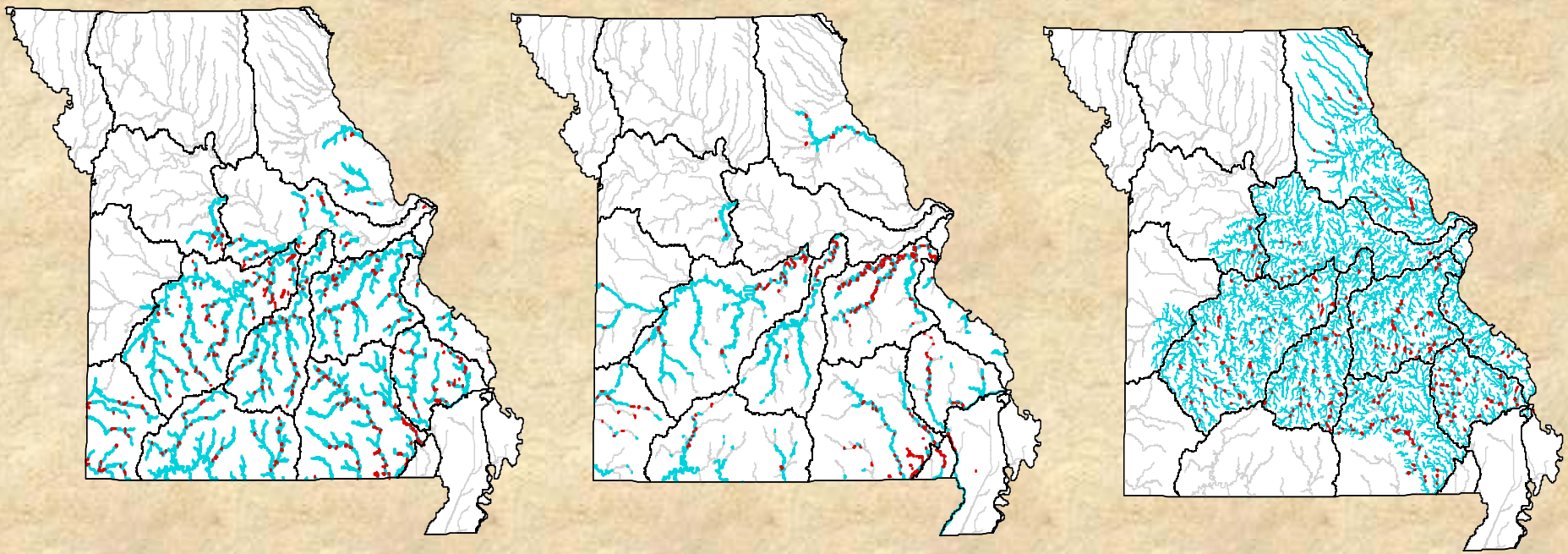
Modeling Distributions of Riverine Biota Using Decision Tree Analyses



USGS National Gap Analysis Meeting

October 6 - 9, 2003

Scott P. Sowa, Gust Annis, Mike Morey, and David D. Diamond

Missouri Resource Assessment Partnership (MoRAP)
University of Missouri

http://www.cerc.cr.usgs.gov/morap

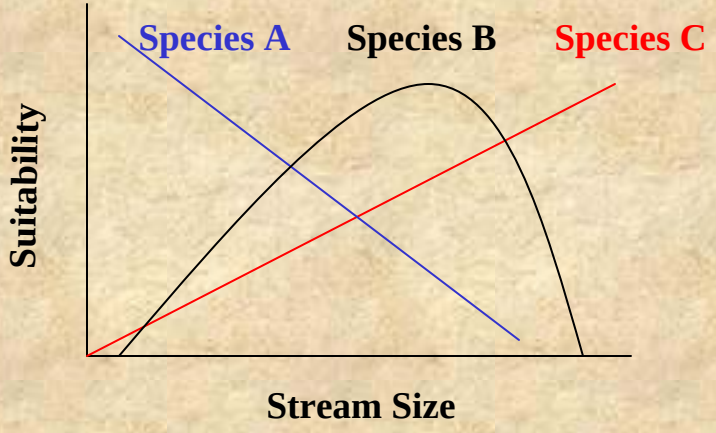
Purpose of Modeling Distributions for GAP

- We cannot directly measure or map biodiversity
  - Species or assemblages serve as surrogate targets for assessing gaps in biodiversity conservation
- We cannot sample everywhere and most sampling data is spatially and temporally biased
  - Provide spatially comprehensive coverage of biological data at the finest resolution of our assessment of conservation gaps (i.e., valley segment scale)
    - A scale at which managers can comprehend and which effective conservation action can take place
- Conservation values of society are largely biologically based
  - The public, legislators, and even scientists can more readily comprehend and relate to biologically-based assessments than other measures of biodiversity (e.g., habitat or processes)

What We Need To Know

- Species has ability to populate suitable habitat
  - Geographic range of the species

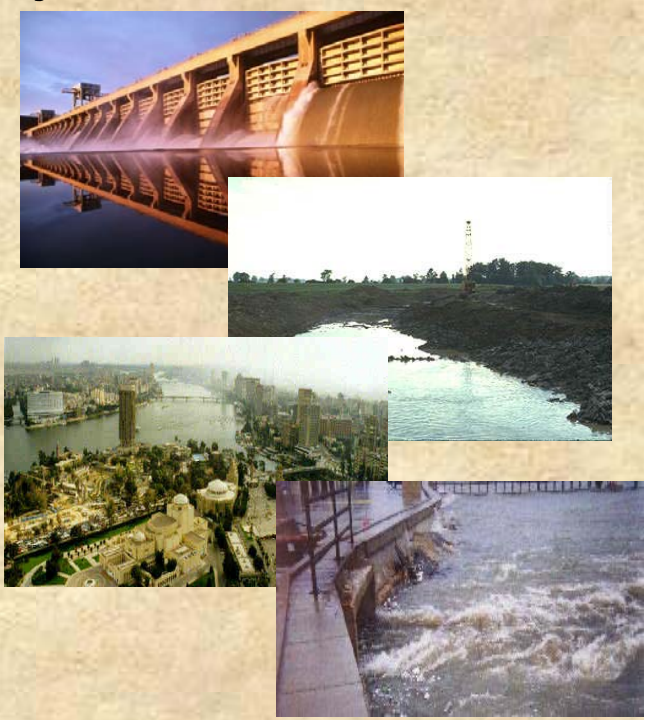
- What constitutes suitable habitat
  - Models of habitat affinities associations



- Where suitable habitat exists within each species geographic range

Important Realizations

- Predicting the present-day biological community of a stream segment is often impossible



  - Impacts to aquatic environments are often diverse, diffuse, distant and hidden

  - The specific response of a given species to a specific human-induced alteration is often unknown

- Habitat affinities often differ with life history stage and season

What We are Trying to Predict

- The Biological Potential of each stream segment
  - The fish, crayfish, and mussel assemblage that would likely occur in a given segment under natural conditions
  - Assume: accuracy of predictions increases as the level of past and present human disturbance decreases
  - No distinctions are made among:
    - reproductive, feeding, rearing, or overwintering habitat
      - **Ex: Dunham and Rieman 1999 (Bull trout)**

