Performing Regression Analysis Using Raster Data

Kevin M. Johnston

Xuguang Wang

Technical Workshop

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**

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

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

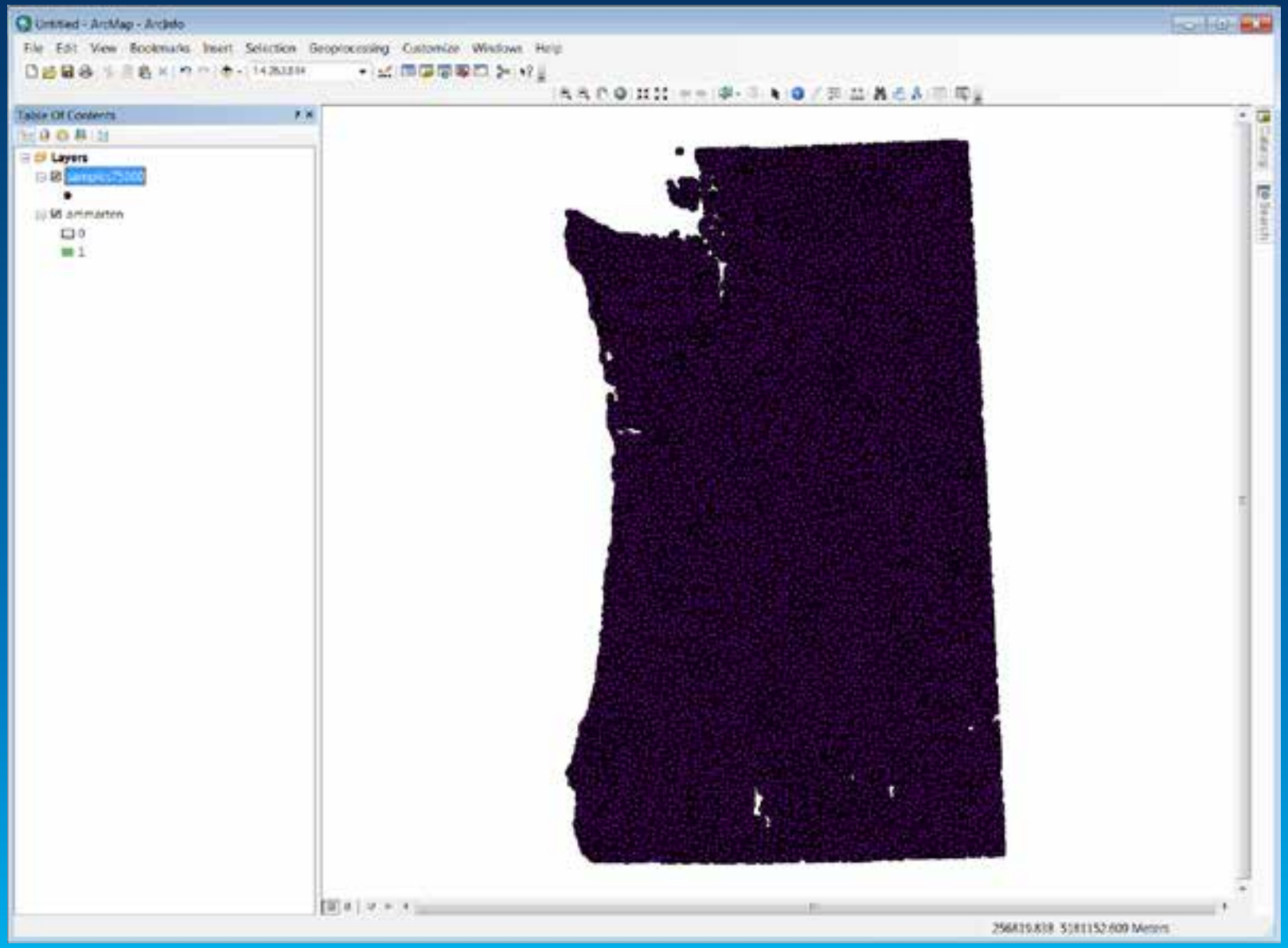
- **Presence/absence model**
- **Create a probability surface**

