Resource materials for a GIS spatial analysis course
Revision of Lectures

By Gary L. Raines(1)
Open-File Report 01-221
Version 1.1
2006

Manuscript approved 2001
U.S. DEPARTMENT OF THE INTERIOR
U.S. GEOLOGICAL SURVEY
(1) U.S. Geological Survey, Reno, NV 89557

Course Outline - Geography 701M

Advanced Geographic Statistical Methods (Spatial Modeling)

Professor: Dr. Gary L. Raines
Office: Laxalt Mineral Research 271, telephone 784-5596, email: graines@usgs.gov

Course Goals and Objectives: Introduction to the techniques of modeling and spatial analysis of non-deterministic processes in GIS for geographers and natural scientists.

The goal of this class is to introduce the concepts of modeling in which multiple categorical and ordered spatial-data sets are combined to predict the distribution or occurrence of the product of some complex process. Examples of the types of applications addressed might be predictive models of animal habitat, occurrence of infectious disease, or undiscovered mineral resources. These types of models all have the characteristic that the processes involved are complex and sometimes poorly understood, that is the models are not prescriptive, but are often fuzzy or probabilistic in nature.

We will use ArcGIS 9.1 and the Spatial Analyst extension with the Spatial Data Modeler extension (ArcSDM 3.1). This will require the student to be familiar with ArGIS 9 and Spatial Analysis. Students will create simple to complex models using software to gain experience in the process of modeling complex natural science processes. Exercises will work toward the types of multi-disciplinary problems that are common in land management or natural resources organizations. Self directed exercises using available data are utilized.

The class will be a combination of lectures and student-lead discussions. In addition, students will present results of exercises to the class.

Schedule

Lecture – 2 hours per week
Laboratory – Three hours per week minimum in 222 or 221 Mackay Science Hall (Geography GIS computer lab. The GIS laboratories are open from 8am to 5pm Monday through Friday. Software and data will be available in both rooms. From Data Works Computer Laboratory in Getchell Library should be able to access the class materials and we plan to get the software there in a few weeks. **Students need to meet with GIS laboratory manager, Patrick Guiberson in room 224 in Mackay Science to get a login for this class and an update on GIS laboratory policy. Patrick has office hours from 11 am to Noon, Monday through Thursday.** All of the exercises for this class can be done with Arcview 3 with the Spatial Analyst, and Spatial Data Modeler (ArcSDM) extensions. The ArcSDM extension is available on the class folder. I am currently developing an ArcMap version of ArcSDM. It should be available for many of the exercises, but it is not yet fully debugged and tested.

Office Hours: to be arranges, 271 Laxalt Mineral Research. I maintain an open door policy. When I am in, the door is open. You are welcome to drop by when you have questions.

Textbook: Bonham-Carter, G.F., 1966, Geographic information systems for geoscientists – modeling in GIS: Elsevier Science Inc., New York, 398p. Besides the textbook, journal articles will be read and discussed in student-led discussions.

Assignments

All students will use modeling tools in an increasingly complex series of exercises. Later exercises will require a group of students with differing science backgrounds to form a team to address a problem that requires expertise in several fields of science. Graduate students will be expected to take a leadership position in these multidisciplinary teams to define the task, the approach, to integrate team members, and to write and present the team report.

Assignment 1 – Using ArcSDM 3.1 in ArcMap 9.1 reproduce weights-of-evidence, logistic-regression, fuzzy-logic, and neural-network models for Carlin deposits. The intent of this exercise is for the student to gain familiarity with ArcSDM, the processing steps, and the decisions necessary to calculate these models.
Assignment 2 – Using various statistical measures, compare the maps prepared in Exercise 1.
Assignment 3 – Prepare and compare models of animal habitat in the Tahoe Basin.

Grading

	Geog 701M
Class Participation	10%
Assignment 1	10%
Assignment 2	10%
Assignment 3 Poster	20%
Assignment 3 Report	30%
Examinations	10%
Discussions	10%

Originality, logic, and overall quality of the models will be the primary consideration in grading; but cartographic and oral presentation will also influence the grade.

Additional Requirements for Students Enrolled in Geography 701M

All graduate students are expected to draw on their experience and knowledge gained elsewhere to enhance the formation of connections between the topics covered in this course as well as related topics not explicitly covered in the course. In a sense, this course addresses a philosophy of creating scientific, spatial models. Thus, the students have to integrate their science, statistics, and GIS background to define the spatial problem, the approach necessary to solve a problem, and then present a solution to the problem. Graduate students will be called upon throughout the semester to lead and participate in class discussion related to advanced concepts of the course material.

Students enrolled in 701M will have to prepare a research project and class presentation in relation to Assignment 3. The report provides an opportunity for investigating course subjects at an advanced level. The graduate students' technical presentation increases their ability to speak in from of an audience, and serves as a synthesis experience, combining explicit class material with external independently research information to develop a greater understanding of the subject.

Lecture, Reading, and Exercise Schedule

Date	Subject	PPT	Exercise 1	Exercise 2	Exercise 3	Reading
29-Aug-06	Finland (Reading Ch 9 & Ch 8 (pg 248-258)					Ch.9
31-Aug-06	Finland					Ch.8 (Pg 250-258)
5-Sep-06	Arrive Home					Carlin Exercise
7-Sep-06	Introduction	1	Carlin - WofE			ArcSDM users manual
						King & Kramer,
12-Sep-06	Redlands					Velleman
14-Sep-06	Overview & Demo	2				
19-Sep-06	Patterns & Discussion (King & Kramer, Vellerman)	3				Nova Scotia
21-Sep-06	Multimap Introduction	4				CI_Agterberg
26-Sep-06	Boolean and Index Overlay models	5				(Exploratory Carlin)
28-Sep-06	WofE1	6	Carlin - LR			(Epithermal Gold)
3-Oct-06	WofE2	7	Carlin - FL			
5-Oct-06	Multi-class Generalization	8	Carlin - NN			
10-Oct-06	Carlin WofE Presentations (Part 1 of Exercise 1)		Expert WofE			
12-Oct-06	Expert WofE, LR, FL, and NN Demo	8B			Final Exercise	Logistic Regression
17-Oct-06	Logistic Regression	9				
19-Oct-06	Fuzzy Logic	10			Form Groups	
24-Oct-06	Neural Networks	11				
26-Oct-06	Miscellany	12			Review Data	
31-Oct-06	Exercise 1 Discussion & Progress on Final Exercise					
2-Nov-06	Ch9 Discussion	13		Correlation		Ch. 8
7-Nov-06	Overlay	14			Define Approach	Kappa
9-Nov-06	Correlation1	15				
14-Nov-06	Correlation2	16				
16-Nov-06	Exercise 2 Presentations & Discussion					Fragstats
21-Nov-06	Fragstats	19				Ch 7.
23-Nov-06	Thanksgiving Holiday					
28-Nov-06	Reclassification	17				
30-Nov-06	Filtering	18			Completed Modeling	
5-Dec-06	Summary	21			Prepare Report/Poster	
7-Dec-06	Spatial-Temporal Modeling?	20				(CA)
12-Dec-06	Final Exercise Presentations					

Additional Reading

Agterberg, F.P., Bonham-Carter, G.F., Cheng, Q. And Wright, D.F., 1993, Weights of evidence modeling and weighted logistic regression for mineral potential mapping *in* Davis, J.C., and Herzfeld, U.C. (eds.), Computers in geology, 25 years of progress: Oxford, Oxford University Press, p. 13-32.

Agterberg, F.P., and Cheng, Q., 2002, Conditional independence test for weights-of-evidence modeling: Natural Resources Research, v. 11, no. 4, p. 249-255.

Bonham-Carter, G.F., Agterberg, F.P., and Wright, D.F., 1988, Integration of geological datasets for gold exploration in Nova Scotia: Photogrammetric Engineering and Remote Sensing, v. 54, no. 11, p. 1585-1592.

Begueria, Santiago, 2006, Validation and evaluation of predictive models in hazard assessment and risk management: Natural Hazards, v. 27, p. 315-329.

Brismar, Jan, 1991, Understanding receiver-operatoring-characteristic curves: a graphic approach: AJR, p. 1119-1121.

Chung, C.F, and Fabbri, A.G., 2003, Validation of spatial prediction models for landslide hazard mapping: Natural Hazards, v. 30, p. 451-472.

Cohen, Jacob, 1960, A coefficient of agreement for nominal scale: Educational and Psychological Measurement, v. 20, no. 1, p. 37-46.

Hudson, W.D., and Ramm, C.W., 1987, Correct formulation of the kappa coefficient of agreement: Photogrammetric Engineering and Remote Sensing, v. 53, no. 4, p. 421-422.

King, J.L., and Kraemer, K.L., 1993 , Models, facts, and the policy process: the political ecology of estimated truth *in* Goodchild, M.F., Parks, B.O., and Steyaert, L.T., Environmental modeling with GIS: New York, Oxford University Press, p. 353-360.

Levin, S.A., 1992, The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture: Ecology, v. 73, no. 6, p. 1943-1967.

Raines, G.L., and Bonham-Carter, G.F., 2006, Exploratory Spatial Modelling Demonstration for Carlin- type deposits, Central Nevada, USA, using Arc-SDM *in* Harris, J.R. (editor), GIS applications in earth sciences: Special Publication, Geological Association of Canada, Special Publication 44, p. 23-52.

Raines, G.L., 1999, Evaluation of.weights of evidence to predict epithermal gold deposits in the Great Basin of the western United States: Natural Resources Research, , v. 8, no. 4, p. 257-276.

Rosenfield, G.H., and Fitzpatrick-Lins, Katherine, 1986, A coefficient of agreement as a measure of thematic classification accuracy: Photogrammetric Engineering and Remote Sensing, v. 52, no. 2, p. 223-227.

Velleman, P.F., 1997, The philosophical past and the digital future of data analysis: 375 years of philosophical guidance for software design on the occasion of John W. Tukey's 80th birthday *in* Brillinger, D.R., Fernholz, L.T., and Morgenthaler, S., The practice of data analysis: essays in honor of John W. Tukey: Princeton, Princeton University Press, p. 317-337.

Source of Spatial Data Modeller Extension

http://www.ige.unicamp.br/sdm/default_e.htm

Lectures for a University Semester Course

Modeling in GIS
Dr. Gary Raines

Insights Through Integration

•Geography 701M – UNR

Gary Raines

- USGS Research Geologist
- Remote Sensing applications to mineral exploration
- Development of techniques for spatial modeling in mineral and environmental applications
- Focus on large areas

Course Outline

- Lecture schedule
- Three Laboratory Assignments
- Examinations
- Reading
  - Geographic Information systems for geoscientists - modeling in GIS: Chapters 7, 8, and 9
  - Additional reading - student lead discussion

Grading

Task	
Class Participation	10%
Assignment 1	10%
Assignment 2	10%
Assignment 3 Poster	20%
Assignment 3 Report	30%
Examinations	10%
Discussions	10%

Laboratory Assignments

- Assignment 1 – Reproduce the weights-of-evidence, logistic-regression, fuzzy-logic, and neural-network models for Carlin deposits.
- Assignment 2 – Using various statistical measures, compare the maps prepared in Exercise 1.
- Assignment 3 – Prepare and compare models of animal habitat in the Tahoe Basin.

Class Participation

- What I know is obvious!
- Your job is to ask questions!

Schedule

Examinations

- Take home
- Short essays
- Probably will be one at the end of Chapter 7 and Chapter 8
- Presentation and report of third exercise will serve as final.

Discussions

- Journal articles will be assigned to enhance material in book.
  - Discussion of these articles will be lead by students.
- Laboratory assignments will be presented and discussed in class by students.

Goals and Expectations

- To introduce the concepts and process of spatial modeling in GIS for geographers and natural scientists.
- Emphasis on probability and favorability models, that is nondeterministic models.
- Students are GIS experts!

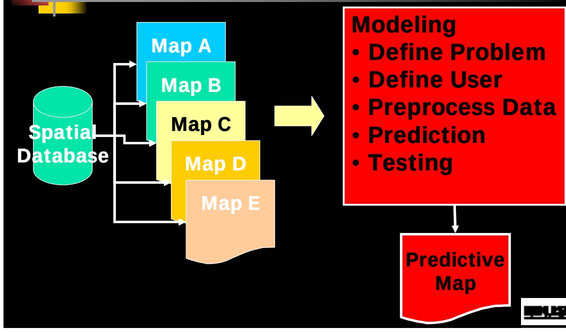
What is a model?

- A simplification of nature.
- A representation of a set of objects and their relationships.
- A model is a way of describing something that cannot be directly observed.
- A model is a way of communicating complex ideas.

Why Model?

- "...when you can measure what you are speaking about and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of science, whatever the matter may be." Lord Kelvin
- GIGO "Garbage In, Garbage Out"

Modeling in GIS



Reading Assignment

- Read King and Kramer
  - Why models?
  - Volunteer to present?
- Read Velleman
  - Philosophy of data analysis
  - Volunteer to present? Select and explain 3 most important points.
- Chapter 9

Laboratory Assignment

- Objective – To gain familiarity with ArcSDM, the processing steps, and the decisions necessary to calculate these models.
- Data – Carlin exercise
  - Arcview 3 – Carlin_AV3.zip
  - ArcMap – Carlin_ArcMap83.zip
  - Carlin Exercise – Carlin_Exercise.pdf

Laboratory Exercise

- Reproduce the weights-of-evidence, logistic-regression, fuzzy-logic, and neural-network models for Carlin deposits.
  - WofE mathematics understood by lecture on Oct 3
- Present your results as a short, 8.5x11 page-size report. Include a concise summary of processing steps for a knowledgeable user.
  - Example report – Nova Scotia and Nevada Epithermal Gold papers
  - WofE presentation Oct 10
  - Final Report due Oct 31

Challenges in this exercise

- How to process the data in ArcGIS and to report the results elegantly.
  - How to use ArcSDM while the lectures and reading give you an understanding of the mathematics and decision process.
- How to concisely summarize the processing sets.
  - Assume a knowledgeable ArcGIS and ArcSDM users, such as yourself.