Mapping Geographic Ranges

Step 1: Compile existing collection records

Mussel Collection Data

Missouri Department of Conservation

Dr. Ronald D. Oesch

Ohio State University

University of Missouri

University of Colorado

University of Michigan Museum of Zoology

Florida State Museum

The Field Museum of Natural History in Chicago

Fish Collection Data

Missouri Department of Conservation

USGS National Water Quality Assessment Program

Ozark National Scenic Riverways

Environmental Protection Agency

Crayfish Collection Data

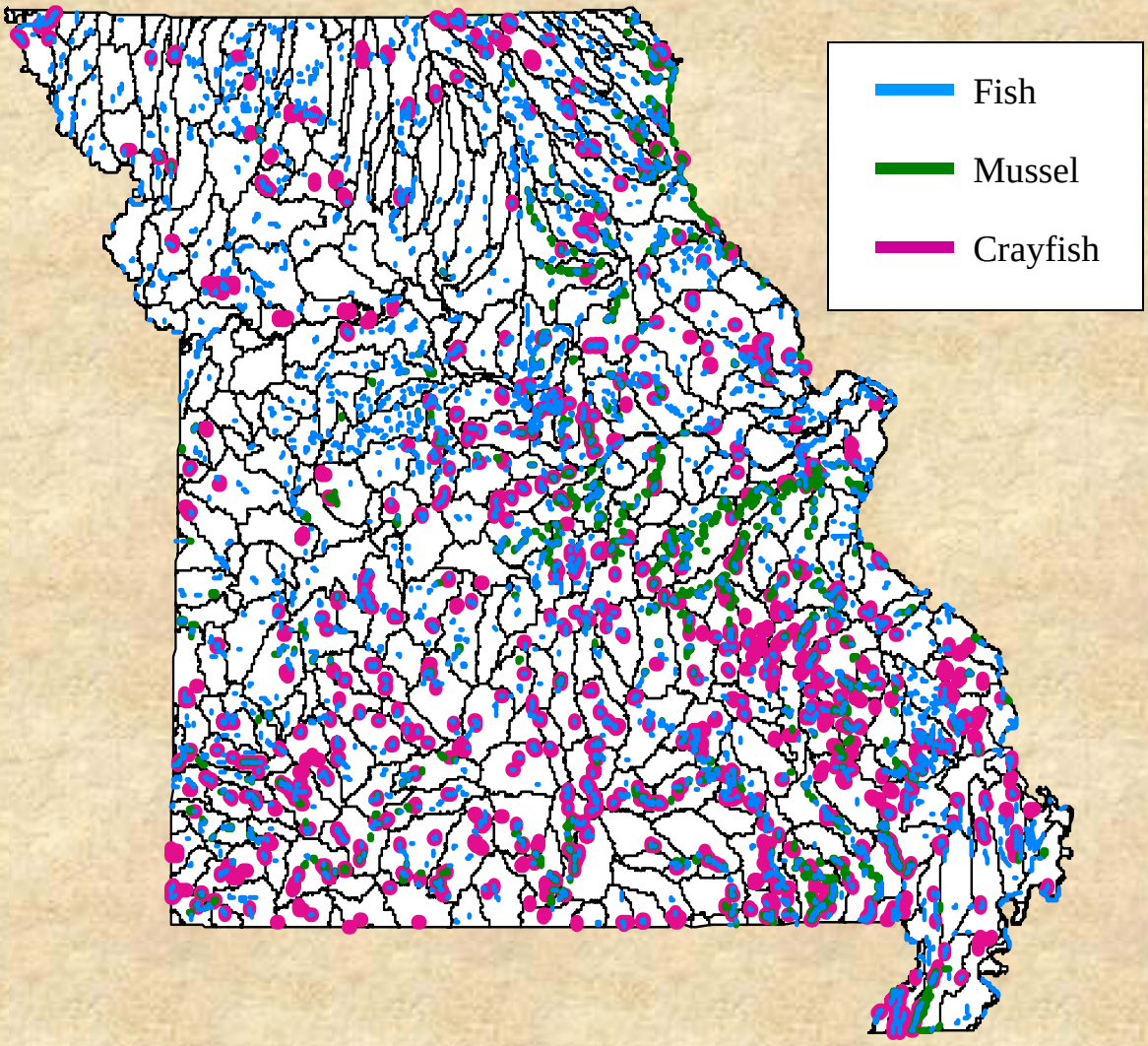
Missouri Department of Conservation

Dr. Ronald D. Oesch

Mapping Geographic Ranges

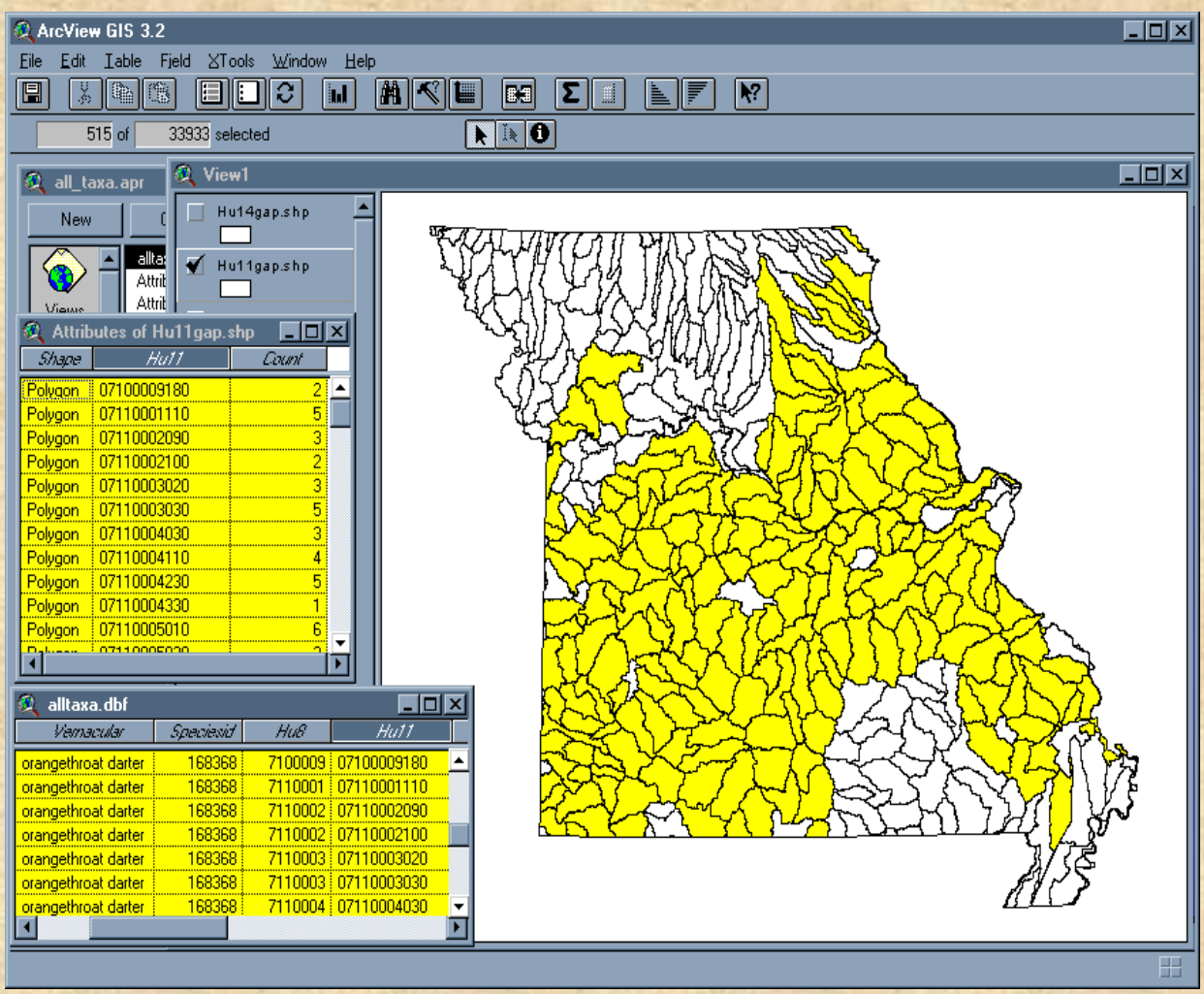
Step 2: Spatially link collection records to NHD and HU coverages

- 5,820 Total records
  - 3,723 Fish
  - 1,157 Mussel
  - 940 Crayfish



Mapping Geographic Ranges

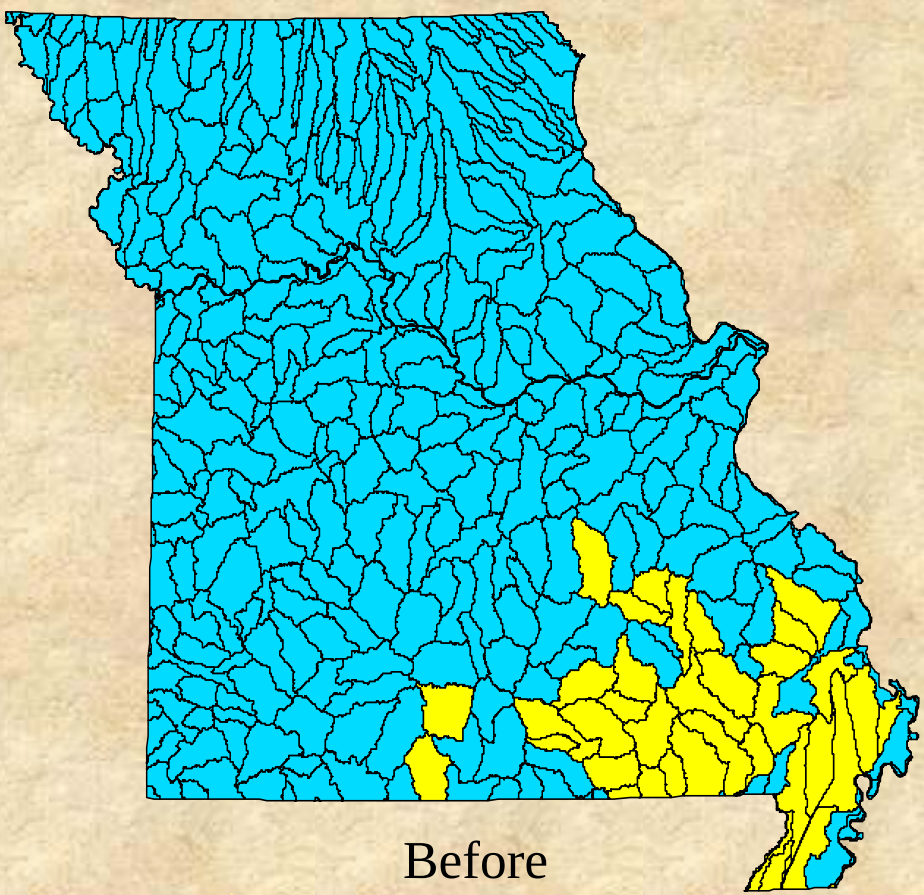
Step 3: Map Geographic Ranges by Hydrologic Unit



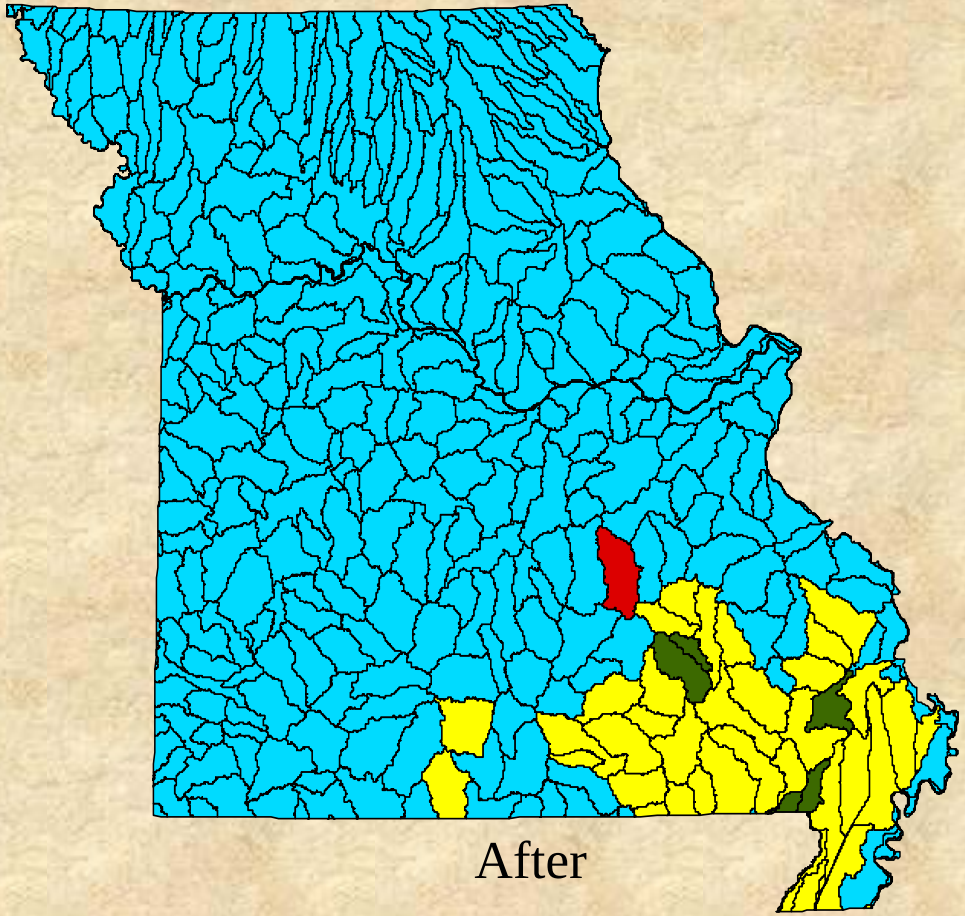
- Accounts for broad-scale environmental constraints on distributions