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

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

- American marten
- Big horn sheep
- Desert rat
- Dusky wood rat
- Fisher
- Flying squirrel
- Gray wolf
- Prong horn sheep
- Pygmy rabbit
- Red tree vole
- Rocky mountain elk
- Sage grouse
- Silver bat
- Spotted owl
- Townsend bat
- White vole

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

ammartendiagmmax.txt

Diag_Name	Diag_Value
Obs	75290
Max Deriv	0.000048
Model L.R.	30899.737182
d.f.	18
P	0
C	0.83951
Dxy	0.679021
Gamma	0.679896
Tau-a	0.337119
R2	0.449892
Brier	0.162851
Penalty	0
AIC	30863.737182

1

(0 out of 13 Selected)

ammartendiagmmax.txt

Logistic Statistics Table

Table

ammartencoefmmax

OID	Variable	Coef	StdError	Wald	Prob
0	Intercept	-9.104343	0.182421	-49.908446	0
1	aspectprj	0.000242	0.000087	2.790799	0.005258
2	clipdemsnp	-0.000233	0.00004	-5.852734	0
3	slopeprj	0.024962	0.000932	26.773582	0
4	ecupoparea	-0.000018	0.000001	-24.719516	0
5	streamsden	-0.025707	0.005805	-4.428518	0.000009
6	histppt970	0.008199	0.000213	38.495107	0
7	histv22	15.771799	0.384474	41.021719	0
8	histv17	8.353126	0.284493	29.361434	0
9	histv16	9.793813	0.235166	41.646298	0
10	histv12	10.78428	0.23449	45.990416	0
11	histv11	15.936118	0.344772	46.222242	0
12	histv8	11.205744	0.231479	48.409268	0
13	histv7	11.968694	0.242448	49.36596	0
14	histv6	12.628216	0.279726	45.144935	0
15	histv2	10.5174	0.551709	19.063296	0
16	roadsden	0.01792	0.005321	3.367785	0.000758
17	histtmpmax	-0.117777	0.006206	-18.978045	0
18	histtmpmin	-0.510137	0.01032	-49.43125	0

1 (0 out of 19 Selected)