Laboratory Grading

- Elegant solution
- Logical thinking
- Quality of writing and cartography
  - Concise writing is a virtue.
  - Tell the reader only what is important.

Background

- Measurement Scales
- Precision
  - Integers versus Real numbers
- Map Scale and Resolution
- Guidelines for modeling

Measurement Scales

Free

Ordered

- Nominal (Categorical)
  - An unordered label of categories or classes.
- Ordinal (Rank)
  - Measurements ordered (ranked) according to relative position on a scale with unequal intervals between classes.
- Interval
  - Measurements that can be labeled and ordered with an equal interval between classes but without a true zero.
- Ratio
  - Measurements that can be labeled and ordered, with an equal interval between classes, and with a true zero.

Examples of Measurement Scales

Scale Type	Examples	Operations	Means
Nominal	Rock type	=	Mode
Ordinal	Relative age	><	Median
Interval	Temperature	+-*/	Mean
Ratio	Distance	+-*/	Mean

Precision = a measure of ability to distinguish between nearly equal numbers.

- The number of significant figures determines how maps can be reclassified and symbolized.
- Integers versus real numbers in ArcGIS
  - Integer and Real valued grids can be classified by various methods that all have assumptions about the data.
  - Integer grids always have VAT or, simply, an attribute table.

Map Scale and Resolution

Map Scale	Base Resolution	Information Resolution	Buffer?
1:2,500,000	1250	2500	5000
1:500,000	250	500	1000
1:250,000	125	250	500
1:100,000	50	100	200

Units - Meters

Base Resolution ~ Scale denominator / 2000

Guidelines for Modeling

- Formal statement of the problem.
- Define the user of the model.
- Specification - preprocess the data to provide useful information, that is evidence.
  - Data exploration
  - Data transformation, filtering, and scaling
  - Reduce the dimensionality by eliminating redundant or correlated information
  - Use the minimum information necessary
- Prediction - combine the evidence to create the model.
  - A type of multidimensional data exploration.
- Testing - evaluate the model and it's properties.

Properties of Evidence

- Selected attributes must discriminate between one or more classes of objects.
- Selected attributes must not be correlated with other attributes to any moderately strong extent.
- Selected attributes must have meaning for humans.

Scientific Method

- Define a problem
- Gather pertinent data
- Form a working hypothesis or explanation
- Do experiments to test the hypothesis
- Interpret the results
- Draw a conclusion and modify the hypothesis as needed.

Occam's Razor

Occam's razor states that a person should not increase, beyond what is necessary, the number of entities required to explain anything, or that the person should not make more assumptions than the minimum needed.

This principle is often called the principle of parsimony.

Questions have been raised, however, as to whether a person can determine without any doubt that given entities or assumptions are not needed in an explanation. Unless this determination can be made, it is impossible to tell with complete certainty when the principle can be applied

Abstracted from the Grolier Encyclopedia.

Spatial Analysis in GIS Overview

- Examples of Nondeterministic Spatial Models
- Demonstration of ArcSDM

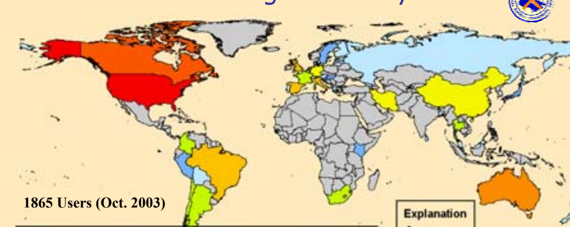
Information

- Email addresses
- Patch for Arcview 3 on XP operating systems

Schedule Revised

Date	Subject	PPT	Exercise 1	Exercise 2	Exercise 3	Reading
1/21/2004	Introduction	1	Carlin - WofE			King&Kramer, Velleman
1/26/2004	Overview & Demo	2				Ch.9
1/28/2004	Patterns & Discussion	3				
2/2/2004	Multimap Introduction	4				Nova Scotia
2/4/2004	Boolean and Index Overlay models	5				(Exploratory Carlin)
2/9/2004	WofE1	6				(Epithermal Gold)
2/11/2004	WofE2	7	Carlin - LR			
2/16/2004	President's Day Holiday		Carlin - FL			
2/18/2004	Multi-class Generalization	8	Carlin - NN			
2/23/2004	Carlin WofE Presentations		Expert WofE			
2/25/2004	Expert WofE, LR, FL, and NN Demo				Final Exercise	
3/1/2004	Logistic Regression	9				
3/3/2004	Fuzzy Logic	10				
3/8/2004	Neural Networks	11				
3/10/2004	Spring Break					
3/15/2004	Spring Break					
3/22/2004	Miscellany	12				
3/24/2004	Exercise 1 Discussion & Progress on Final Exercise					Ch. 8
3/29/2004	Ch9 Discussion	13		Correlation		Kappa
3/31/2004	Overlay	14				
4/5/2004	Correlation1	15				
4/7/2004	Correlation2	16				
4/12/2004	Correlation Presentations & Discussion					Ch 7.
4/14/2004	Reclassification	17				
4/19/2004	Filtering	18				
4/21/2004	Fragstats	19				FRAGSTATS
4/26/2004	Spatial-Temporal Modeling?	20				CA
4/28/2004	Summary	21				
5/3/2004	Final Exercise Presentations					

ArcSDM Usage Summary 2003



1865 Users (Oct. 2003)

Applications	Count	Applications	Count
Anthropology	3	Hazards/landslides/earthquakes	67
Archeology	133	Hydrology	71
Business applications	15	Land Management	91
Crime prevention	11	Marine Research	16
Ecology/Habitat studies	254	Medical Geography	12
Education	190	Mineral Exploration	236
Epidemiology	46	Mineral Potential Mapping	111
Forestry	46	Not Specified - other	298
Geochemistry	18	Other Geological Application	92
Geology	166	Public health	16
Geophysics	25	Soils	25

Additional Materials

- Raines, G.L., 2001, Resource materials for a GIS spatial analysis course: U.S. Geological Survey Open File Report 01-221, http://geopubs.wr.usgs.gov/open-file/of01-221/, 216p, four zip files of software and class exercises, and a zip file of student posters.
- Exploratory Carlin zip file

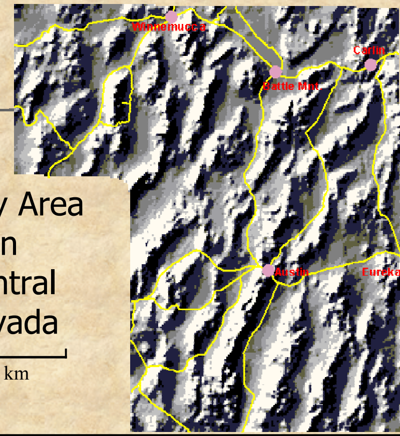
Points of Demonstration

- What to do with data and why.
- There may be no right way to analyze any particular data!
- There are often several ways to analyze data that are good!
- Data analysis is like doing an experiment.

Paraphrased from Velleman, 1997

Demonstration of ArcSDM

- Weights of Evidence and Logistic Regression
- Fuzzy Logic
- Neural Networks



Study Area in Central Nevada

75 km

Spatial Analysis in GIS Overview Continued

•Modeling - Pattern Recognition

•Discussion of King & Kramer and Velleman

Data Exploration

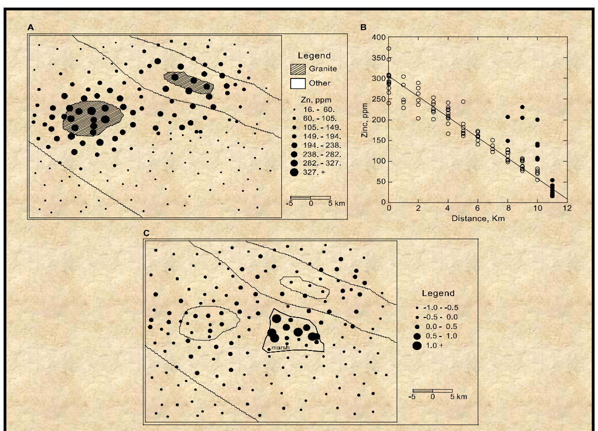
- Process of seeking patterns on maps that help predict spatial phenomena.
  - Visualization leads to recognition of a pattern and the association of the pattern with something of interest.
  - A model is proposed that describes the association.

Data Exploration

- Seeking patterns involves:
  - Measurement
  - Statistical Summary
  - Visualization
  - Description
  - Understanding of processes causing pattern
- Foundation is data model.

Pattern

- An area having a consistent, recognizable characteristics associated with some object or process.
  - A pattern is something that deviates from the norm.
  - A pattern is associated with a particular scale of observation!
  - It is a primitive.
- Association of patterns and their causes are the bricks of scientific knowledge.



Bonham-Carter, 1996

Types of Recognition

- Classification is the process of grouping objects together in classes according to perceived similarities.
- Identification is the recognition of an individual object as a unique singleton class.
- Discrimination is the recognition that an individual object as different from a class.

Recognition of a Pattern

- Task - Determine what the appropriate level of aggregation and simplification is for the problem at hand, a problem of reclassification.
  - Aggregation and simplification are tied to scale of observation.
  - There is no single scale at which to view a system.
  - Does not mean that all scales serve equally well or there are not scaling laws.
- Description of patterns is the starting point.
- Spatial models start with an assemblage of patterns and associated processes.

Measurement Scales

- Nominal (Categorical)
  - An unordered label of categories or classes.
- Ordinal (Rank)
  - Measurements ordered (ranked) according to relative position on a scale with unequal intervals between classes.
- Interval
  - Measurements that can be labeled and ordered with an equal interval between classes but without a true zero.
- Ratio
  - Measurements that can be labeled and ordered, with an equal interval between classes, and with a true zero.

King and Kramer Modeling Continuum

Engineering	↔ Science ↔	Public Policy Decision Making
How to build a bridge at location X?	Which location is best?	Do we need a bridge somewhere?
Facts Clear Precise	↔	Facts Unclear Imprecise
Deterministic	↔	Probabilistic Fuzzy

King and Kramer

- Models are most useful when the right answer is not clear.
- Modeling clarifies the issues of debate in evaluation of an answer.
- Modeling enforces a discipline of analysis, discourse, and consistency.
- Models provide a powerful form of “advice”, that is not “truth”, but a refined result of a particular viewpoint.

Velleman – Top 3 Points

- Aphorism 3 – Iterative learning leading to understanding.
- Aphorism 7 – Keep it simple!
- Aphorism 14 – Multiple working hypotheses.

Tools for Map Analysis
Multiple Maps

Boolean Logic
Index Overlay (Weighted Overlay)
Fuzzy Logic
Weights of Evidence
Logistic Regression
Neural Networks

Reading Assignment

- Chapter 9
  - Look over whole chapter and pages 250-258.
  - Boolean Logic
  - Index Overlay (Weighted Overlay)
  - Bayesian Models (Weights of Evidence)
  - Logistic Regression
  - Fuzzy Logic
- Other Papers
  - Nova Scotia: Lecture 9/28 (WofE_NovaScotia.pdf)
  - Logistic Regression (WofE_LogisticRegression.pdf)
  - Fuzzy knowledge representation (Fuzzy Logic Chapter - Report.pdf)

USGS

Additional Reading

- Epithermal Gold (Nevada_Epithermal_Gold.pdf)
- Exploratory Carlin: (060117_GIS44-2.pdf)
- Fuzzy Logic (060117_GIS44-2.pdf)
- Neural networks (RBFLN_ArcSDM1.pdf)

Guidelines for Modeling

- Formal statement of the problem.
- Define the user of the model.
- Specification - preprocess the data to provide useful information, that is evidence.
  - Data exploration
  - Data transformation, filtering, and scaling
  - Reduce the dimensionality by eliminating redundant or correlated information
  - Use the minimum information necessary
- Prediction - combine the evidence to create the model.
- Testing - evaluate the model and it's properties.

USGS

Purpose of GIS Projects

- Combine data from diverse sources
- To describe and analyze interactions
- To make predictions, that is models
- To provide support for decision makers

USGS

Properties of Evidence

- Selected attributes must discriminate between one or more classes of objects.
- Selected attributes must not be correlated with other attributes to any moderately strong extent.
- Selected attributes must have meaning for humans.

USGS

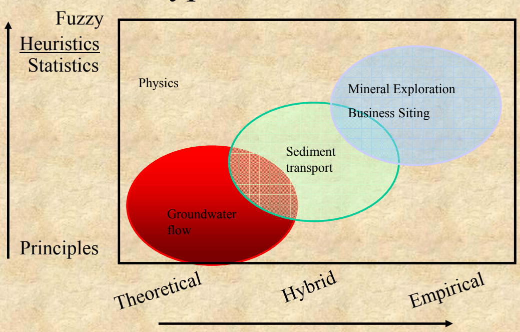
Scientific Method

- Define a problem
- Gather pertinent data
- Form a working hypothesis or explanation
- Do experiments to test the hypothesis
- Interpret the results
- Draw a conclusion and modify the hypothesis as needed.