Mapping Geographic Ranges

Step 4: Professional Review of Range Maps



Before

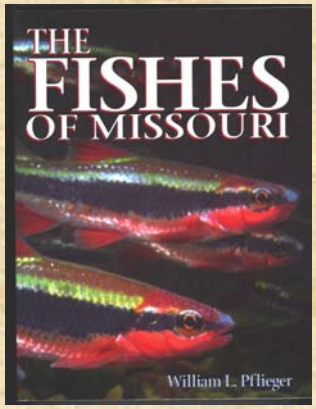


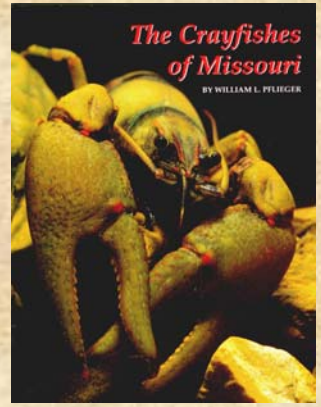
After

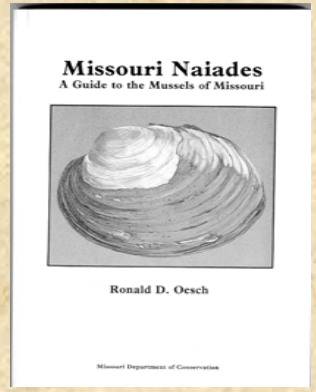
Geographic Range by 10-digit Hydrologic Unit
(Redspotted sunfish)

Developing Models of Habitat Affinities/Associations

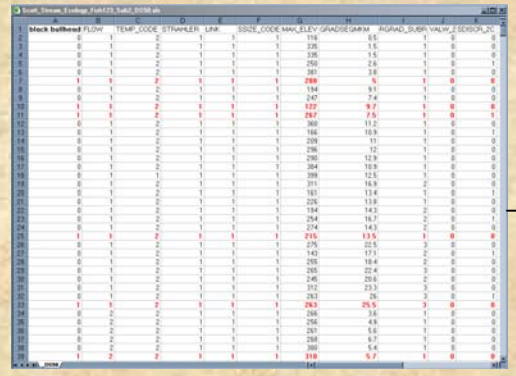
- Use information extracted from the literature

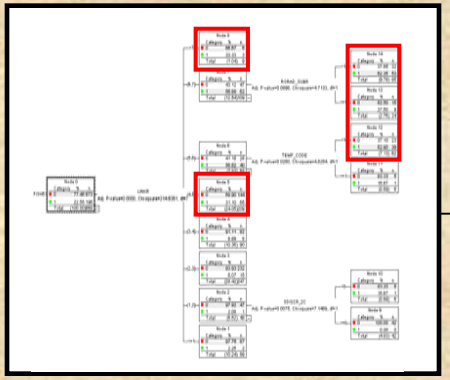


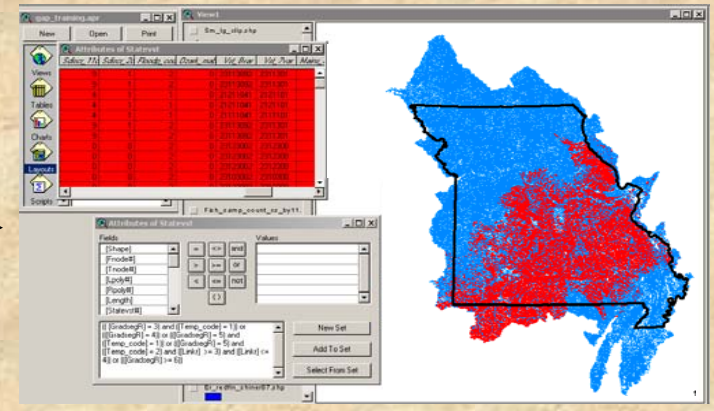




- Empirically generate models using collection records







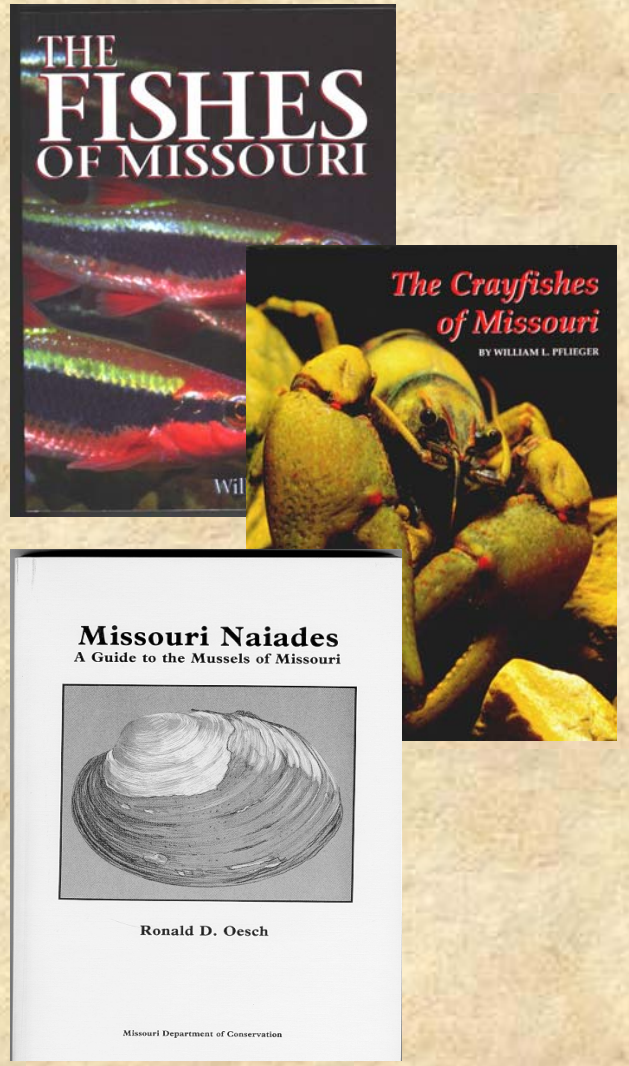
Documenting Habitat Affinities

Creek chub

(Semolitus atromaculatus)

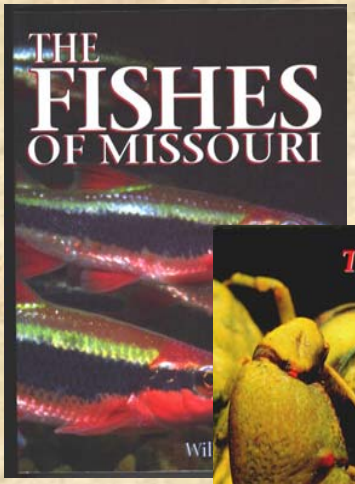
Habitat Affinities

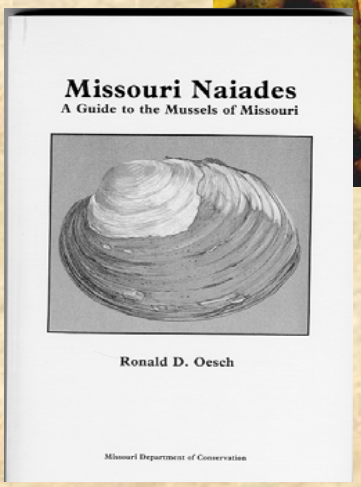
- Inhabits small headwater creeks with gravel riffles.
- Tolerates a wide range of turbidity.
- Most abundant in small headwater creeks
- Associated with moderate to swift currents when spawning
- Gravel bottom streams with minimal silt.
- In MO found where there was 3-6% silt.
- Associated with shallow water depths.
- Found from 101' to 2000' elevations
- Usually spawn just above riffle where silting is minimal.
- Prefers pH of around 7.0
- Feeds in slow currents
- Found in all stream orders, but primarily first through sixth order
- In MO, 58% of collections were in stream orders 3 and 4.
- In MO, spawns from April to late May.
- Substrate may be a limiting factor in MO's small streams and ditches.



