Esri International User Conference | San Diego, CA

Technical Workshops | **********

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
  - Deer habitat
  - Species distributions and climate change

The problem : Linear regression

- From field data a raster surface has been created defining the percent canopy damage caused by spruce bud worm (an insect)
- There is an assumption that where the insect has caused greater canopy damage, there are more favorable features located there
- We know what features the insect is responding to but it is too complex to quantify the relationship
- We would like to predict the damage the spruce budworm might cause on other locations (from the features located at the locations)

Regression analysis in GIS

- Establishes the relationship of many features and values
- Presents the relationship of these features in a concise manner
- Allows for further exploration of the data

Regression analysis in GIS

- The analysis output format is conducive to the GIS environment
- Can make assumptions from samples and apply them to the entire population (or 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 of it 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

- Synergistic use of a statistical package with Spatial Analyst
- Why do we need the statistical package?
- Basic assumption–independent observations
- Utilizing the results from the models in the GIS

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 if an independent variable has a positive or negative influence and the magnitude of the influence

Creating the preference surface

- **Linear regression**

$$Z = a_0 + x_1a_1 + x_2a_2 + x_3a_3 \ldots x_na_n$$

Creating the preference surface

- **Output from a regression**

Coef#	Coef
0	1.250
1	-0.029
2	0.263

- **Creating the prediction surface with Map Algebra**

Outgrid = 1.25 + (-0.029* elevation) + (0.263 * 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

Regression analysis: Problem two

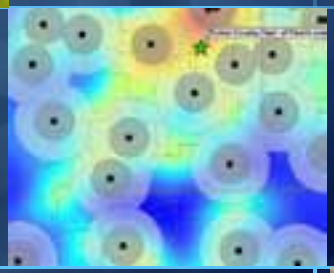
- We know where deer are located
- We have psuedo absence where they are not
- We believe that there are certain attributes that the species prefers at the locations they are at
- We want to predict the preference by the species for each location in the study area

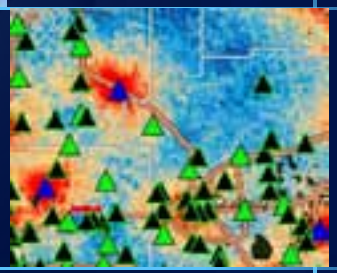
Logistics regression

- **Presence/absence model**
- **Sample**
- **Derive coefficients**
- **Create a probability surface**

$$Z = 1 / 1 + \exp (- S a_i x_i)$$

Demo 1: Regression analysis





Linear

Logistic

Spatial autocorrelation

Problem 3: Logistics regression – True absence

- We want to examine the potential affects of climate change on the distribution of animal species
- We have the known current locations of the distributions of the species
- We have a series of independent variables including
  - Vegetation type (as dummy variables)
  - Elevation, slope, and aspect
  - Distance from roads and cities
  - Etc.

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

The model

- Sample points and associate the raster values for the dependent and independent variables
- Tools created to run R logistics regression
- Fit model
- Coefficients and diagnostics statistics
- Use coefficients to create a raster surface

Creating the raster surface

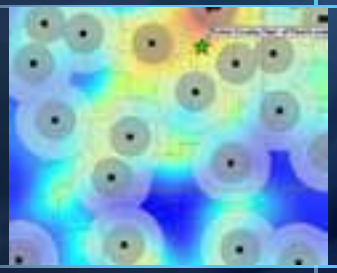
- Apply the logistics formula with coefficients

1 / (1 + exp(-1 * (9.595857 + (-1.28212 * tmp1991) + (-0.003687 * ppt1991) + (0.426121 * veg8_10) + (-0.560821 * veg7_10) + (-2.077026 * veg6_10) + (-2.941375 * veg2_10) + (-0.496024 * veg17_10) + (-1.740473 * veg16_10) + (0.557113 * veg12_10) + (-7.103907 * veg10_10) + (0.016223 * slope) + (-0.000674 * elevation) + (-0.000555 * aspect) + (-0.000062 * disthigh) + (0.000049 * distcity))))

- Select for probability of .5 or greater

- Repeat for each model, for each scenario, and for time period

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

ArcGIS Spatial Analyst Technical Sessions

- **An Introduction - Rm 1 A/B**

Tuesday, July 12, 8:30AM – 9:45AM
Thursday, July 14, 10:15AM – 11:30AM

- **Suitability Modeling - Rm 1 A/B**

Tuesday, July 12, 1:30PM – 2:45PM
Thursday, July 14, 8:30AM – 9:45AM

- **Dynamic Simulation Modeling – Rm 5 A/B**

Wednesday, July 13, 8:30AM – 9:45AM

- **Raster Analysis with Python – Rm 6C**

Tuesday, July 12, 3:15PM – 4:30PM
Wednesday, July 13, 3:15PM – 4:30PM

- **Creating Surfaces – Rm 5 A/B**

Wednesday, July 13, 1:30PM – 2:45PM

ArcGIS Spatial Analyst Short Technical Sessions

- **Creating Watersheds and Stream Networks – Rm 6A**

Tuesday, July 12, 10:40AM – 11:00AM

- **Performing Image Classification – Rm 6B**

Tuesday, July 12, 8:30AM – 8:50AM

- **Performing Regression Analysis Using Raster Data – 6B**

Tuesday, July 12, 8:55AM – 9:15AM

Demo Theater Presentations – Exhibit Hall C

- **Modeling Rooftop Solar Energy Potential**

Tuesday, July 12, 3:30PM – 4:00PM

- **Surface Interpolation in ArcGIS**

Wednesday, July 13, 9:00AM – 10:00AM

- **Getting Started with Map Algebra**

Wednesday, July 13, 10:00AM – 11:00AM

- **Agent-Based Modeling**

Wednesday, July 13, 5:30PM – 6:00PM

Open to Questions

…Thank You!

Please fill the evaluation form.

www.esri.com/sessionevals