Types of Models

- Prescriptive or Deterministic
  - Application of good technical practices
  - Process: Boolean rules, Equations, Index Overlay
  - Output: Binary map (yes or no), User defined range such as 0 to 10
- Predictive
  - Application of mathematics to represent how people think about the evidence but cannot represent as equations.
  - Process: weighting of evidence and combination of weights
  - Output: Favorability, probability, or fuzzy map [0 to 1]

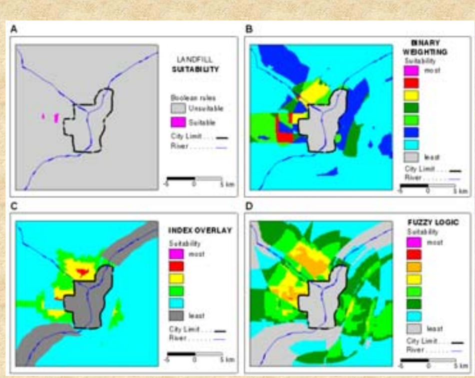
Types of Models



Knowledge Driven Methods

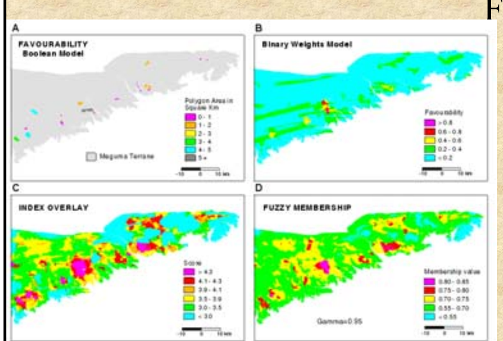
- **Boolean Logic** - True/False representation of maps with all maps rated equally. Simple method with True/False answer.
- **Index Overlay with Binary Maps** - Maps are given different weights. Linear combination of maps. Can use Weighted Overlay tool.
- **Index Overlay with Multi-Class Maps** - Maps are given different weights as well as the classes of the maps are given different weights. Linear combination of maps. Can use Weighted Overlay tool
- **Fuzzy Logic** - More flexible weighting of maps and map classes. Nonlinear combination of maps.

Fig 9-2

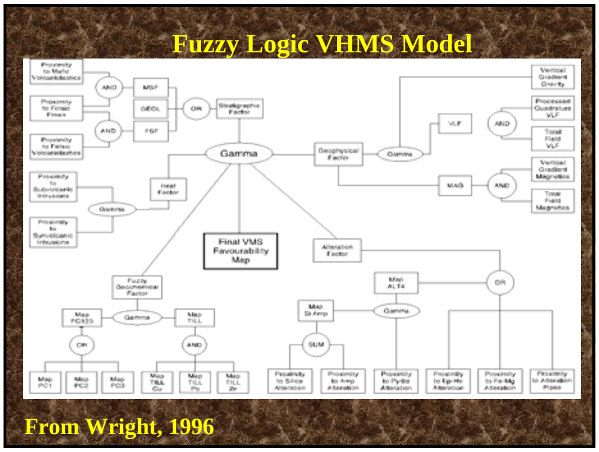


A: Boolean B: Binary Index Overlay C: Multi-class Index Overlay D: Fuzzy logic

Fig 9-4

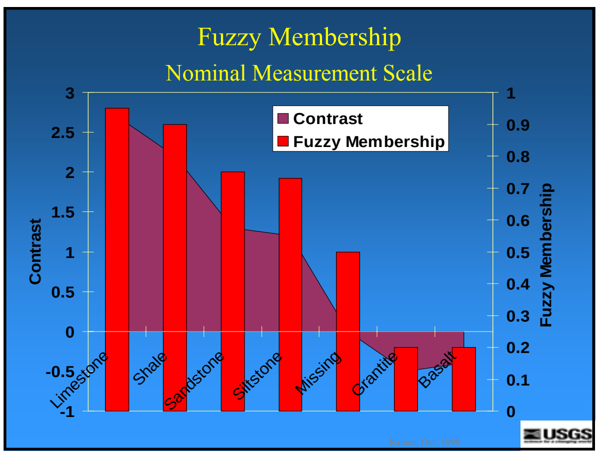


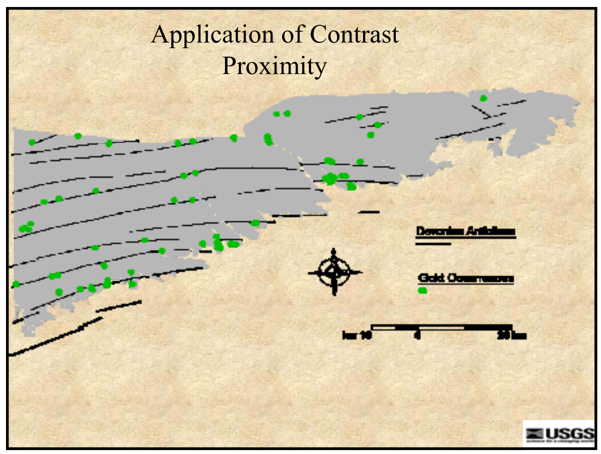
A: Boolean B: Binary Index Overlay C: Multi-class Index Overlay D: Fuzzy logic

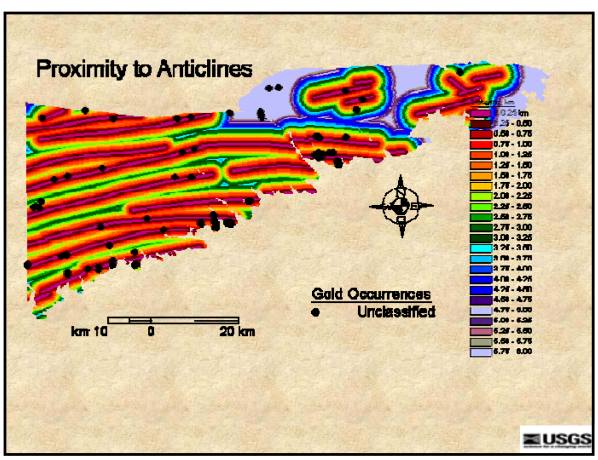


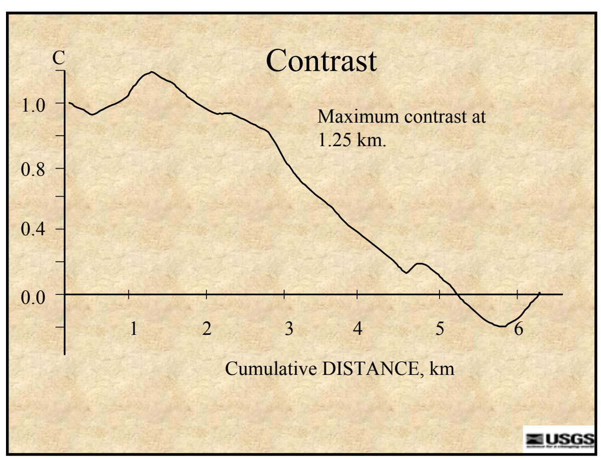
Data Driven Methods

- Weights of Evidence
  - log linear combination of maps, simplest with binary maps.
  - Classifies areas by probability or favorability of occurrence of a training site.
  - Model parameters easy to understand.
- Logistic Regression
  - log regression combination of binary maps
  - Classifies areas by probability of occurrence of a training site.
  - Model parameters complex.
- Neural networks
  - Experimental, nonlinear combination of fuzzy or rescaled maps
  - Classifies areas by fuzzy membership in training set.
  - Can also be self organizing to produce fuzzy membership.
  - Model parameters complex.

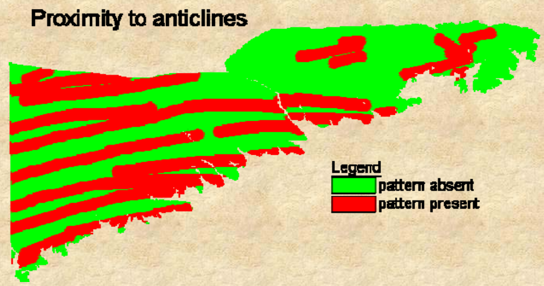






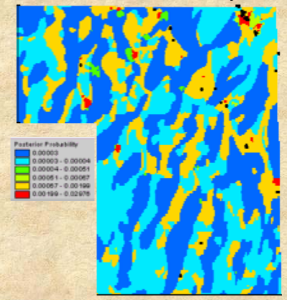


Binary Reclassification of Anticline Proximity

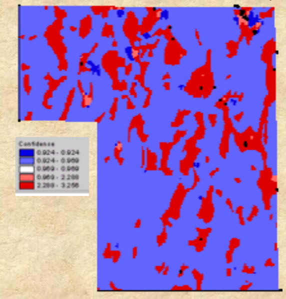


Variance and Uncertainty

Posterior Probability



Studentized Posterior Probability

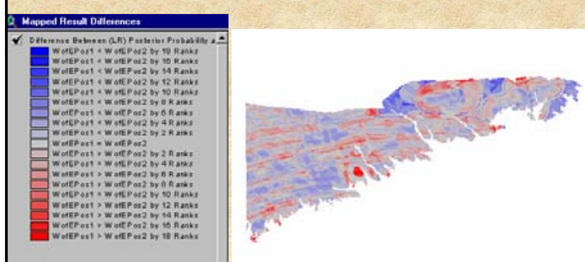


Compare Results

- ArcSDM Post Processing (Classes)
  - Spearman Correlation Coefficienct
  - Map of Rank Differences
  - Quantile-Quantile Plot
- Spatial Analyst Tools/Multivariate
  - Band Collection Statistics
    - Covariance and Pearson's Correlation Coefficient (aka Product Moment Corelation Coefficient) matrices

Compare Results

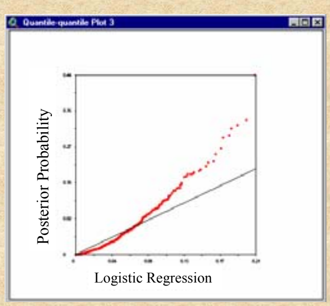
Map of Rank Differences



Bonham-Carter, 1999

Compare Results

Quantile-quantile Plot



Bonham-Carter, 1999

Tools for Map Analysis
Multiple Maps

Boolean Logic
Index Overlay
Fuzzy Logic
Weights of Evidence
Logistic Regression
Neural Networks

Knowledge Driven (Boolean Logic, Index Overlay, Fuzzy Logic)

Data Driven (Weights of Evidence, Logistic Regression, Neural Networks)

Boolean Operators

- And - Returns True (= 1) only if all are true
  - Logical intersection
- Or - Returns False (= 0) if all are false, otherwise returns True (= 1)
  - Logical union
- Xor - Returns True (1) if one and only one is true.
- Not - Negates the operation

Examples

- 1 and 1 = 1
- 1 and 0 = 0
- 0 and 0 = 0
- 1 or 1 = 1
- 1 xor 1 = 0
- 1 or 0 = 1
- 1 xor 0 = 1
- 0 or 0 = 0
- 0 xor 0 = 0
- 0 or (not 0) = 1

- For Boolean operators, an input of zero (0) equals False.
- Any other number is True.
- -3 and 2 = 1
- 2 and 0 = 0
- -3 and 2 and 12 = 1
- -3 or 2 or 12 = 1
- -3 or 2 or 0 = 1
- -3 xor 0 xor 0 = 1

Landsite Selection
Statement of the Problem

1. Be in an area where unconsolidated surficial material is more than a minimum thickness, AND
2. Be in material that has a low permeability, AND

etc.

Example on page 272 of text.

Boolean Map Algebraic
Statement of the Problem

: At current location, determine if conditions for each input are satisfied
: The conditions, C1 to C2 are either TRUE (=1) or FALSE (=0)
: See Table 9-5 for a summary of the map classes
C1 = class('OVERTHIK')>4
etc.
C10 = class('ECOLOG') = = 1
:Combine conditions with Boolean "AND" operator
: The variable OUTPUT is either TRUE (=1) or False (=0)
OUTPUT = C1 AND C2 AND … AND C10
: Map results as a binary 2-class map
RESULTS(OUTPUT)

Portion of calculation on page 273.