Pros and Cons of Using Habitat-Affinity Information to Generate Predictive Models

- Pros
  - Easy to interpret basis of predictions
  - Can be used with very few samples
  - Helps qualitatively validate empirical models
- Cons
  - Literature searches are difficult
  - Habitat-affinity information is often lacking
  - Habitat-affinity information mostly pertains to microhabitat variables
  - Information often lacks context
  - Subjective selection of predictors and quantitative model constraints

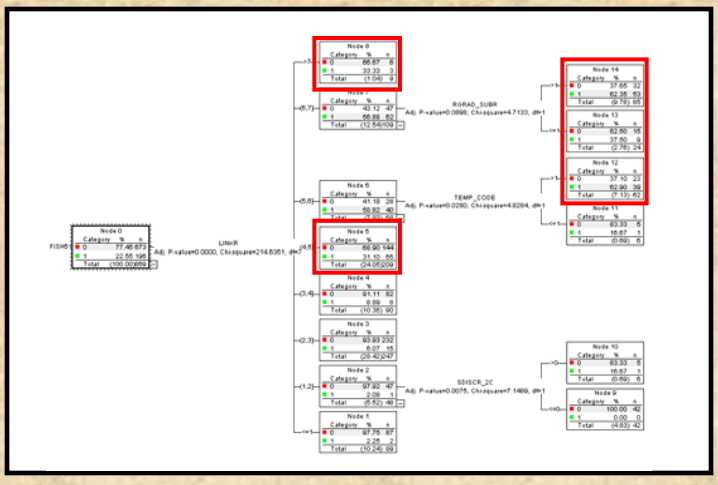




Different Empirical Approaches to Modeling Species Distributions

- Decision Tree Analysis
- Logistic Regression
- Discriminant Analysis
- Generalized Linear Models
- Generalized Additive Models
- K-nearest Neighbor
- Neural Networks
- Bayesian probability, logical inference, genetic algorithms (GARP models)

Pros and Cons of Using Decision Tree Analyses

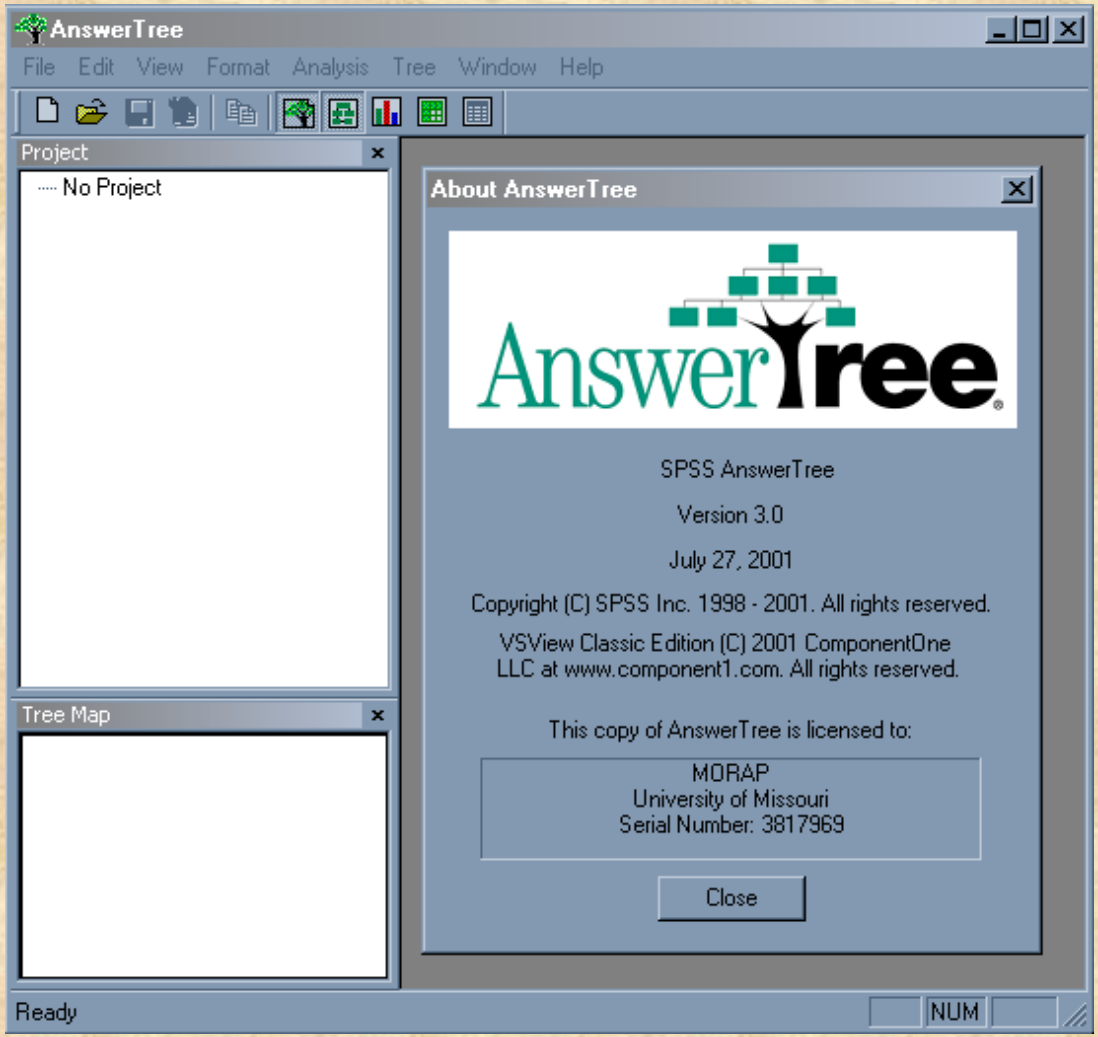


- Pros
  - Easy to perform analyses
  - Easy to interpret models
  - Literature suggests CART models are quite accurate
  - Can handle non linear relationships
  - Can handle various data types (binary, nominal, ordinal)
- Cons
  - Data hungry
  - Not as accurate as more complex methods

Statistical Software Used

Qualities

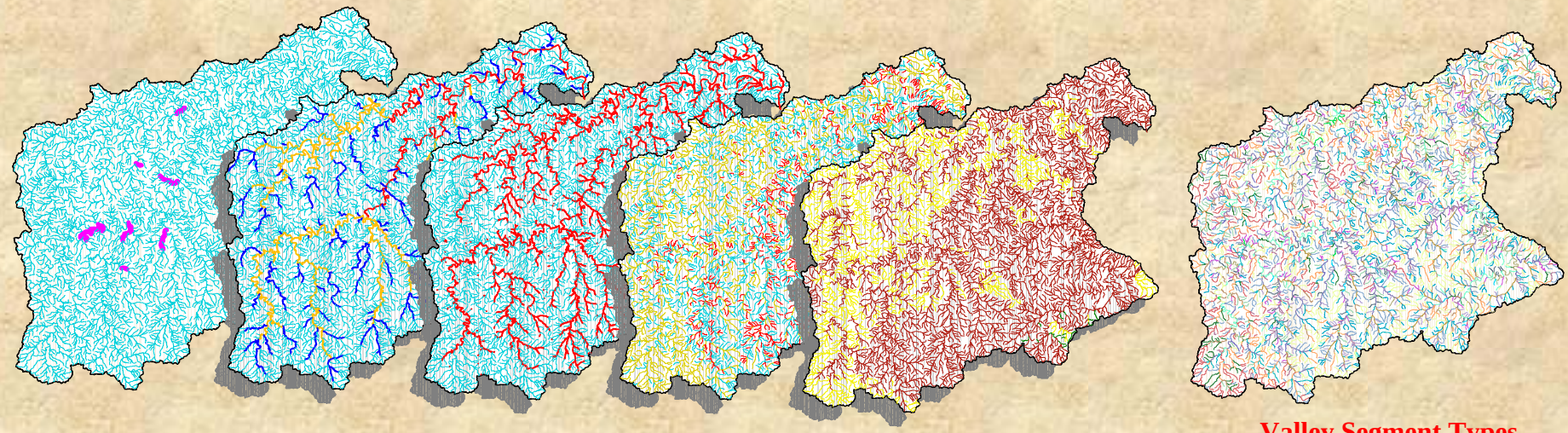
- User friendly
- Compatible with Excel
- Performs 4 different analyses
  - CHAID
  - Exhaustive CHAID
  - CART
  - QUEST



Predictor Variables Generated from Statewide Valley Segment Coverage

Individual Variables

Unique Valley Segment Types



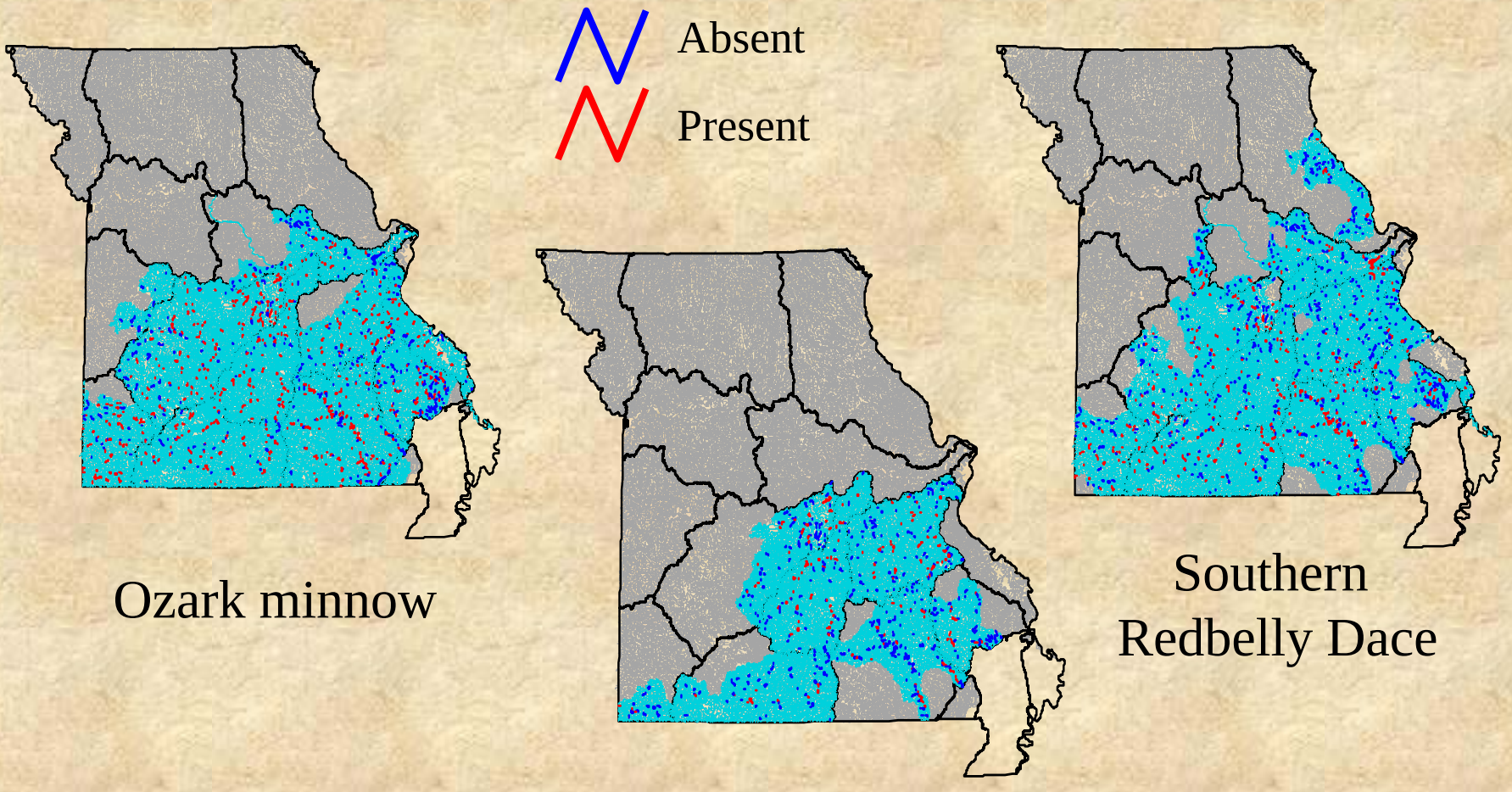
Temperature Stream Size Flow Gradient Geology