ammartendiagmmax.txt ammartencoefmmax

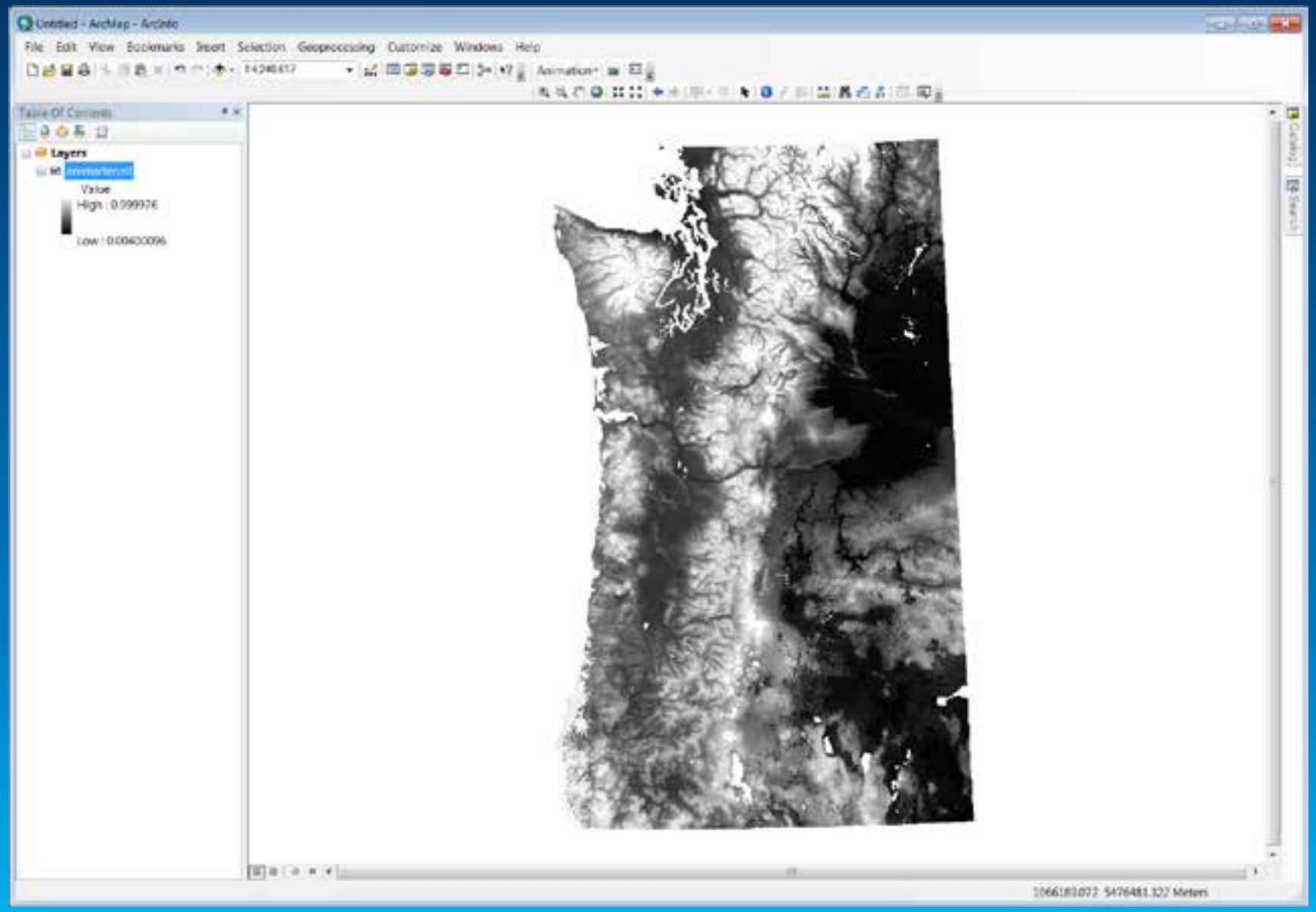
Logistics regression

- 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

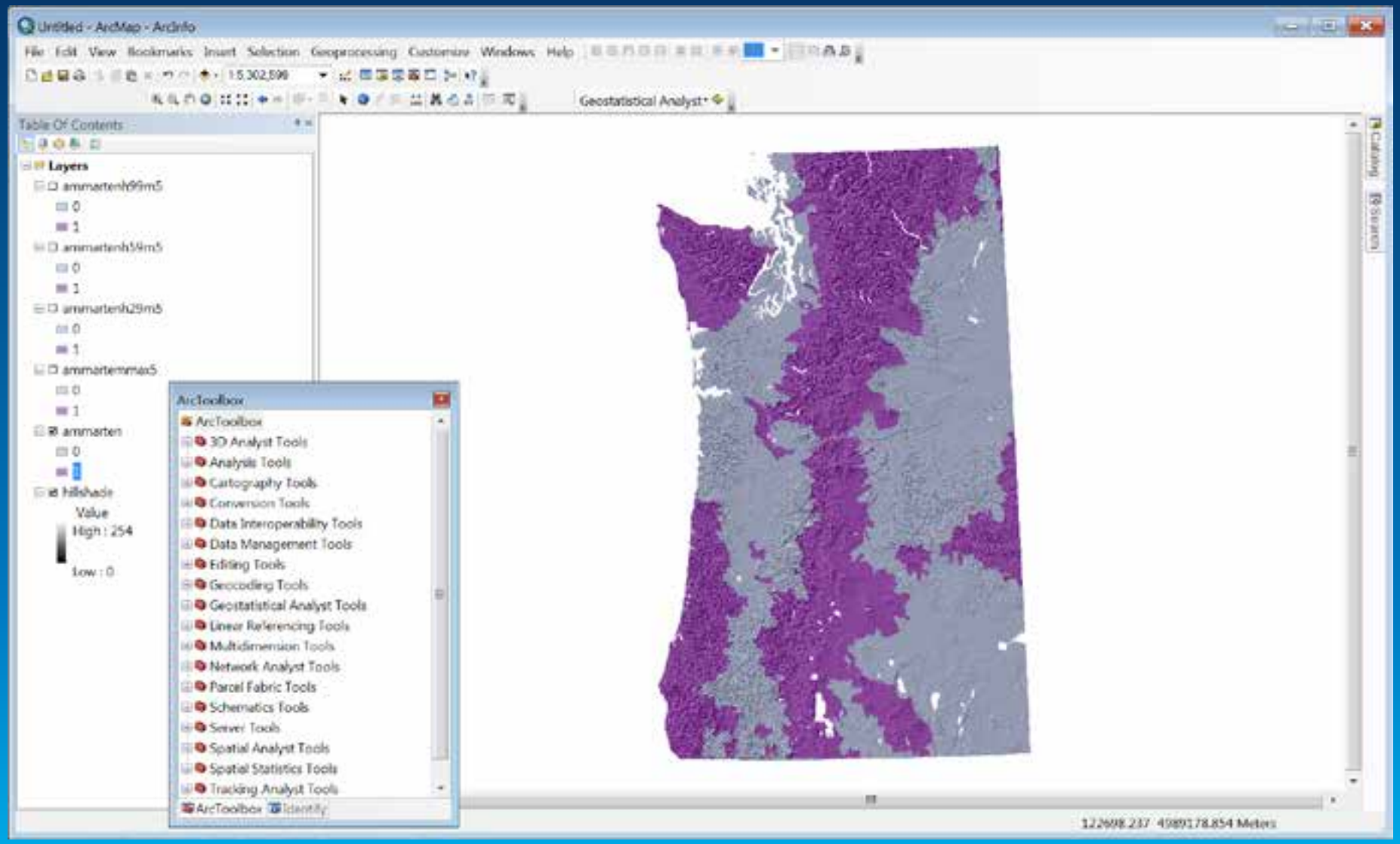
Map Algebra Statement – Probability Surface

```
outraster = 1 / (1 + Exp(-1 * (-9.10434320922079 +
(0.00024204847633 * Raster(aspectdata)) + (-0.00023299740223 *
Raster(demdata)) + (0.02496203077777 * Raster(slopedata)) + (-
0.00001843603003 * Raster(eucpopareasdata)) + (0.01792043302346 *
Raster(roaddendata)) + (-0.02570716609559 *
Raster(streamsdendata)) + (-0.11777663211510 *
Raster(tempmaxdata)) + (-0.51013740745761 * Raster(tempmindata)) +
(0.00819861349118 * Raster(precpptdata)) + (15.77179850521890 *
Raster(veg22data)) + (8.35312593006918 * Raster(veg17data)) +
(9.79381294103217 * Raster(veg16data)) + (10.78427995755950 *
Raster(veg12data)) + (15.93611773165130 * Raster(veg11data)) +
(11.20574409634300 * Raster(veg8data)) + (11.96869401781200 *
Raster(veg7data)) + (12.62821578657110 * Raster(veg6data)) +
(10.51739997038930 * Raster(veg2data)))))
```