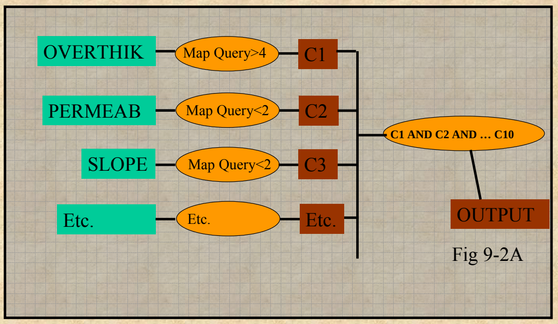
Translate class into ArcGIS

- **C1 = class('OVERTHIK') > 4**
  - OVERTHIK is an integer grid
  - Returns TRUE (= 1) if OVERTHIK > 4; otherwise returns FALSE (= 0)
  - **ArcMap 9.1:**
    - Spatial Analysis/Raster Calculator

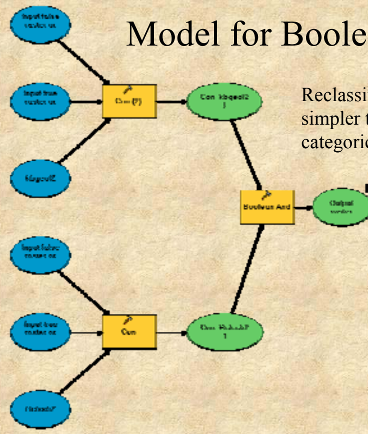
      CON([OVERTHIK] > 4, 1, O)
    - Spatial Analyst/Raster Calculator

      [OVERTHIK] > 4
    - Spatial Analyst/Reclassify
    - Geoprocessing – Weighted Overlay

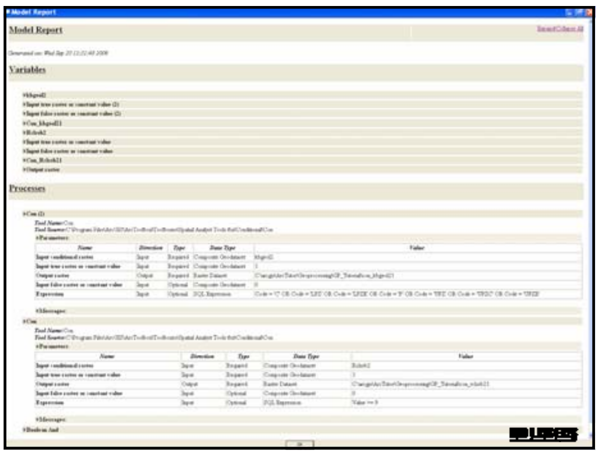
Inference Net for Landfill Site Boolean Logic



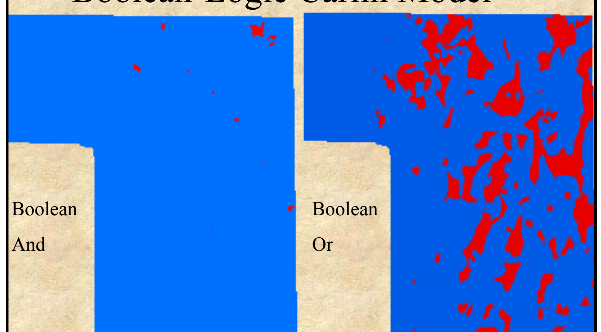
Model for Boolean Logic



Reclassify is probably a simpler tool than Con for categorical evidence.

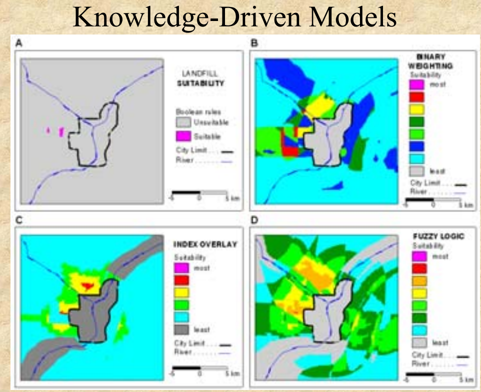


Boolean-Logic Carlin Model



Red = 1 Blue = 0

Knowledge-Driven Models



A: Boolean B: Binary Index Overlay C: Multi-class Index Overlay D: Fuzzy logic

Bonham-Carter, 1996, Fig 9-2

Decisions for Boolean Logic

<u>Reclassify Attributes and Map Interactions</u>

- Thresholds
  - Greater than some value
    - Distance from some feature
    - Some high measured value (e.g. slope > 20)
  - Less than some value
    - Some measured low value (e.g. thickness < 4)
- Equal or Not Equal to some named class
- How the criteria (maps) interact
  - AND, OR, XOR, NOT

Boolean Logic Summary

- Advantages
  - Models are simple.
  - Where prescriptive guidelines from law, Boolean combinations are practical and easily applied.
- Disadvantages
  - All evidence (Maps) are treated equally.
  - A weak representation of how people think about spatial problems
  - Output is binary, either Suitable or Not Suitable.

Index Overlay

$$Score = \frac{\sum_{i=1}^{n} w_i * s_{ij}}{\sum_{i=1}^{n} w_i}$$

Where

w_i = weight of Map I

For binary-class maps, s_{ij} is either 1 for true or present or 0 for false or absent. Score ranges between 0 and 1.

For multi-class maps, s_{ij} is the score or weight assigned to a particular attribute. Score is average score ranging between minimum and maximum weights.

High scores indicate more favorable places.

Index Overlay Algebraic Statement of the Problem

```
: Calculate normalization sum
        SUMW = 3 + 4 + 5 + 3 + 2 + 4 + 5 + 4 + 2 + 1
: Define a variable to name the row
        ROW = class('BASIN')
:  For current location, determine map weights
        M1 = 3 * (class('GEOL') = = 1 OR class('GEOL') = = 2)
        M2 = 4 * table('BASIN', ROW,'AS') > 30
        M3 = 5 * table('BASIN', ROW, 'SB') > 0.8
        etc.
: Calculate normalized sum of weight factors
        NEW = (M1 + M2 + M3 … + M10)/SUMW
: Classify and map output
        NEWMAP = CLASSIFY(NEW,'BINWT')
        RESULTS(OUTPUT)
```

Portion of calculation on page 287, Mineral model.

Translate table into ArcGIS

- **M1 = 4 * table('BASIN', ROW,'AS') > 30**
  - Basin is an integer grid with multiple attributes. ArcGIS 9.1 does not do this. Will be in ArcGIS 9.2
  - Returns TRUE (= 4) if AS > 30; otherwise returns False (= 0)
  - Arcview 3.0 (Something like this in ArcGIS 9.2)
    - Analysis/Map Query

      ([BASIN.AS] > 30.AsGrid)*4.AsGrid
    - Returns 4 if TRUE and 0 if FALSE, but will be labeled TRUE(1) and FALSE (0), respectively.
  - ArcMap 9.1 (Arsenic raster)
    - If had a real or float grid, that is only one attribute (Value), can use the same procedure. If want an integer result, may have to appropriately use **Int()** in the equation.
    - Can also use the longer form in the Raster Calculator of the Boolean example (con statement).

      CON([AS] > 30, 1, 0)*4
      - Returns 4 if TRUE and 0 if FALSE. May need to use **Int()** function
    - Spatial Analysis/Reclassify, specially for categorical data
    - Geoprocessing Weighted Overlay tool

From Mineral Model page 287.

Inference Net for Landfill Site Binary Index Overlay

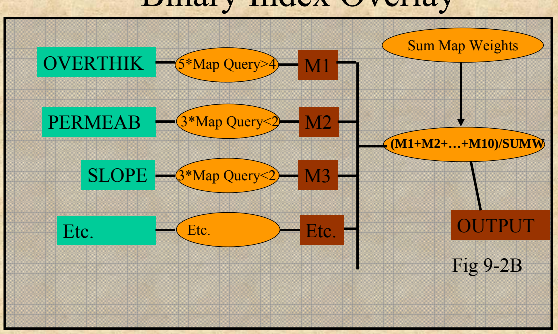
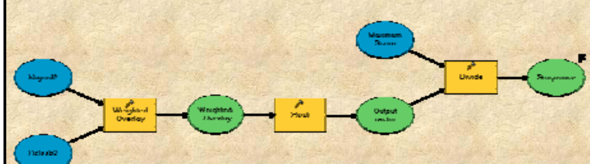


Fig 9-2B

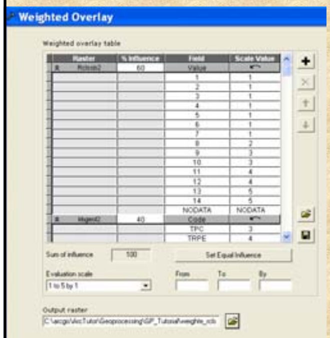
Model for Multi-class Index Overlay



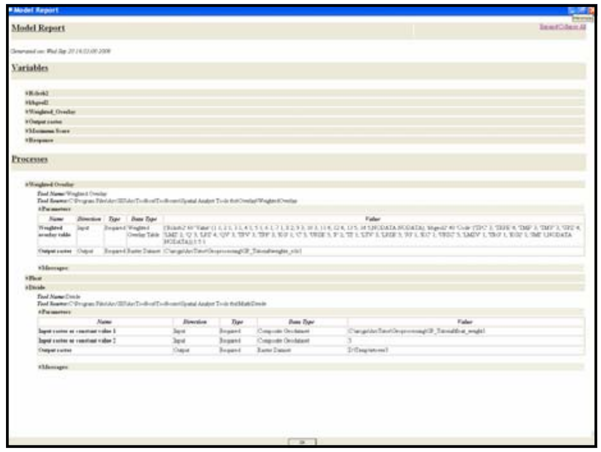
Used Scores [1,5]

Divide by 5 to rescale [0,1]

Weighting Scheme Example

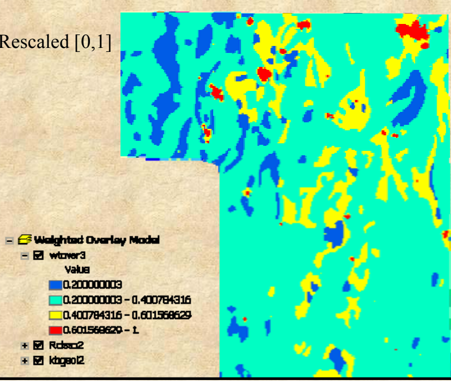


- Multi-class weighting or scoring scheme
- Influence defines the importance of each evidence layer
- Note influence is percent and scoring scheme is same for all evidence. Thus maximum response is fixed, 5 in this case.



Weighted Overlay Response

Rescaled [0,1]

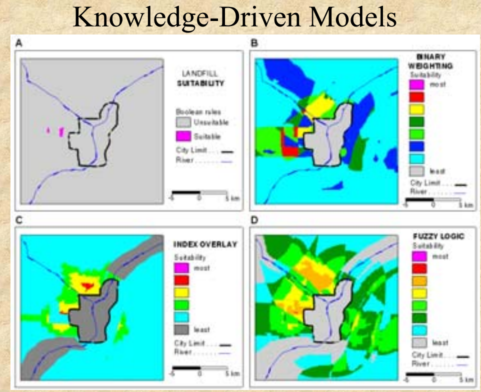


Decisions for Index Overlay

Weights for Attributes and Maps

- Thresholds
  - Greater than some value
    - Distance from some feature
    - Some high measured value (e.g. slope > 20)
  - Less than some value
    - Some measured low value (e.g. thickness < 4)
- Equal or Not Equal to some named class
- How the criteria (maps) interact
  - Weight individual maps. What is the value of each criteria (map)?
  - Summation

Knowledge-Driven Models



A: Boolean B: Binary Index Overlay C: Multi-class Index Overlay D: Fuzzy logic

Bonham-Carter, 1996, Fig 9-2

Index Overlay Summary

- Advantages
  - Weights for individual maps and attribute values allows for better representation of experts opinion of the data.
  - By adjusting weights of maps and attributes can evaluate many different scenarios.
  - Output is a ranking of suitability, which gives decision makers more flexibility.
  - Scaling of Output is by reclassification, an expert decision.
- Disadvantages
  - Linear additive nature is greatest disadvantage.

Model Complexity

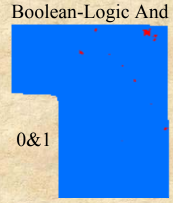
Simple

- Boolean Logic does binary, logical reclassification of evidential layers (maps).
- Binary Index Overlay adds relative weighting of evidential layers (maps).
- Multi-Class Index Overlay adds relative weighting of an attribute or attributes of each evidential layer (map).

More Complex

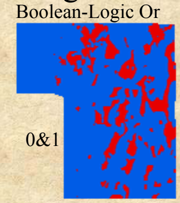
Knowledge-Driven Models

Boolean-Logic And



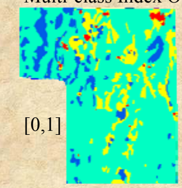
0&1

Boolean-Logic Or



0&1

Multi-class Index Overlay

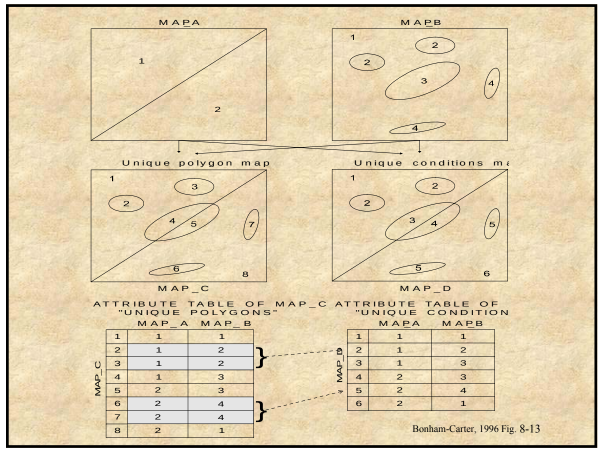


[0,1]

Tools for Map Analysis

Multiple Maps

Boolean Logic Index Overlay Fuzzy Logic	**Knowledge Driven**
Weights of Evidence (Part 1) Logistic Regression Neural Networks	**Data Driven**



Counts to Probability

	A	$\overline{A}$			A	$\overline{A}$	
B	345 (T_{11})	382 (T_{12})	727 $(T_{1\bullet})$	B	0.117 (p_{11})	0.130 (p_{12})	0.247 $(p_{1\bullet})$
$\overline{B}$	141 (T_{21})	2077 (T_{22})	2218 $(T_{2\bullet})$	$\overline{B}$	0.049 (p_{21})	0.705 (p_{22})	0.753 $(p_{2\bullet})$
	486 $(T_{\bullet 1})$	2459 $(T_{\bullet 2})$	2945 $(T_{\bullet\bullet})$		0.165 $(p_{\bullet 1})$	0.835 $(p_{\bullet 2})$	$(p_{\bullet\bullet})$