Valley Segment Types

Predictor Variables and Classes

- Temperature
  - Binary (cold vs. warm)
- Flow
  - Binary (perennial vs. intermittent)
- Stream Size
  - 5 classes (headwater, creek, sm. river, lg. river, great river)
  - 10 classes
- Gradient
  - 10 classes (actual gradients)
  - 3 classes (relative gradients)
- Size Discrepancy
  - Binary (yes or no)

Selecting Input Data



Ozark minnow

Wedgespot shiner

Southern
Redbelly Dace

Input Data Matrix

Scott_Stream_Ecology_Fish123_Sub2_DS50.xls

	A	B	C	D	E	F	G	H	I	J	K
1	**black bullhead**	FLOW	TEMP_CODE	STRAHLER	LINK	SSIZE_CODE	MAX_ELEV	GRADSEGMKM	RGRAD_SUBR	VALW_2	SDISCR_2C
2	0	1	2	1	1	1	116	0.5	1	0	0
3	0	1	2	1	1	1	335	1.5	1	0	0
4	0	1	2	1	1	1	335	1.5	1	0	0
5	0	1	2	1	1	1	250	2.6	1	0	1
6	0	1	2	1	1	1	381	3.8	1	0	0
7	**1**	**1**	**2**	**1**	**1**	**1**	**288**	**5**	**1**	**0**	**0**
8	0	1	2	1	1	1	194	9.1	1	0	0
9	0	1	2	1	1	1	247	7.4	1	0	0
10	**1**	**1**	**2**	**1**	**1**	**1**	**122**	**9.7**	**1**	**0**	**0**
11	**1**	**1**	**2**	**1**	**1**	**1**	**267**	**7.5**	**1**	**0**	**1**
12	0	1	2	1	1	1	360	11.2	1	0	0
13	0	1	2	1	1	1	166	10.9	1	0	1
14	0	1	2	1	1	1	209	11	1	0	0
15	0	1	2	1	1	1	296	12	1	0	0
16	0	1	2	1	1	1	290	12.9	1	0	0
17	0	1	2	1	1	1	384	10.9	1	0	0
18	0	1	1	1	1	1	399	12.5	1	0	0
19	0	1	2	1	1	1	311	16.9	2	0	0
20	0	1	2	1	1	1	161	13.4	1	0	1
21	0	1	2	1	1	1	226	13.8	1	0	0
22	0	1	2	1	1	1	194	14.3	2	0	0
23	0	1	2	1	1	1	254	16.7	2	0	1
24	0	1	2	1	1	1	274	14.3	2	0	0
25	**1**	**1**	**2**	**1**	**1**	**1**	**215**	**13.5**	**1**	**0**	**0**
26	0	1	2	1	1	1	275	22.5	3	0	0
27	0	1	2	1	1	1	143	17.1	2	0	1
28	0	1	2	1	1	1	255	18.4	2	0	0
29	0	1	2	1	1	1	265	22.4	3	0	0
30	0	1	2	1	1	1	245	20.6	2	0	0
31	0	1	2	1	1	1	312	23.3	3	0	0
32	0	1	2	1	1	1	263	26	3	0	1
33	**1**	**1**	**2**	**1**	**1**	**1**	**263**	**25.5**	**3**	**0**	**0**
34	0	2	2	1	1	1	266	3.6	1	0	0
35	0	2	2	1	1	1	256	4.9	1	0	0
36	0	2	2	1	1	1	261	5.6	1	0	0
37	0	2	2	1	1	1	268	6.7	1	0	0
38	0	2	2	1	1	1	380	5.4	1	0	0
39	**1**	**2**	**2**	**1**	**1**	**1**	**318**	**5.7**	**1**	**0**	**0**

DS50

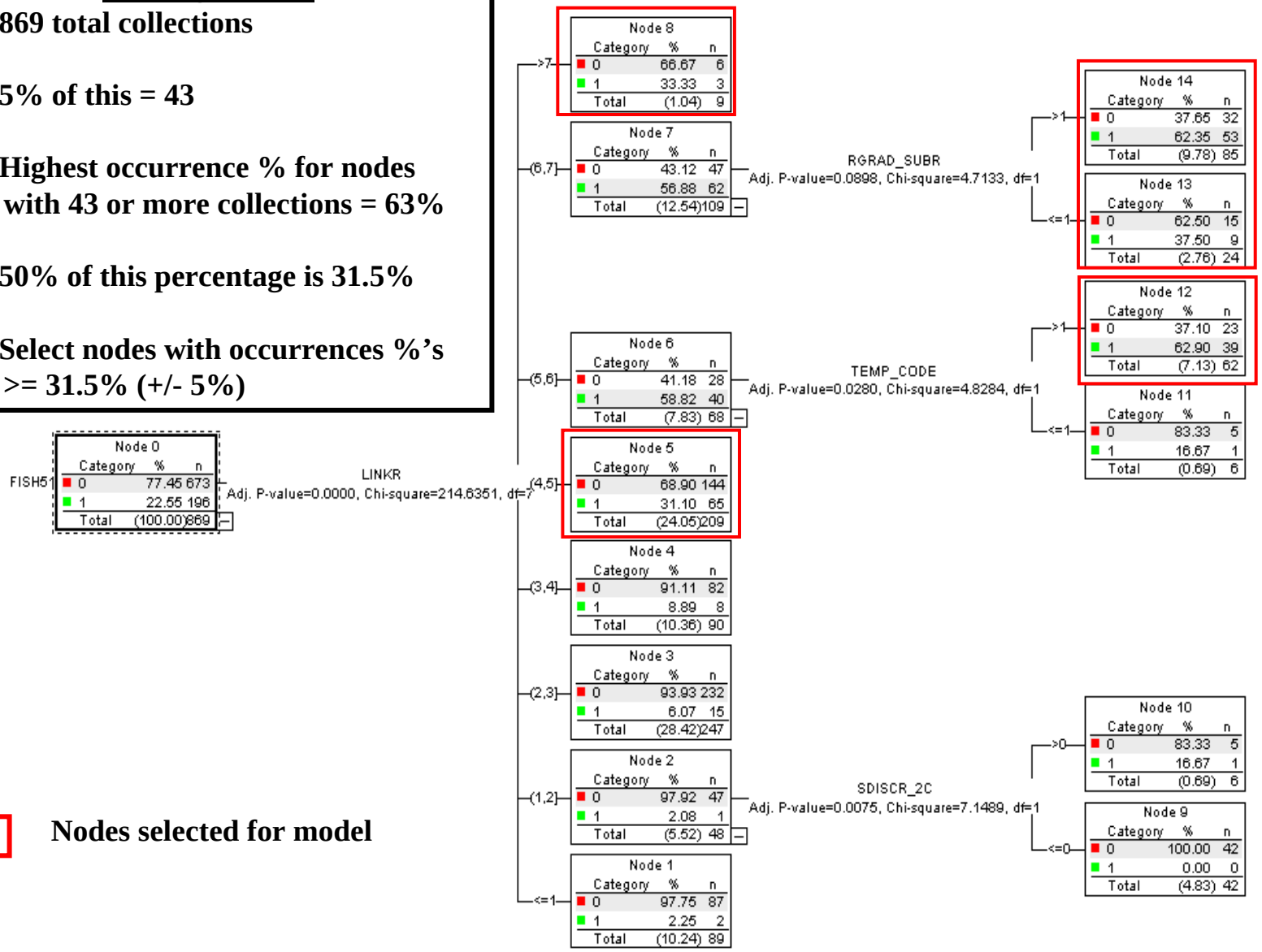
Model Selection and Pruning Criteria

- Selection criteria
  - Method – Exhaustive CHAID
  - Alpha = 0.1 using Bonferoni correction
    - Increased to 0.2 if no model developed
  - Maximum number of levels = 10
  - Minimum number of records = 1
- Pruning criteria
  - "Relative 50% Rule"
  - Find highest occurrence % among nodes containing at least 5% of the total collection records
    - e.g., 500 collections in input dataset, find highest occurrence% for nodes containing 25 or more collections
  - Divide highest occurrence% by 2 and select all nodes having occurrence% $\geq$ to this value
    - e.g., highest = 40%, select all nodes with occurrence% $\geq$ 20% (+/- 5%)

