1732

Online – [www.esri.com/ucsessionsurveys](http://www.esri.com/ucsessionsurveys)

Paper – pick up and put in drop box