Area Tabulation $\longrightarrow$ **Probability Tabulation**

Cross-Tabulation Table

	A	$\overline{A}$	
B	345 (T_{11})	382 (T_{12})	727 $(T_{1\bullet})$
$\overline{B}$	141 (T_{21})	2077 (T_{22})	2218 $(T_{2\bullet})$
	486 $(T_{\bullet 1})$	2459 $(T_{\bullet 2})$	2945 $(T_{\bullet\bullet})$

Area Tabulation

$\text{Area}(A \cap B) = T_{11} = 345$

$\text{Area}(A \cap \overline{B}) = T_{21} = 141$

$\text{Area}(\overline{A} \cap B) = T_{12} = 382$

$\text{Area}(\overline{A} \cap \overline{B}) = T_{22} = 2077$

Version 1, January 2000

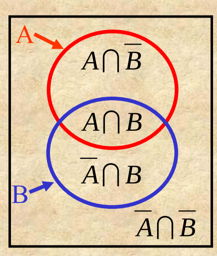
Venn Diagram

$\text{Area}(A \cap B) = T_{11} = 345$

$\text{Area}(A \cap \overline{B}) = T_{21} = 141$

$\text{Area}(\overline{A} \cap B) = T_{12} = 382$

$\text{Area}(\overline{A} \cap \overline{B}) = T_{22} = 2077$



Version 1, January 2000

Probability Table

	A	$\overline{A}$	
B	0.117 (p_{11})	0.130 (p_{12})	0.247 $(p_{1\bullet})$
$\overline{B}$	0.049 (p_{21})	0.705 (p_{22})	0.753 $(p_{2\bullet})$
	0.165 $(p_{\bullet 1})$	0.835 $(p_{\bullet 2})$	$(p_{\bullet\bullet})$

Proportional-Area Tabulation

$$p_{ij} = \frac{T_{ij}}{T_{\bullet\bullet}}$$

$$P\{A\} = p_{\bullet 1}$$

$$P\{B\} = p_{1\bullet}$$

$$P\{A \cap B\} = p_{11}$$

$$P\{A \cap \overline{B}\} = p_{21}$$

$$P\{\overline{A} \cap B\} = p_{21}$$

$$P\{\overline{A} \cap \overline{B}\} = p_{22}$$

Version 1, January 2000

Conditional Probability

$$P\{B \mid A\} = \frac{P\{B \cap A\}}{P\{A\}} = \frac{p_{11}}{p_{\bullet 1}} = \frac{T_{11}}{T_{\bullet 1}}$$

$$P\{\text{GraniteTil l} \mid \text{Granite}\} = \frac{345}{486} = 0.7098$$

$$P\{\text{GraniteTil l}\} = p_{1\bullet} = \frac{T_{1\bullet}}{T_{\bullet\bullet}} = 0.247$$

If Granite is present, then the probability of Granite Till also being present is 0.7098

Probability and Odds

P = probability

O = odds

$$O = \frac{P}{1-P}$$

$$P = \frac{O}{1+O}$$

P	O	lnO
0.0	0	-∞
.1	1/9	-2.20
.2	1/4	-1.39
.4	2/3	-0.41
.5	1/1	0.00
.6	3/2	0.41
.8	4/1	1.39
.9	9/1	2.20
1.0	∞	∞

Conditional Odds

$$O\{B\} = \frac{P\{B\}}{1-P\{B\}} = \frac{T_{1\bullet}/T_{\bullet\bullet}}{1-\frac{T_{1\bullet}}{T_{\bullet\bullet}}} = \frac{T_{1\bullet}}{T_{\bullet\bullet}-T_{1\bullet}}$$

$$O\{B \mid A\} = \frac{P\{B \mid A\}}{1-P\{B \mid A\}} = \frac{P\{B \mid A\}}{P\{\overline{B} \mid A\}}$$

$$O\{B \mid A\} = \frac{p_{11}/p_{\bullet 1}}{p_{21}/p_{\bullet 1}} = \frac{p_{11}}{p_{21}} = \frac{T_{11}}{T_{21}}$$

Conditional Odds

Odds of B given A does occur

$$O\{B \mid A\} = \frac{p_{11}/p_{\bullet 1}}{p_{21}/p_{\bullet 1}} = \frac{p_{11}}{p_{21}} = \frac{T_{11}}{T_{21}}$$

Odds of B given A does not occur

$$O\{B \mid \overline{A}\} = \frac{p_{12}}{p_{22}} = \frac{T_{12}}{T_{22}}$$

Example - Conditional Odds

$$O\{GraniteTill\} = \frac{727}{2945-727} = 0.328$$

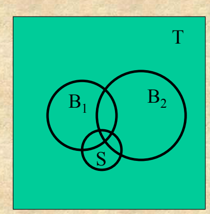
or 3 to 10

$$O\{GraniteTill \mid Granite\} = \frac{345}{141} = 2.45$$

or 25 to 10

If Granite is present, then the odds of Granite Till also being present is 25 to 10

Numerical Example



Venn diagram of point and grid intersections, not draw to scale.

Fig. 9-9

N{T} = 10,000	W^{+}_{1} = 0.9474
N{B_1} = 3600	W^{-}_{1} = -1.8734
N{B_2} = 5000	W^{+}_{2} = 0.3447
N{S} = 200	W^{-}_{2} = -0.5189
N{$B_1 \cap S$} = 180	C_1 = 2.8208
N{$B_2 \cap S$} = 140	C_2 = 0.8636

$C_1 > C_2$:Therefore Pattern B_1 is a better predictor that Pattern B_2!

Weights

- Define the area to be studied
  - Count its area in unit cells = N{Study Area}
- Count the number of training sites in the study area = N{Training Sites}= N{S}
- Count the area of the pattern B = N{B}
- Prior probability = P{S} = N{Training Sites}/N{Study Area}
- Conditional Probability: Posterior Probability of a training site given the presence of a binary pattern B and absence of B.

$$P\{S|B\} = \frac{P\{S \cap B\}}{P\{B\}} = \frac{N\{S \cap B\}}{N\{B\}} = P\{S\} * \frac{P\{B|S\}}{P\{B\}}$$

$$P\{S|\bar{B}\} = P\{S\} * \frac{P\{\bar{B}|S\}}{P\{\bar{B}\}}$$

Odds Formulation

$$P\{S|B\} = \frac{P\{S \cap B\}}{P\{B\}} = \frac{N\{S \cap B\}}{N\{B\}} = P\{S\} * \frac{P\{B|S\}}{P\{B\}}$$

$$O\{S|B\} = O\{S\} \frac{P\{B|S\}}{P\{B|\bar{S}\}}$$

$$\ln O\{S|B\} = \ln O\{S\} + \ln\left\{\frac{P\{B|S\}}{P\{B|\bar{S}\}}\right\}$$

$$\operatorname{logit}\{S|B\} = \operatorname{logit}\{S\} + \operatorname{logit}\left\{\frac{P\{B|S\}}{P\{B|\bar{S}\}}\right\} = \operatorname{logit}\{S\} + W^{+}$$

$$P\{S|\bar{B}\} = P\{S\} * \frac{P\{\bar{B}|S\}}{P\{\bar{B}\}}$$

$$\operatorname{logit}\{S|\bar{B}\} = \operatorname{logit}\{S\} + \operatorname{logit}\left\{\frac{P\{\bar{B}|S\}}{P\{\bar{B}|\bar{S}\}}\right\} = \operatorname{logit}\{S\} + W^{-}$$

Cross-Tabulation Table

	A	$\bar{A}$	
B	345 (T_{11})	382 (T_{12})	727 ($T_{1\bullet}$)
$\bar{B}$	141 (T_{21})	2077 (T_{22})	2218 ($T_{2\bullet}$)
	486 ($T_{\bullet 1}$)	2459 ($T_{\bullet 2}$)	2945 ($T_{\bullet\bullet}$)

Area $(A \cap B) = T_{11} = 345$

Area$(A \cap \bar{B}) = T_{21} = 141$

Area$(\bar{A} \cap B) = T_{12} = 382$

Area$(\bar{A} \cap \bar{B}) = T_{22} = 2077$

Area Tabulation

Version 1, January 2000

Weights Calculation Formula

$$W^{+} = \ln\left[\frac{T_{11} * T_{2\bullet}}{T_{21} * T_{1\bullet}}\right] \quad \text{eq. 8-20}$$

$$W^{-} = \ln\left[\frac{T_{12} * T_{2\bullet}}{T_{22} * T_{1\bullet}}\right] \quad \text{eq. 8-21}$$

Binary Patterns!

Bayes' Theorem

P{Rain|Time-of-Year} = P{Rain} * Time-of-Year Factor

P{Rain|Evidence} = P{Rain} * Evidence1 * Evidence 2 etc.

P{Rain} = Prior Probability, the probability before considering the evidence

P{Rain|Evidence} = Posterior Probability, the probability after considering the evidence.

- The evidence can increase or decrease the prior probability
- Applied to maps, the evidence is a pattern!

Bayes' Theorem and Training Sites

- Used here to predict the presence or absence of a set of point objects.
  - Points objects used include mineral deposits, animal habitat, human disease, etc.
  - Points represent a small unit of area, the unit cell, relative to the area studied and the resolution of the evidence.
  - Points are the training sites.
  - Assumes one training site per unit cell.
- Assumes conditional independence of evidence with regards to training sites.

Logit Form of Baye's Theorem

- This allows for summation of the weights for all patterns as opposed to products
- W^+ is weight for inside the pattern, B
- W^- is weight for outside the pattern, not B
- Positive W^+ and negative W^- indicates a positive correlation between training sites and the pattern
- Contrast = W^+ - W^-
  - Relative measure of correlation - larger the contrast the greater the correlation
  - Can use contrast to help define best pattern!

Multiple Patterns = Multiple Weights

- Objective is to combine all the evidence to obtain a combined posterior probability.
  - Use Bayes' Theorem to combine patterns
  - Assumes conditional independence of patterns with regards to the training sites.

Conditional independence implies

$$P\{B_1 \cap B_2 \mid S\} = P\{B_1 \mid S\} * P\{B_2 \mid S\}$$

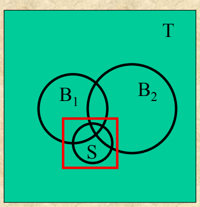
This allows

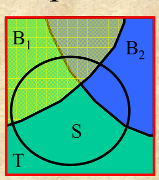
$$P\{S \mid B_1 \cap B_2\} = P\{S\} * \frac{P\{B_1 \mid S\}}{P\{B_1\}} * \frac{P\{B_2 \mid S\}}{P\{B_2\}}$$

or

$$\operatorname{logit}\{S \mid B_1 \cap B_2\} = \operatorname{logit}\{S\} + W_1^+ + W_2^+$$

Conditional Independence





Conditional Independence is satisfied if :

$$\frac{N\{B_1 \cap S\}}{N\{S\}} * \frac{N\{B_2 \cap S\}}{N\{S\}} = \frac{N\{B_1 \cap B_2 \cap S\}}{N\{S\}}$$

Using numbers from Fig 9-9

$$\frac{180}{200} * \frac{140}{200} = 0.63 \Rightarrow \text{for CI } \frac{N\{B_1 \cap B_2 \cap S\}}{N\{S\}} = \frac{\boxed{126}}{200} = 0.63$$

What if there were three patterns?

Old Overall Test for Conditional Independence

$$N\{S_{Calc}\} = \sum_{k=1}^{m} P_k * (\text{unit cell})$$

where m = total number of unit cells.

$$\text{CI Ratio} = \frac{N\{S\}}{N\{S_{Calc}\}}$$

- **Unit cell is a constant in the grid implementation of Weights of Evidence.**
- **CI Ratio is typically less than 1.**
- **If CI Ratio is less than .90 to .85 then a serious CI problem has occurred. Now considered too conservative.**
- **Replaced by Agterberg-Cheng CI Test**

Agterberg-Cheng CI test

Part of the WofE table

	N	P	S(P)	N*P	N2*S(P)2
					Variance
Area_sqm	Area (KM2)	Post_Prob	Uncertainty	T	T
3.27E+09	3272.75	0.002132	0.000712	6.978845	5.435925
3.17E+10	31687.19	0.000239	8E-05	7.58116	6.432526
1.03E+10	10271.94	1.31E-05	1.38E-05	0.134254	0.020036
1.2E+09	1204.625	0.000117	0.000123	0.14058	0.021975
2.29E+08	228.875	0.000511	0.000541	0.116973	0.015316
6.84E+08	683.875	0.009275	0.003188	6.343248	4.754615
8.03E+09	8026.25	0.000521	0.000197	4.181596	2.504673
1.13E+09	1130.625	0.004633	0.001747	5.2384	3.901065
2.69E+08	269.0625	0.019985	0.007631	5.377319	4.215386
73375000	73.375	0.000614	0.000104	0.045076	5.8E-05
1937500	1.9375	3.36E-05	3.41E-05	6.5E-05	4.35E-09
				36.13752	27.30158

Test Statistic for confidence that the predicted number is greater than the expected number (35).

$$\frac{T - 35}{Std(T)} = 0.218$$

So less than 0.253 (60% confidence); therefore **Accept CI**

See CI_Agterberg.pdf in Documentation or Reading folders for details.