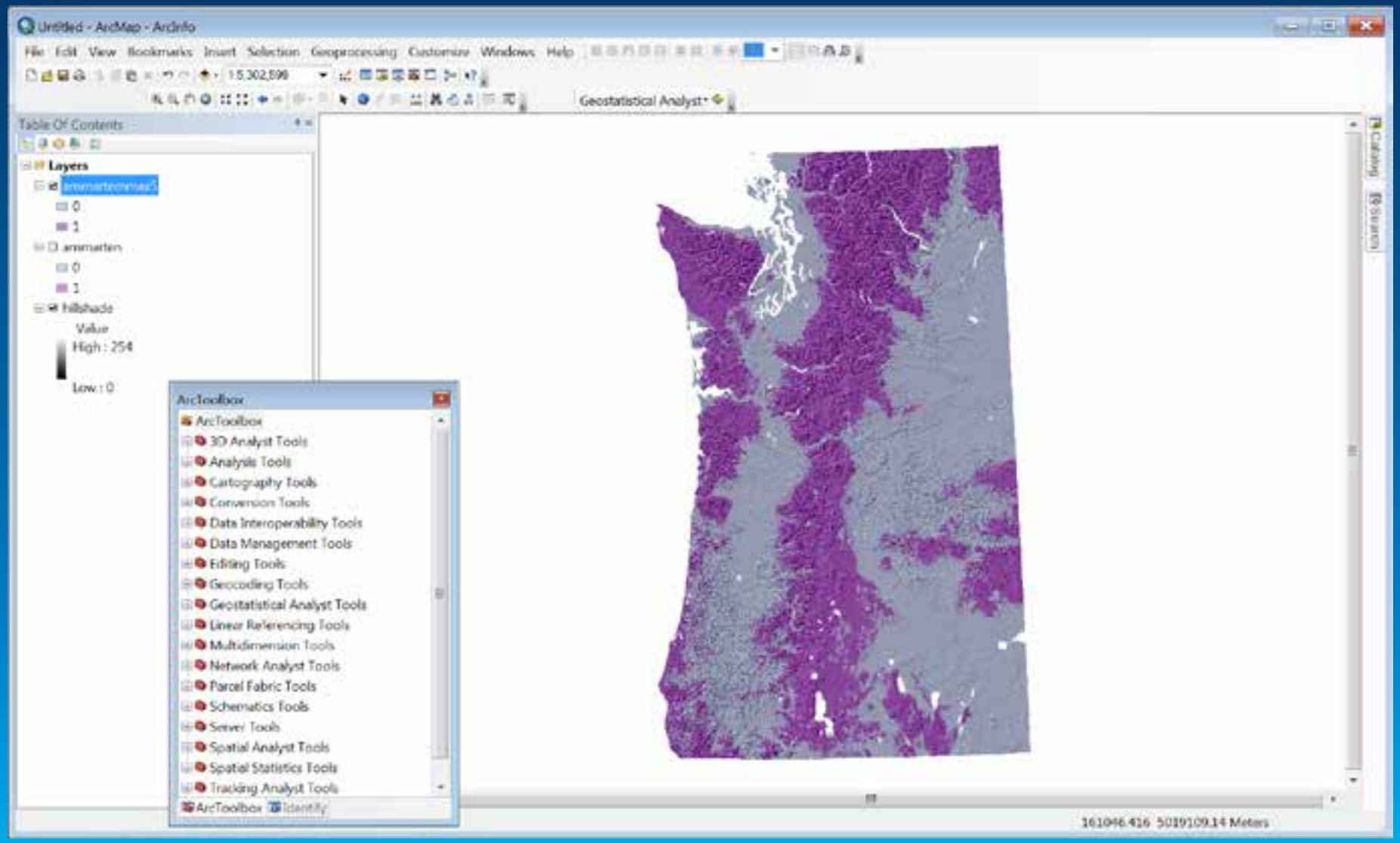
Probability Surface



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

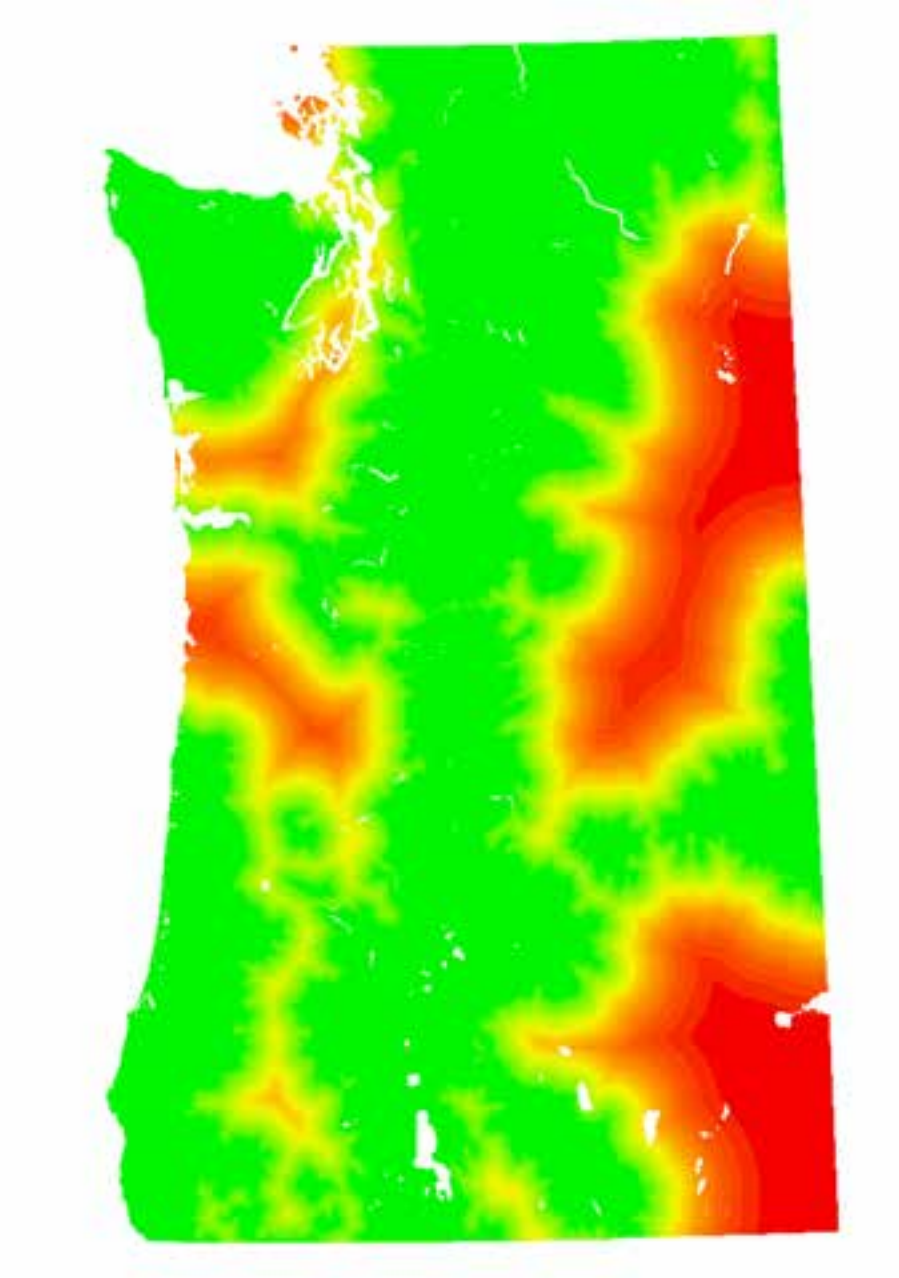
Spatial evaluation

- **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**

Spatial evaluation

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

Paper – pick up and put in drop box

Understanding our world.