Classification Tree Diagram for the Wedgespot Shiner

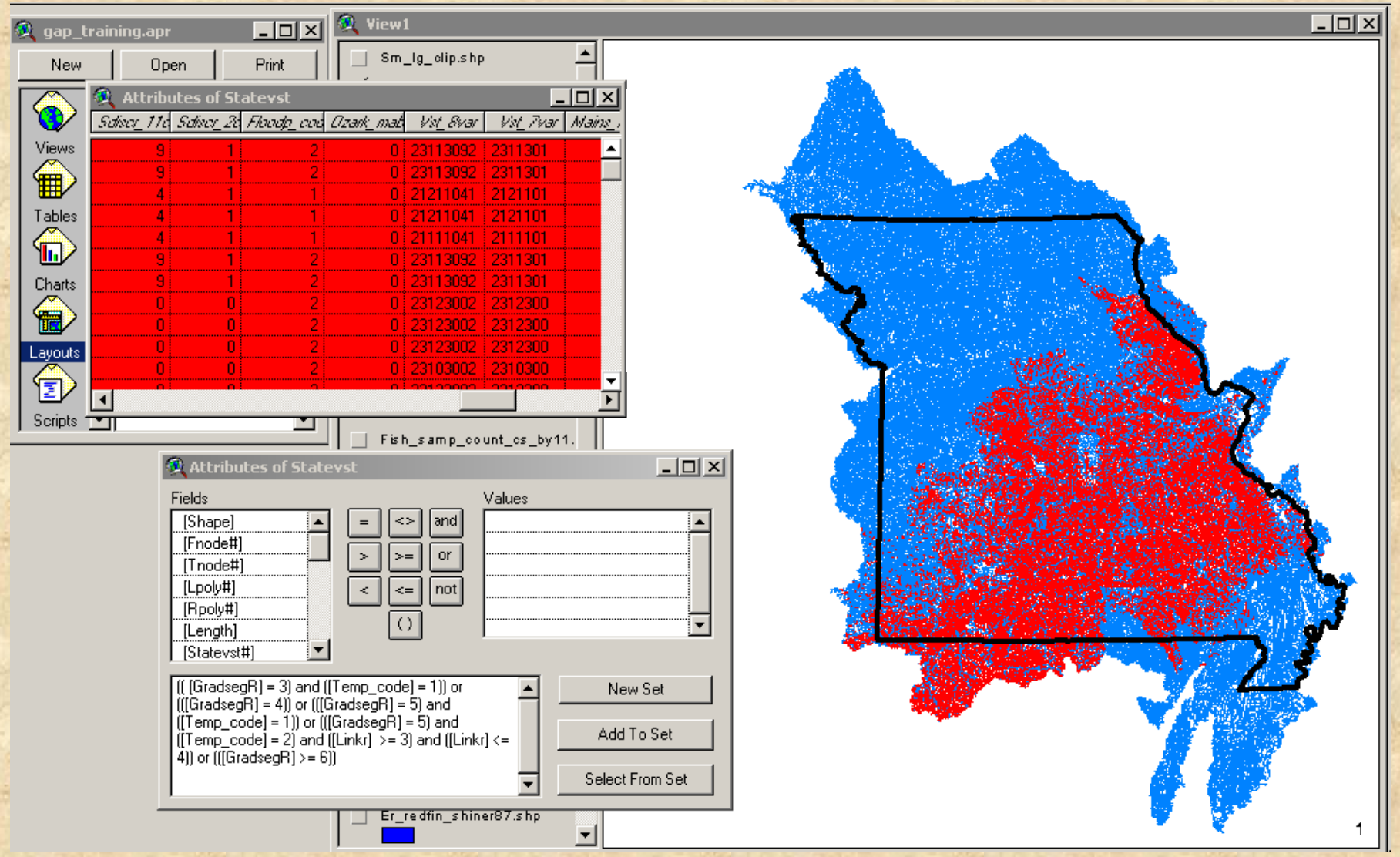
Pruning the tree

- **869 total collections**
- **5% of this = 43**
- **Highest occurrence % for nodes with 43 or more collections = 63%**
- **50% of this percentage is 31.5%**
- **Select nodes with occurrences %'s >= 31.5% (+/- 5%)**

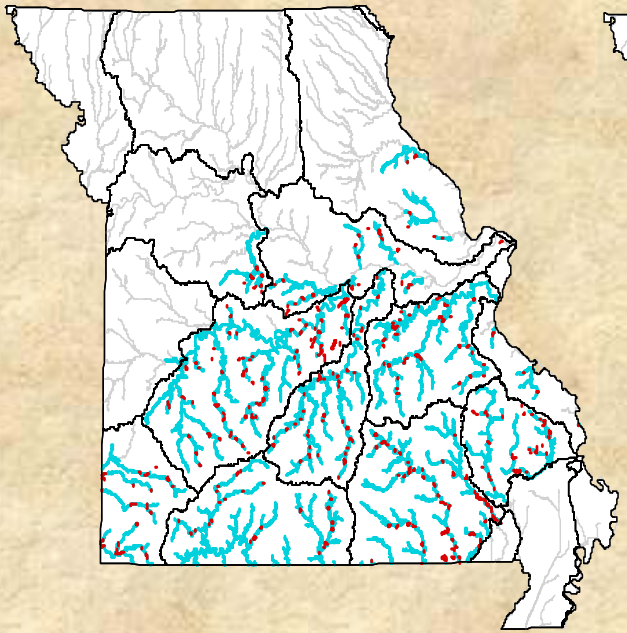


Nodes selected for model

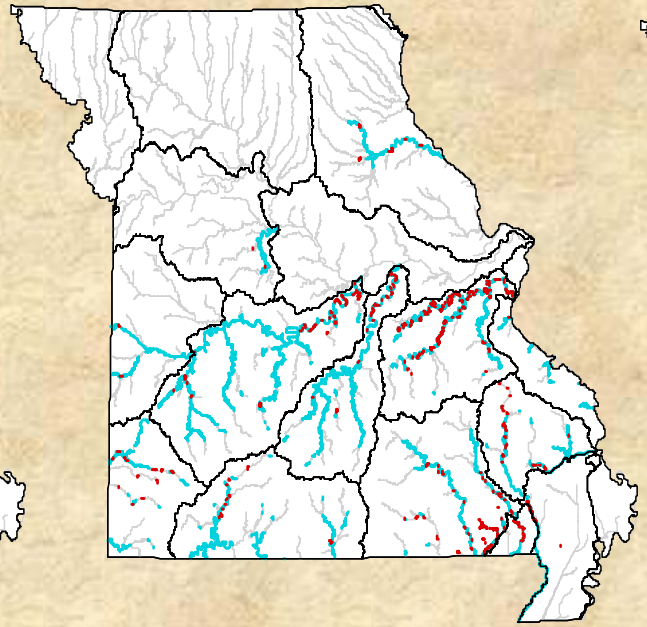
Use Model to Select Suitable Stream Segments in ArcView



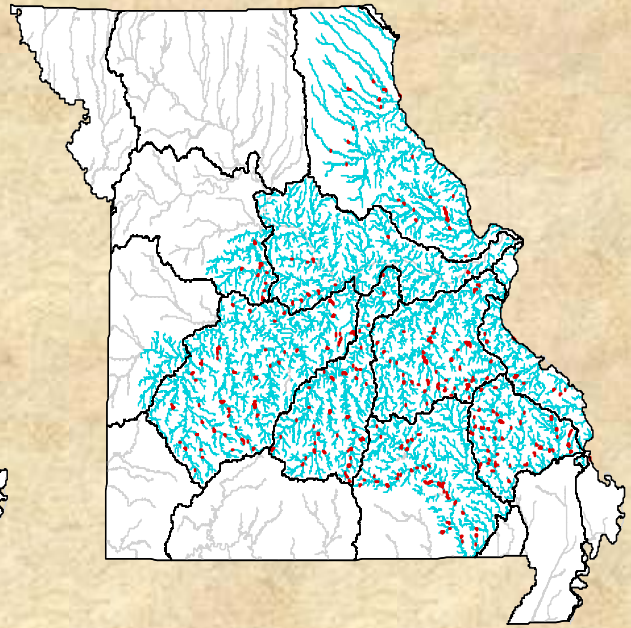
Constructed Models Separately for Each Species