Testing for Sources of Conditional Dependency

- Pair-wise Chi-squared test
  - A weak test of pairs only.
  - Not implemented in ArcSDM 3.
- Replaced by multiple Agterberg-Cheng tests.
  - Make models of pairs, triplets, etc. of evidence layers and get the Agterberg-Cheng results.
  - Identify combinations causing the problem. Note may be a triplet or larger combination.

Solutions to CI Problems

- Combine group of evidence causing the CI problem in some logical fashion or delete one evidence and recalculate the model.
- If still have CI problem, must consider the WofE Posterior Probability distorted.
  - Treat the "posterior probability" as favorability, an ordinal measurement-scale number.
  - Call it favorability even though the software labels it posterior probability.
  - Define ranks.
- Use Logistic Regression Posterior Probability.

Variance of Weights and Contrast

$$s^2(W^+) = \frac{1}{N\{B \cap S\}} + \frac{1}{N\{B \cap \overline{S}\}}$$

$$s^2(W^-) = \frac{1}{N\{\overline{B} \cap S\}} + \frac{1}{N\{\overline{B} \cap \overline{S}\}}$$

$$s^2(\text{Contrast}) = s^2(W^+) + s^2(W^-)$$

Version 1, January 2000

Total Variance of Posterior Probability

$$s^2(P_{Posterior}) = \left[\frac{1}{N(S)} + \sum_{i=1}^{n} s^2(W_j^k)\right] * P_{Posterior}^2$$

where

k is + and - and

n is the number of patterns

$$s_i^2(\text{missing}) = \{P(S \mid B_i) - P(S)\}^2 * P(B_i) + \{P(S \mid B_i) - P(S)\}^2 * P(\overline{B}_i)$$

where i is a pattern with missing data

$$s^2(\text{total}) = s^2(P_{Posterior}) + \sum_{i=1}^{m} s_i^2(\text{missing})$$

where m is the number of layers with missing data.

Version 1, January 2000

Revised Variance of Missing Data

$$s_i^2(\text{Missing}) = \sum_{j=1}^{m_i} [(P_j^* - P)^2 \frac{a_{ij}}{a\_\text{data}_i}]$$

where

i = a layer with missing data

j = one of m_i classes in layer i.

P_j^* = updated posterior probability by the weight for class j of a cell with missing data

P = the posterior probabilty of a cell with missing data

a_{ij} = the area of class j in layer i

a_data_i = the total area of data in layer i,
that is total study area - area of missing data in layer i

Note this is a cell based calculation, which is applied to cells with missing data!

Revised Sept. 2006

Studentized Value

- Studentized Contrast = Contrast/s(C)
- Studentized Posterior Probability = Post. Prob./s(total Post. Prob.)
- An informal test of the hypothesis that value tested is zero. If Studentized value greater than 2 then can assume that the value tested is not equal to zero with approximately 98% confidence.
- Use in a relative sense and to structure decision making.

Version 1, January 2000

Student T Values

Confidence	Test Value
99.5%	2.576
99%	2.326
97.5%	1.96
95%	1.645
90%	1.282
80%	0.842
70%	0.542
60%	0.253

Because Studentized test applied here is only approximate, use these values as a guide. If you can accept more risk, then you can use lower confidence values!

Version 1, January 2000

Decisions for Weights of Evidence

- Define the study area
- Define the training set
- Select confidence level for contrast
- Select the evidential maps
  - Use Contrast and Studentized Contrast to evaluate.
  - Binary Reclassification
  - Thresholds maximum, minimum, or grouping of nominal classes
- These decisions define objective, binary reclassification
  - Needed measurements: Area of study, Area of the pattern, Number of training sites, Number of training sites inside the pattern

Weights of Evidence

- Advantages
  - Objective assignment of weights, which reflect the importance of the class and the layer.
  - Multiple patterns combined simply
  - Binary reclassification to optimize contrast gives insights into spatial relationships
  - Deals with missing data
  - Measures aspects of uncertainty that can be mapped
- Disadvantages
  - Assumption of conditional independence
  - Requires a training set of sufficient size.

Tools for Map Analysis
Multiple Maps

Boolean Logic
Index Overlay
Fuzzy Logic
} **Knowledge Driven**

Weights of Evidence (Part 2)
Logistic Regression
Neural Networks
} **Data Driven**

Weights-of-Evidence Method

- Originally developed as a medical diagnosis system
  - relationships between symptoms and disease evaluated from a large patient database
  - each symptom either present/absent
  - weight for present/weight for absent (W+/W-)
- Apply weighting scheme to new patient
  - add the weights together to get result

Weights of Evidence

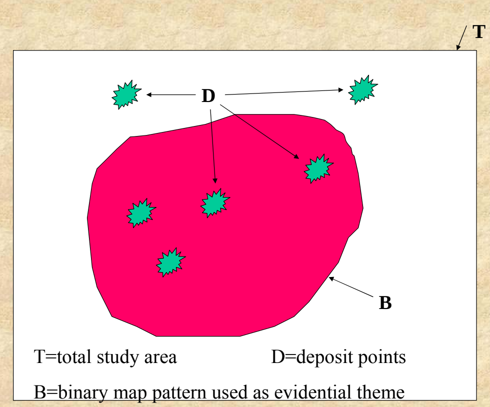
- Data driven technique
  - Requires training sites
- Statistical calculations are used to derive the weights based upon training sites.
- Evidence (maps) are generally reclassified into binary patterns.

Weights-of-Evidence Terms

- Weights for patterns
  - W+ - weight for inside the pattern
  - W- - Weight for outside the pattern
  - 0 - Weights for areas of no data
- Contrast - a measure of the spatial association of pattern with sites
- Studentized Contrast - a measure of the significance of the contrast

Weights of Evidence

- Binary maps to define favorable areas
  - Can use multi-layer patterns
- Measurements
  - Area of study
  - Area of Pattern
  - Number of training sites
  - Number of training sites inside the pattern



Bonham-Carter, personal comm. 2002

Preprocessing
Nominal Measurement Scale

- For example - Geological map
  - select particular stratigraphic units or class
  - generalize by reclassification
  - extract and buffer boundaries between units

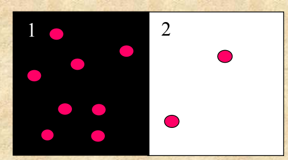
Preprocessing
Continuous Measurement Scale

- Histogram transformations
- Physical properties processing
- Filter
  - separate anomaly/background
- Spatial interpolation (e.g. surfaces, krige)
- Logical combinations (merging, boolean, fuzzy logic)
- Summarize by zonal statistics
  - separate anomaly/background
  - define a residual
  - multivariate analysis
    - principal components analysis and others

Overlay combination

- In vector
  - create polygon overlay and associated PAT
  - create unique conditions overlay and associated PAT
  - Topological selections
- In raster
  - superimpose grids

Application to Binary Evidence

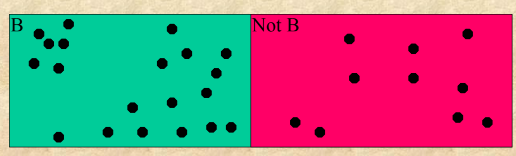


Class	Area	#sites	Relative density	Weight
1	50	8	0.8/0.5=1.6	ln(1.6)= + 0.47
2	50	2	0.2/0.5=0.4	ln(0.4)= - 0.92
Total	100	10		

Expected Values of Weights

- If sites occur randomly,
  - Relative density (RD)=1.0
  - Weight (W) = ln(RD) =0.0
- If sites occur more frequently than chance
  - RD > 1.0, W is positive
- If sites occur less frequently than chance
  - RD < 1.0, W is negative

Example – More Points Than Chance



N(T) = 1000 unit cells (area of study region)

N(B) = 500 unit cells (area of theme B present)

N(B&D) = 20 (count of number of training points on B)

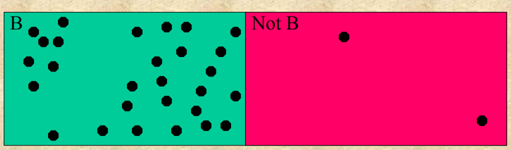
N(D) = 30 (count of total number of training points)

$W^+ = 0.2980$ $W^- = -0.4157$ $C = 0.7138$

More points on theme than would be expected due to chance

Bonham-Carter, personal comm. 2002

Example – Many More Points



N(T) = 1000 unit cells (area of study region)

N(B) = 500 unit cells (area of theme B present)

N(B&D) = 28 (count of number of training points on B)

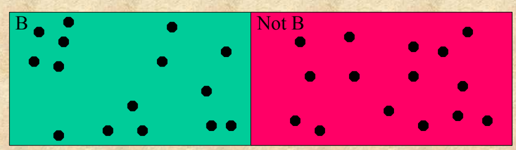
N(D) = 30 (count of total number of training points)

$W^+ = 0.6513$ $W^- = -2.0414$ $C = 2.6927$

Many more points on theme than would be expected due to chance

Bonham-Carter, personal comm. 2002

Example – Equal Pattern and Points



N(T) = 1000 unit cells (area of study region)

N(B) = 500 unit cells (area of theme B present)

N(B&D) = 15 (count of number of training points on B)

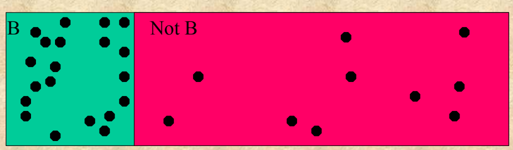
N(D) = 30 (count of total number of training points)

$W^+ = 0.0$ $W^- = -0.0$ $C = 0.0$

Number of points on theme equals that expected due to chance

Bonham-Carter, personal comm. 2002

Example – Small Pattern and Many Points



N(T) = 1000 unit cells (area of study region)

N(B) = 250 unit cells (area of theme B present)

N(B&D) = 20 (count of number of training points on B)

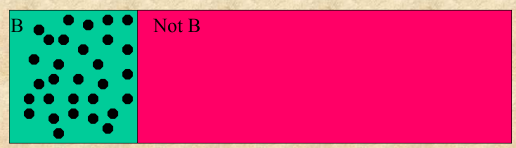
N(D) = 30 (count of total number of training points)

$W^+ = 1.0338$ $W^- = -0.8280$ $C = 1.8617$

Many more points on theme than would be expected due to chance

Bonham-Carter, personal comm. 2002

Example - Weights Undefined



N(T) = 1000 unit cells (area of study region)

N(B) = 250 unit cells (area of theme B present)

N(B&D) = 30 (count of number of training points on B)

N(D) = 30 (count of total number of training points)

$W^+ = \text{inf}$ $W^- = -\text{inf}$ $C = \text{inf}$

Undefined: practical solution--assign fraction of point to (not B)

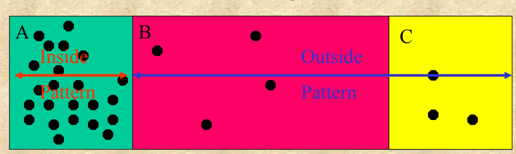
Bonham-Carter, personal comm. 2002

Multi-class Themes

- Maps (themes) with unordered classes (categorical) e.g. geological map. Calculate weights for each class and then group classes (reclassify) as needed.
- Maps (themes) with ordered classes (contour maps e.g. geochemical or geophysical field variables). Usually calculate weights based on successive contour levels, cumulatively. Then reclassify.

Bonham-Carter, personal comm. 2002

Multi-class – Categorical Classes



N(T) = 1000 unit cells (area of study region)

N(A) = 250 , N(B) = 500, N(C) = 250,

N(A&D) = 23, N(B&D) = 4, N(C&D) = 3,

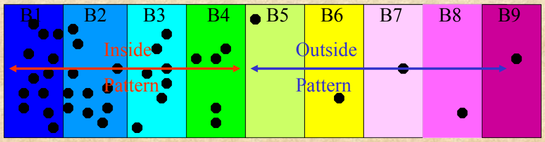
N(D) = 30 (count of total number of training points)

$W1 = 1.1866$ $W2 = -1.3442$ $W3 = -0.9347$ $C_{max} = 2.5308$

Three classes, e.g. rock types (categorical scale of measurement)

Bonham-Carter, personal comm. 2002

Ordered Classes - Cumulative

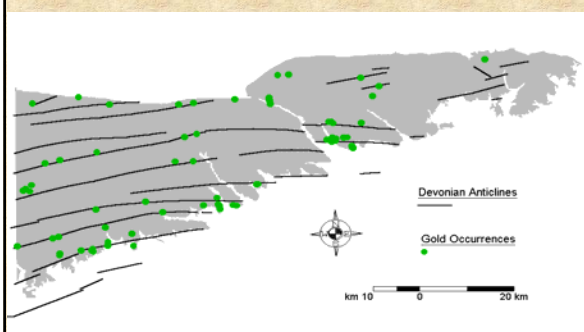


$N(B_i)$	100	100	100	100	100	100	100	100	100
Cum	100	200	300	400	500	600	700	800	900
N(D)	12	11	7	5	1	1	1	1	1
Cum	12	23	30	35	36	37	38	39	40
W+	1.08	1.03	0.87	0.72	0.51	0.35	0.21	0.10	--
W-	-0.25	-0.63	-1.01	-1.53	-1.53	-1.53	-1.53	-1.53	--
C	1.33	1.66	1.88	2.25	2.04	1.88	1.74	1.64	--

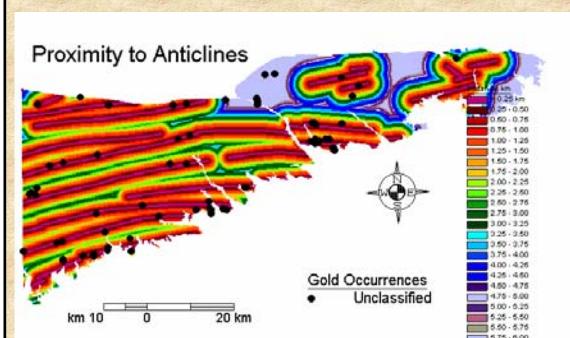
Bonham-Carter, personal comm. 2002

Weights Calculations

- Choose a small unit cell – affects the prior probability but only a little on the weights
- Can have multi-class maps but often not enough training points to get stable weights.
  - Use Studentized contrast to evaluate stability of weights.
- Contrast can be used to define optimal thresholds.
  - Use Studentized contrast to evaluate stability of contrast.

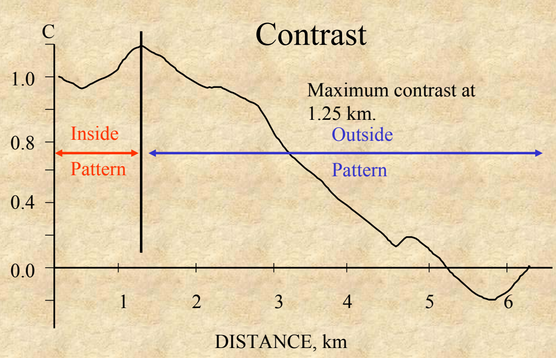


Bonham-Carter, 1996

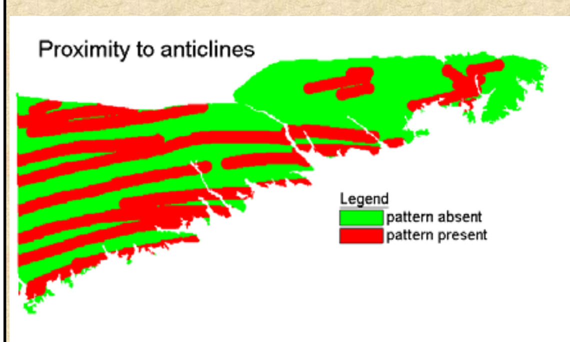


Bonham-Carter, 1996

Contrast

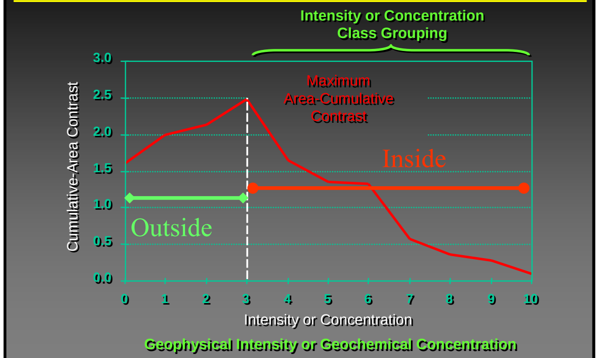


Bonham-Carter, 1996



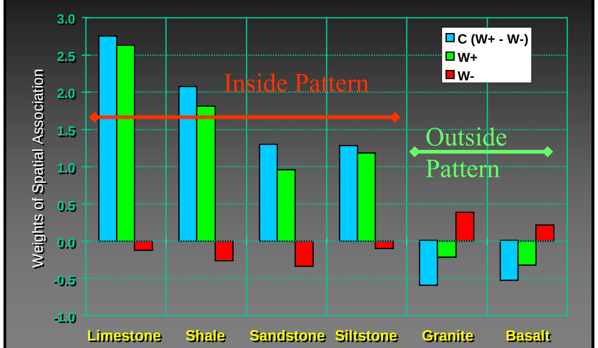
Bonham-Carter, 1996

Cumulative-descending Reclassification



Modified from Mihalasky, 1999

Categorical-Weighting Reclassification



Modified from Mihalasky, 1999

Handling Uncertainty

- Uncertainty due to weights – variance of weights.
- Uncertainty due to missing data – estimate of variance due to missing data
- Other measures of uncertainty?
- For Response Map can combine the various uncertainty measures to obtain a total variance.
- Studentized posterior probability (PP/s(PP)) can provide a useful measure of confidence.

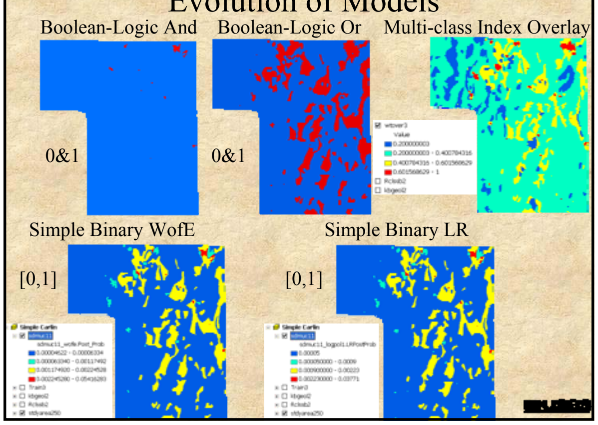
Decisions for Weights of Evidence

- Define unit area for counting area (Unit Cell)
- Define the study area
- Define the training set
- Select confidence level for contrast
- Select the evidential maps
  - Use Contrast and Studentized Contrast to evaluate.
  - Reclassification (Binary or Multi-class)
  - Thresholds maximum, minimum, or grouping of nominal classes
- These decisions define objective, binary reclassification
  - Needed measurements: Area of study, Area of the pattern, Number of training sites, Number of training sites inside the pattern

Weights of Evidence

- Advantages
  - Objective assignment of weights
  - Multiple patterns combined simply
  - Reclassification to optimize contrast gives insights into spatial relationships
  - Deals with missing data
  - Measures aspects of uncertainty that can be mapped
- Disadvantages
  - Assumption of conditional independence
  - Requires a training set of sufficient size.

Evolution of Models



Multi-Class Generalization

- Boolean Logic
- Index Overlay
- Fuzzy Logic

Knowledge Driven

- Weights of Evidence (Part 3)
- Logistic Regression
- Neural Networks

Data Driven

Student T Values

Confidence	Test Value
99.5%	2.576
99%	2.326
97.5%	1.96
95%	1.645
90%	1.282
80%	0.842
70%	0.542
60%	0.253

Because Studentized test applied here is only approximate, use these values as a guide. If you can accept more risk, then you can use lower confidence values!

Version 1, January 2000

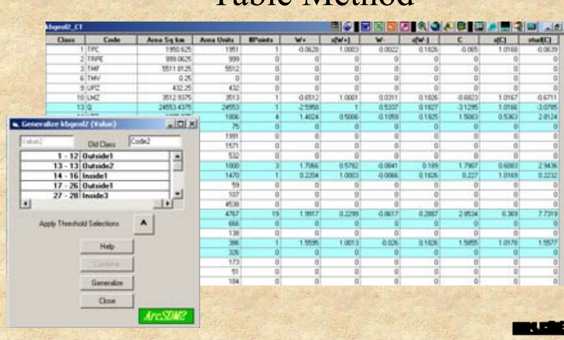
Categorical Weights Data

Class	Code	Area Sq km	Area Units	#Points	W+	s(W+)	W-	s(W-)	C	s(C)	stud(C)
38	LPZE	4766.9375	4767	19	1.9917	0.2299	-0.8617	0.2887	2.8534	0.3690	7.7319
27	C	999.6250	1000	3	1.7066	0.5782	-0.0841	0.1890	1.7907	0.6083	2.9436
45	UPZC	385.9375	386	1	1.5595	1.0013	-0.0260	0.1826	1.5855	1.0178	1.5577
14	LPZ	1805.8750	1806	4	1.4024	0.5006	-0.1059	0.1925	1.5083	0.5363	2.8124
29	UPZE	1469.8750	1470	1	0.2204	1.0003	-0.0066	0.1826	0.2270	1.0169	0.2232
1	TPC	1950.6250	1951	1	-0.0628	1.0003	0.0022	0.1826	-0.0650	1.0168	-0.0639
10	LMZ	3512.9375	3513	1	-0.6512	1.0001	0.0311	0.1826	-0.6823	1.0167	-0.6711
13	Q	24553.4375	24553	1	-2.5958	1.0000	0.5337	0.1827	-3.1295	1.0166	-3.0785
2	TRPE	999.0625	999	0							
3	TMF	5511.8125	5512	0							
6	TMV	0.2500	0	0							
9	UPZ	432.2500	432	0							
17	QV	74.8125	75	0							
18	TPV	1991.2500	1991	0							
20	TPF	1570.8750	1571	0							
25	KG	531.8125	532	0							
32	P	58.9375	59	0							
35	TI	106.8125	107	0							
36	LTV	4537.7500	4538	0							
39	JG	665.5625	666	0							
43	KC	138.5000	138	0							
47	LMZV	325.5625	326	0							
48	TRG	173.3750	173	0							
49	KG2	50.9375	51	0							
50	JMI	184.2500	184	0							

Categorical Data (Zoomed)

Class	Code	#Points	C	s(C)	stud(C)
38	LPZE	19	2.8534	0.3690	7.7319
27	C	3	1.7907	0.6083	2.9436
45	UPZC	1	1.5855	1.0178	1.5577
14	LPZ	4	1.5083	0.5363	2.8124
29	UPZE	1	0.2270	1.0169	0.2232
1	TPC	1	-0.0650	1.0168	-0.0639
10	LMZ	1	-0.6823	1.0167	-0.6711
13	Q	1	-3.1295	1.0166	-3.0785
2	TRPE	0			

Multi-Class Generalization Table Method



Editing and Checking

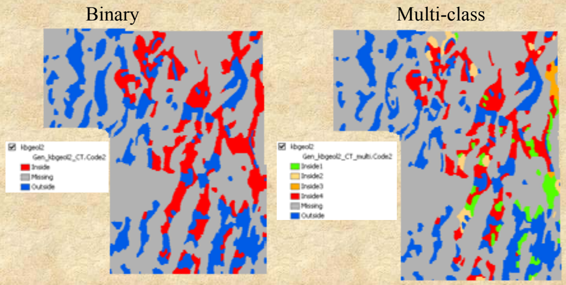
S_value3	Area (km²)	Number Points	W+	Std W+	W-	Std W-	Contrast	Std Contrast	Confidence
Outside	11717.75	1	-2.9076	1	0.7788	0.1771	-3.6864	1.0156	-3.6298
Moderate	4275.375	8	0.1818	0.3539	-0.0518	0.2001	0.2337	0.4066	0.5747
High	385.9375	2	1.2038	0.7089	-0.0442	0.1797	1.248	0.7314	1.7063
VHigh	4766.938	22	1.0873	0.2137	-0.8441	0.3016	1.9314	0.3696	5.2251
Younger	35840.44	2							

Class	S_value3	Area_Sq_km	Area_Units	No_Points	Wplus	s_Wplus	Wminus	s_Wminus	Contrast	s_Contrast	stud_Cnt
3	Inside1	999.625	999.625	3	1.6752	0.5782	-0.0807	0.1857	1.7559	0.6073	2.8913
6	Inside2	1805.875	1805.875	4	1.3707	0.5006	-0.1013	0.189	1.472	0.5351	2.751
4	Inside3	4766.9375	4766.9375	19	1.9599	0.2299	-0.8134	0.2774	2.7734	0.3603	7.6984
8	Inside4	385.9375	385.9375	2	2.2236	0.7089	-0.0578	0.1826	2.2814	0.7321	3.1163
2	Outside1	24287.25	24287.25	3	-1.518	0.5774	0.4598	0.1858	-1.9778	0.6065	-3.2608
1	Outside3	24553.4375	24553.4375	1	-2.6276	1	0.5348	0.1797	-3.1624	1.016	-3.1125
-99	Missing Data	187.375	187.375	0							

Categorical-Weights Generalization

Binary

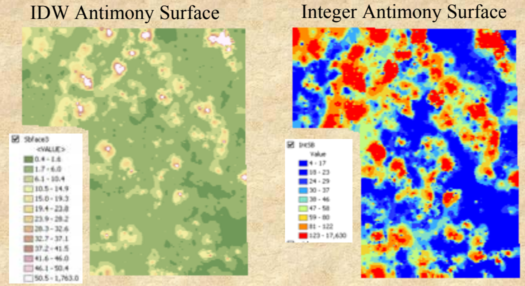
Multi-class



Ordered-Data Preprocessing

IDW Antimony Surface

Integer Antimony Surface

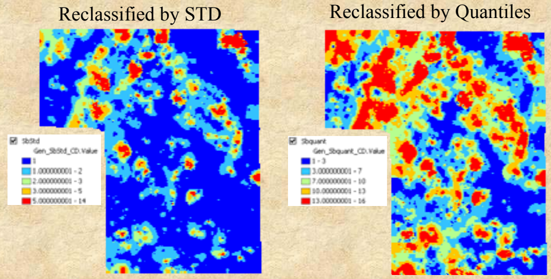


Int(([Sb] * 10) + 0.5))