Black redhorse



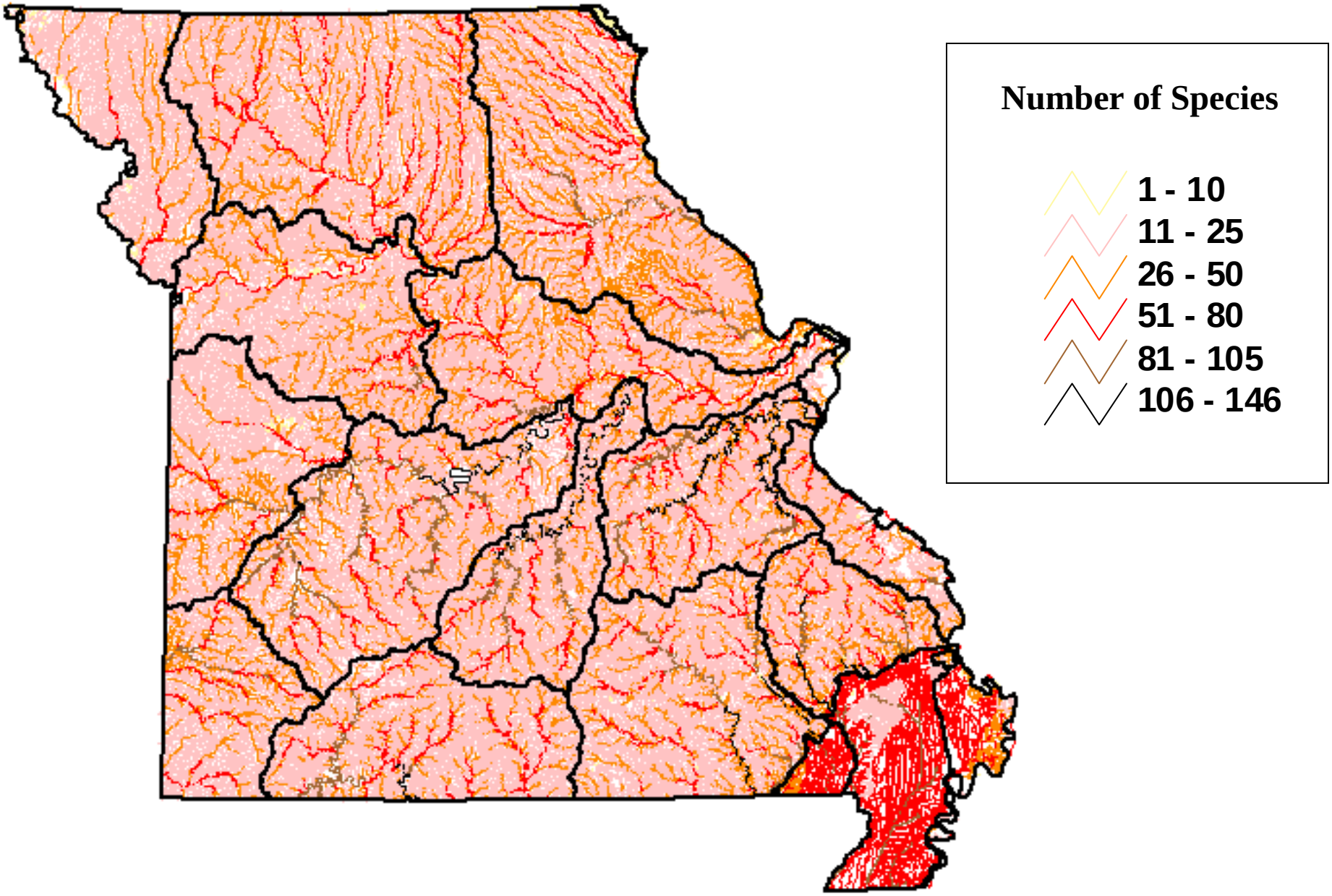
Round pigtoe



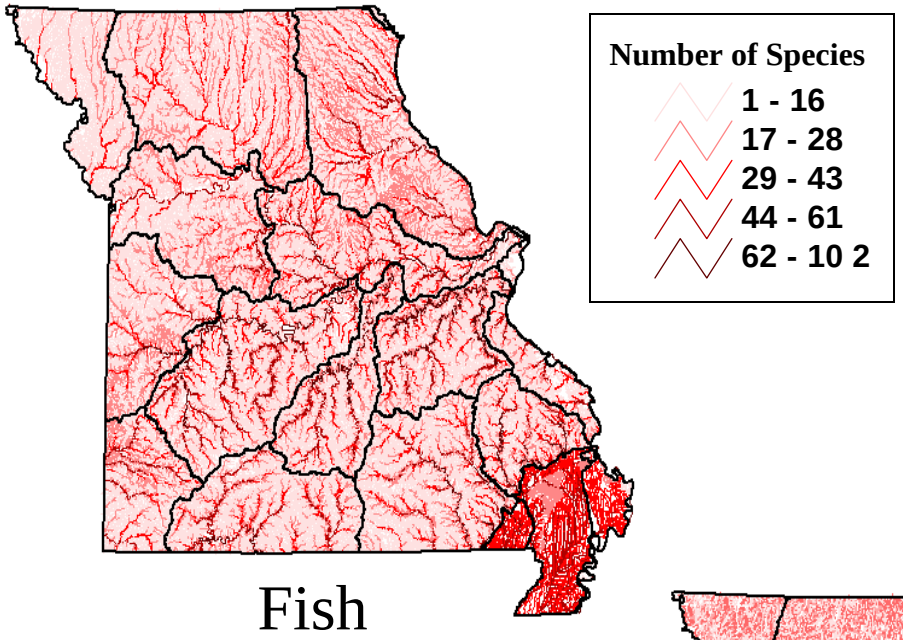
Golden crayfish

- 571 total models constructed for 315 different species
  - Anywhere from 1 to 4 models developed for each species

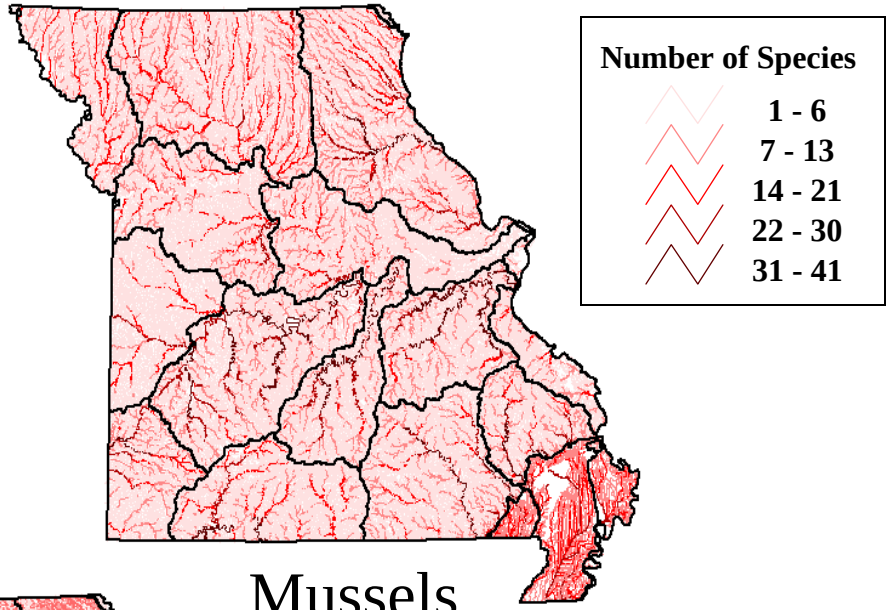
Predicted Species Richness (all fish, mussel, crayfish)



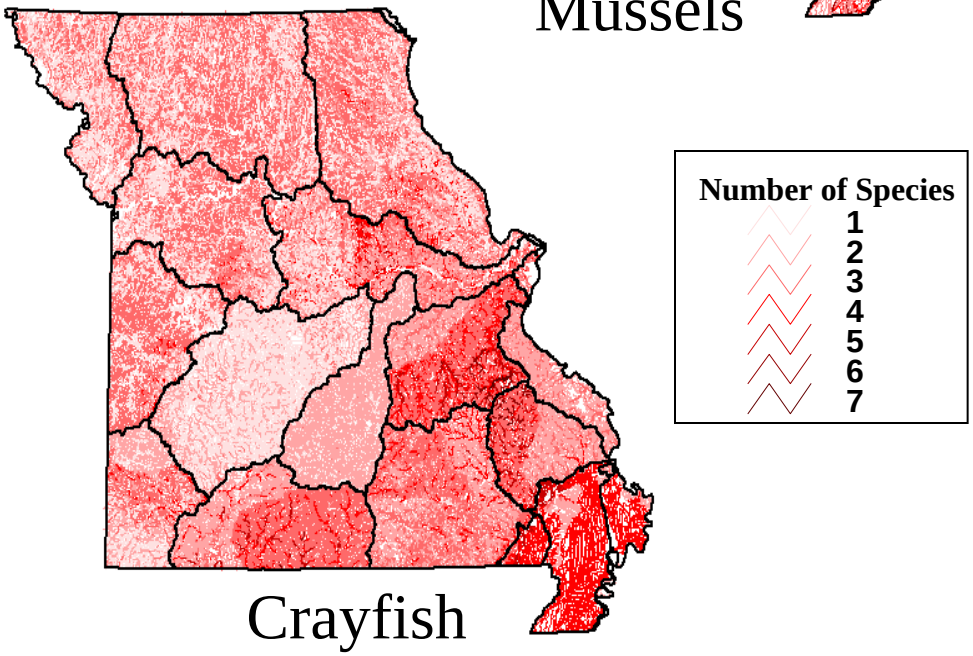
Comparing Patterns of Richness



Fish

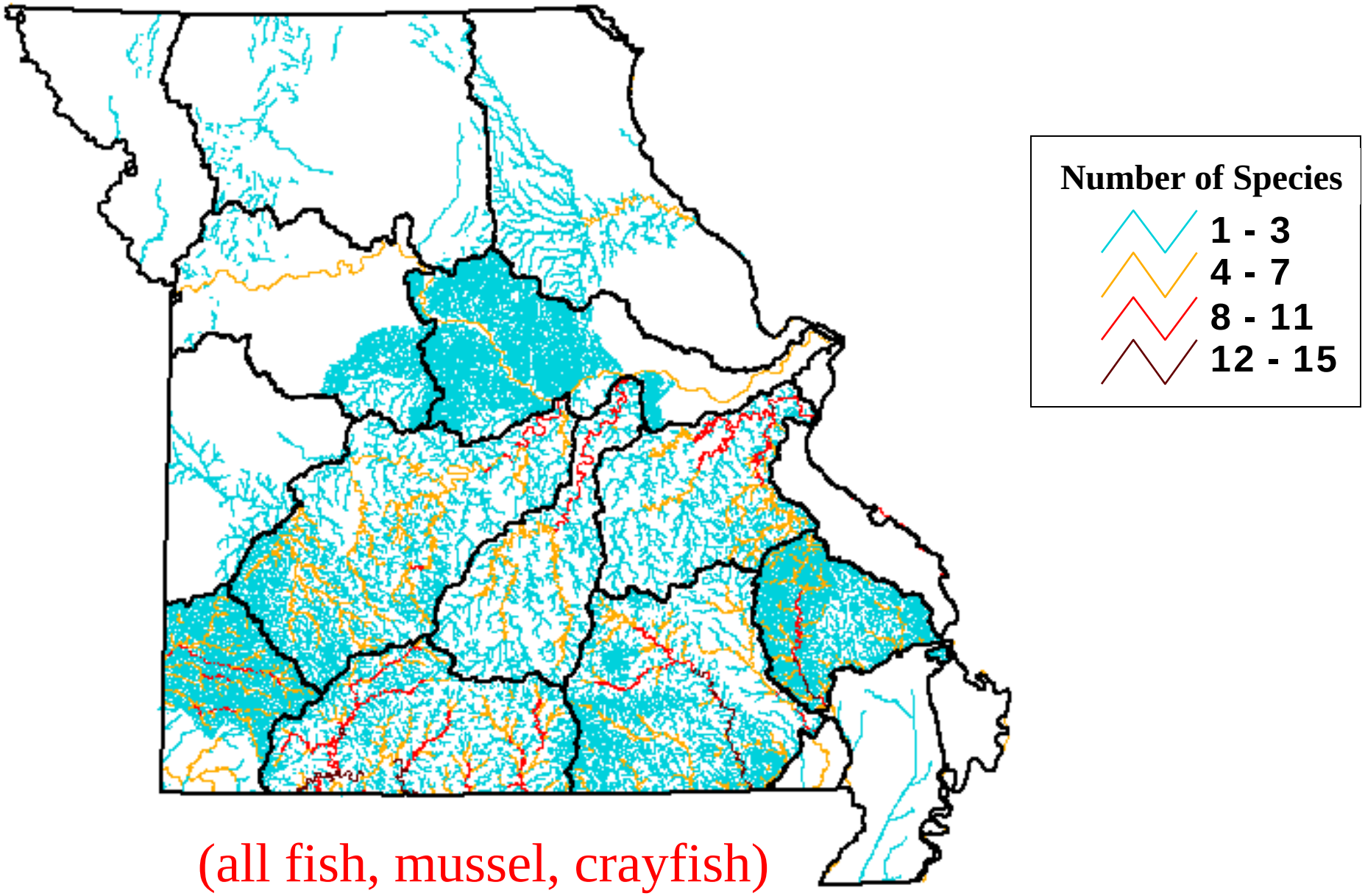


Mussels



Crayfish

Predicted Richness for all
Globally Rare, Threatened, and Endangered Species



(all fish, mussel, crayfish)

Predicted Biological Potential

Plains: 10 Species

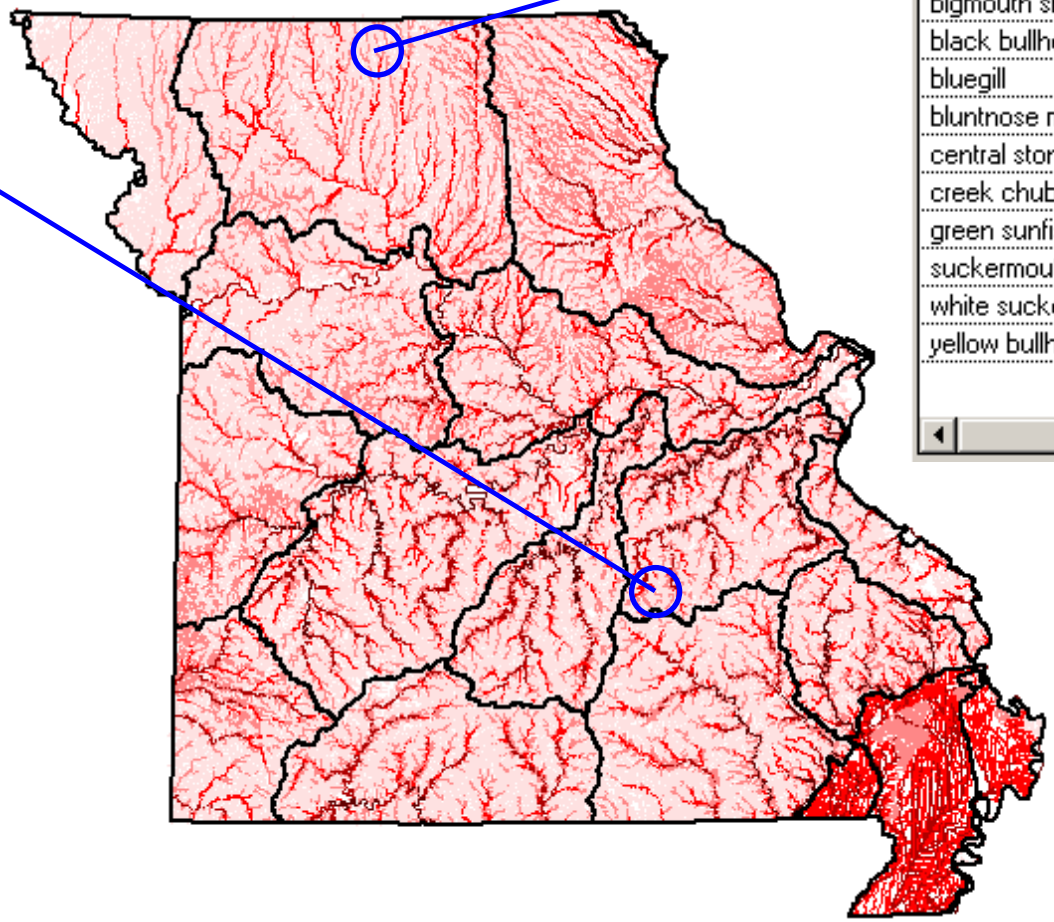
sum7.dbf

Vernacular
bigmouth shiner
black bullhead
bluegill
bluntnose minnow
central stoneroller
creek chub
green sunfish
suckermouth minnow
white sucker
yellow bullhead

Ozarks: 27 Species

sum6.dbf

Vernacular
Ozark minnow
bigeye chub
black bullhead
blackspotted topminnow
bleeding shiner
bluegill
bluntnose minnow
brook silverside
central stoneroller
common carp
creek chub
creek chubsucker
golden redhorse
golden shiner
green sunfish
greenside darter
largescale stoneroller
logperch
longear sunfish
northern hog sucker
northern studfish
plains topminnow
rainbow darter
slender madtom
smallmouth bass



Accuracy Assessment

- Fish
  - 80 independent samples (2001-2002)
- Crayfish
  - 147 independent samples (1991-2002)
- Mussels
  - 51 independent samples (2000-2002)

Taxa	Commission	Omission
Fish	48%	10%
Crayfish	52%	11%
Mussels	64%	6%

Problems with Accuracy Assessment

- Virtually none of samples cover an entire valley segment
  - Typically 100-300m for fish and often shorter stream lengths are surveyed for the other taxa
- All are single samples from one point in time
  - Do not account for temporal differences (e.g., seasonal changes or natural disturbance)
- Sampling efficiencies often very low
- Many samples came from highly degraded streams
  - We are predicting biological potential
- Because of these and other problems we should expect high commission errors

Utility of Predicted Distribution Coverage

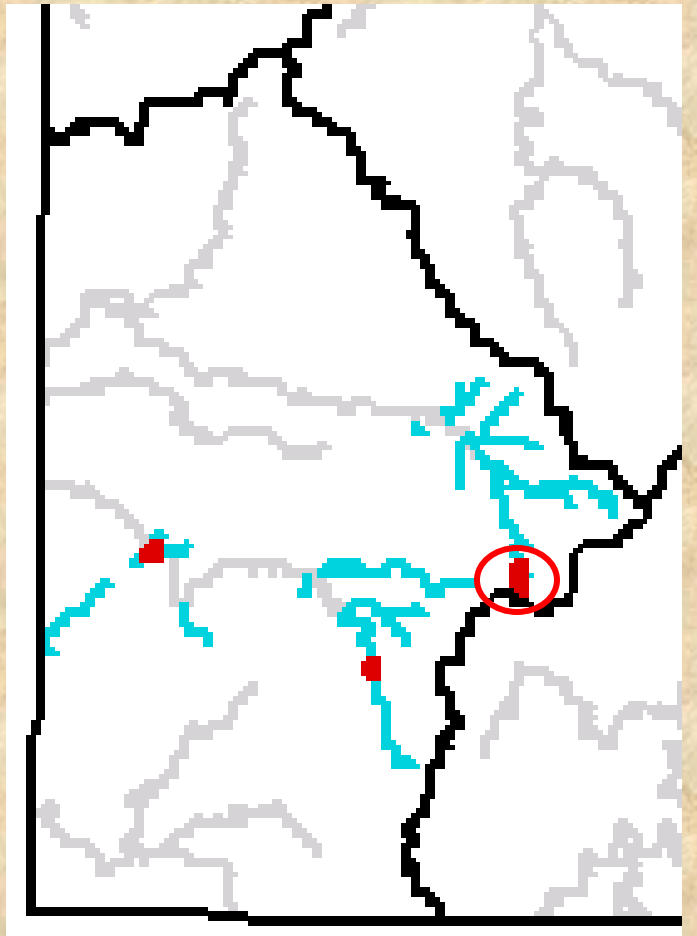
- Identify biotic targets and assess representation of targets
- Part of a decision support system for planning and enforcing regulations (e.g., CWA 401 & 404 permits)
- Enhancing understanding of habitat requirements of species in a standardized fashion
- Identifying information gaps
- Serving as a general reference condition for biomonitoring or restoration efforts
- Identifying additional populations of endangered species

A Success Story

Plains Topminnow
(*Fundulus sciadicus*)

- Found in most headwater streams of the central Ozarks
- 3 historic populations in Neosho EDU
- Only 1 thought to remain
- Predicted to occur in 15 other streams
- Biologists made collections from 11
- Found 4 additional populations

Predicted Distribution
Neosho EDU



Limitations/Opportunities for Improvement

- Lack of sampling data
- Lack of detailed environmental data
  - Especially temperature
- Models use proximate variables and are purely associative
- Do not incorporate abundance data
- Do not account for watershed variables
- Do not account for human disturbance
- Do not account for biotic interactions
- Do not account for ontogenetic shifts in habitat use
- Do not account for spatial autocorrelation

Summary

- We cannot directly measure biodiversity so distributional data for select species provide us with surrogate targets
- We need spatially comprehensive data at the finest spatial grain of our assessment (valley segment)
- We cannot sample the biota of every segment and most existing data is spatially and temporally biased
- We must rely on predictive modeling
- Many different approaches to modeling species distributions
- Nearly impossible to develop literature-based models
- Decision Tree Analysis proved to be relative efficient and ?accurate?
- There is significant room for improvement
- Species distribution models are purely correlative, but they move us beyond making informed decisions only at sampled locations

Acknowledgements

USGS National Gap Analysis Program

U.S. Department of Defense Legacy Program

Missouri Department of Conservation

University of Missouri

Missouri Cooperative Fish and Wildlife Research Unit

MoRAP Partner Agencies and Personnel