Ordered-Data Reclassification from Integer Source

Reclassified by STD

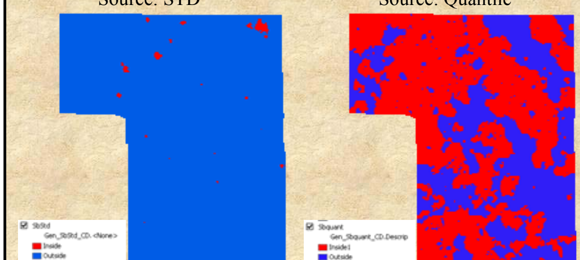
Reclassified by Quantiles



WofE Binary Generalization from Integer Source

Source: STD

Source: Quantile



Ratio Weights Data Cumulative Descending

Class	Area Sq km	Area Units	#Points	W+	s(W+)	W-	s(W-)	C	s(C)	stud(C)
14	416.0000	416	4	2.8777	0.5024	-0.1309	0.1925	3.0085	0.5380	5.5917
13	469.0000	469	4	2.7566	0.5021	-0.1299	0.1925	2.8866	0.5378	5.3676
12	526.0000	526	5	2.8661	0.4494	-0.1667	0.1962	3.0327	0.4903	6.1854
11	618.0000	618	6	2.8874	0.4102	-0.2043	0.2000	3.0917	0.4564	6.7738
10	730.0000	730	7	2.8749	0.3798	-0.2431	0.2042	3.1180	0.4312	7.2312
9	868.0000	868	9	2.9538	0.3351	-0.3277	0.2132	3.2815	0.3972	8.2622
8	1137.5000	1138	9	2.6805	0.3347	-0.3228	0.2132	3.0034	0.3968	7.5685
7	1567.5000	1568	11	2.5598	0.3026	-0.4104	0.2236	2.9702	0.3763	7.8941
6	2310.5000	2310	11	2.1701	0.3022	-0.3969	0.2236	2.5670	0.3760	6.8273
5	3746.0000	3746	15	1.9960	0.2587	-0.5934	0.2500	2.5894	0.3598	7.1968
4	7282.5000	7282	21	1.6666	0.2185	-0.9945	0.3163	2.6611	0.3844	6.9225
3	16892.3125	16892	27	1.0752	0.1926	-1.6950	0.5000	2.7703	0.5358	5.1700
2	50663.0625	50663	31							
1	56779.0625	56779	31							
-99	0.5000	0	0							0.0000

Ratio Data – Categorical Weights Table Method

Class	Area Sq km	Area Units	#Points	W+	s(W+)	W-	s(W-)	C	s(C)	stud(C)
-99	0.5000	0	0							0.0000
1	6116.0000	6116	0							
2	33770.7500	33771	4	-1.5911	0.5000	0.7748	0.1858	-2.3659	0.5334	-4.4352
3	9609.8125	9610	6	0.0717	0.4084	-0.0152	0.1925	0.0869	0.4515	0.1925
4	3536.5000	3536	6	1.0725	0.4086	-0.1364	0.1925	1.2090	0.4517	2.6767
5	1435.5000	1436	4	1.5693	0.5007	-0.1037	0.1857	1.6730	0.5340	3.1326
6	743.0000	743	0							
7	430.0000	430	3	2.4917	0.5794	-0.0878	0.1826	2.5794	0.6075	4.2461
8	269.5000	270	0							
9	138.0000	138	2	3.2303	0.7123	-0.0601	0.1797	3.2904	0.7346	4.4793
10	112.0000	112	2	3.4425	0.7135	-0.0606	0.1797	3.5031	0.7358	4.7611
11	92.0000	92	1	2.9390	1.0055	-0.0292	0.1768	2.9681	1.0209	2.9073
12	57.0000	57	1	3.4245	1.0089	-0.0298	0.1768	3.4543	1.0243	3.3724
13	53.0000	53	0							
14	416.0000	416	4	2.8151	0.5024	-0.1219	0.1857	2.9370	0.5357	5.4830

Checking Weights from Generalization

	Class	Gen_SbStd	Area (Sq km)	Area (Units)	No. Points	Weight +	s Weight +	Weight -	s Weight -	Contrast	s Contrast	stud Cnt
▸	1	Outside	32273.1875	32273	3	-1.7713	0.5774	0.7396	0.1891	-2.511	0.6076	-4.1329
	2	Inside1	23449.25	23449	19	0.3946	0.2295	-0.4162	0.2887	0.8109	0.3688	2.1985
	3	Inside2	431.5	432	3	2.5491	0.5794	-0.0942	0.189	2.6433	0.6094	4.3374
	4	Inside3	600	600	6	2.9168	0.4103	-0.2046	0.2	3.1214	0.4565	6.8381

Input Sb reclassified by Standard Deviation.

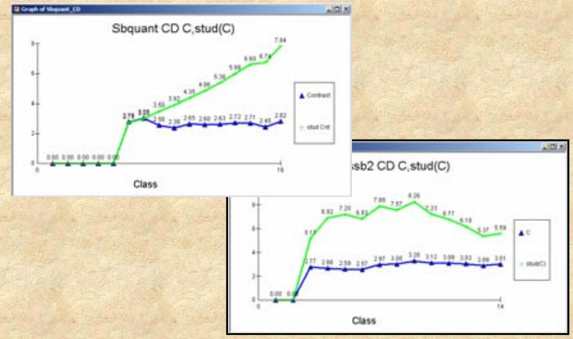
Used in subsequent models.

	Class	Area (Sq k	Area (Units)	No. Points	Weight +	s Weight +	Weight -	s Weight -	Contrast	s Contrast	stud Cnt
▸	-99	0.5	0	0	0	0	0	0	0	0	0
	1	39877.125	39877	5	-1.5038	0.4472	1.044	0.1926	-2.5478	0.487	-5.2321
	2	18146.3125	18146	21	0.8366	0.2184	-0.7334	0.3016	1.5699	0.3723	4.2167
	3	730	730	6	2.6871	0.4099	-0.1948	0.1962	2.8819	0.4545	6.3415

Record: 1 Show: All Selected Records (0 out of 4 Selected) Options

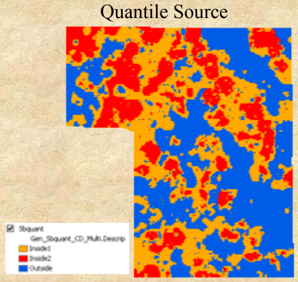
Input Sb reclassified by Natural Breaks with breaks at 10 and 2.

Ratio Data – Chart Method

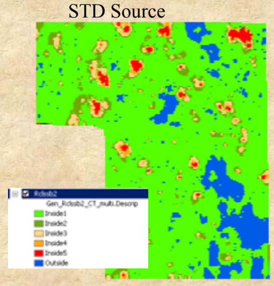


Ratio Data – Multi-Class Generalization

Quantile Source

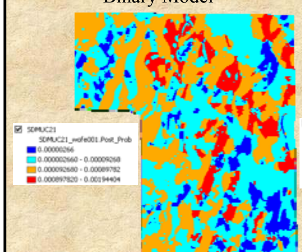


STD Source

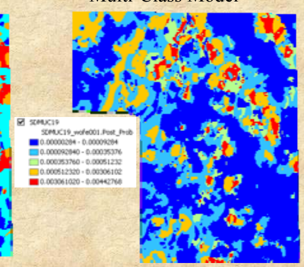


Comparison of Models

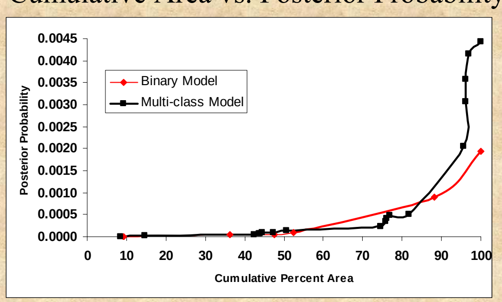
Binary Model



Multi-Class Model

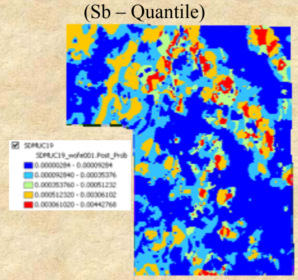


Multi-Class vs. Binary Models Cumulative Area vs. Posterior Probability

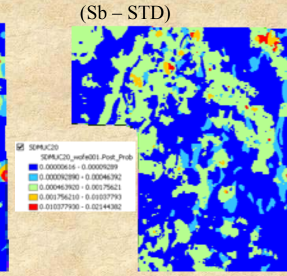


Comparison of Multi-Class Models

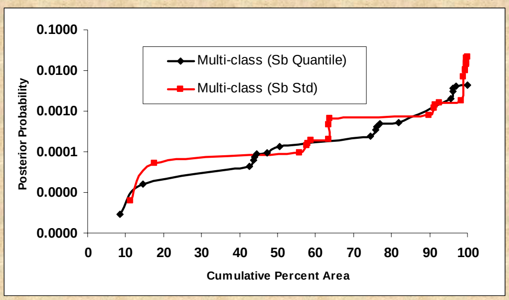
Multi-Class Model
(Sb – Quantile)



Multi-Class Model
(Sb – STD)



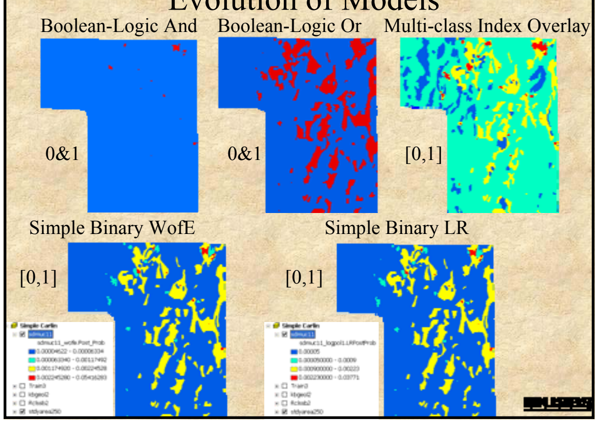
Multi-Class vs. Multi-Class Models Cumulative Area vs. Posterior Probability



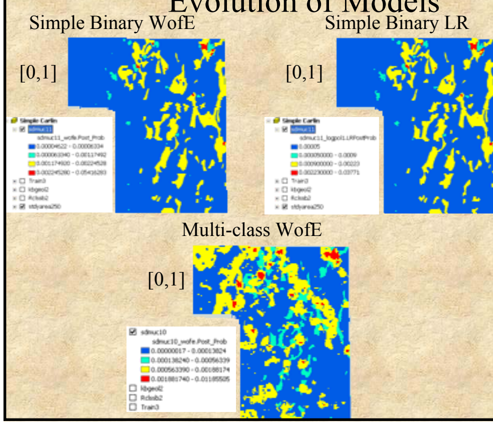
Multi-Class Generalizations

- Many approaches: a complex knowledge driven process guided by the statistics
  - Evaluate Studentized Contrast
  - Pick significant values of contrast and Studentized Contrast
  - Classes should make sense as measured by contrast and logic of process being modeled.
- Test the generalization by calculating weights for the generalization
- Evaluate differences between models

Evolution of Models



Evolution of Models



Logistic Regression Method

Graeme Bonham-Carter

Bonham-Carter, 1999
Modified by Raines 2006

Introduction

- "Data-driven" method applicable where training set of mineral sites is available
- The response variable is dichotomous (binary), e.g. presence/absence of mineral site
- The explanatory variables (evidential themes) are ordered or dichotomous (not multi-class categorical).

Bonham-Carter, 1999

In ordinary regression, the response variable is continuous, unbounded and measured on an interval or ratio scale

In situations where the response variable is binary (present/absent) this causes a problem, because the predicted response must be in the interval [0,1].

The response variable can be assumed to be P(Y=1), from which we also know P(Y=0)=1-P(Y=1)

Bonham-Carter, 1999

The solution to the problem of forcing the response variable to be in the range [0,1] is to use the logit transform.

Logits = natural logs of odds

Odds = Probability/(1-Probability)

$$\text{Logit}(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

Where the b's are unknown coefficients and the X's are the explanatory variables

Bonham-Carter, 1999

Logistic Regression Vs. Weights of Evidence

$$\text{Logit}(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

(simultaneous solution of b's)

$$\text{Logit}(Y) = \text{Prior Logit} + W_1 + W_2 + W_3 + \ldots + W_k$$

(solution for W's theme by theme, not simultaneous)

Note that the b_0 term in LR is comparable to the prior logit in WofE, and the b's are comparable to the W's. However, instead of 1 coefficient, there are 2 (or more) weights, depending on the number of classes. Therefore, the b's are more comparable to the contrast values

Bonham-Carter, 1999

Solution to Logistic Regression Equation

- The coefficients cannot be solved by ordinary least squares (a direct matrix inversion), because the equation is non-linear
- The method of maximum likelihood is used to maximize the value of a log-likelihood function
  - This requires an iterative solution
- So coefficients are obtained simultaneously without an assumption of conditional independence.

Bonham-Carter, 1999

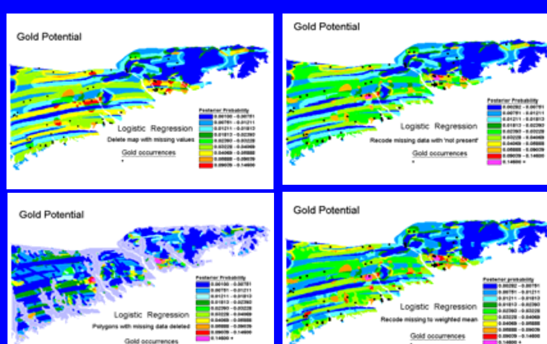
Practicalities

- Can calculate the logistic regression coefficients using the same unique conditions table as for WofE
  - Muti-class themes must be split into binary themes in unique conditions table.
- In ArcSDM deal with missing data and multi-class problem automatically.
- In Arc/Info does not deal with missing data and has another input format.

Problem of Missing Data

- Deleting all unique conditions with missing values in any of the evidential themes.
- Deleting themes that have missing data totally.
- Replacing missing values with zero, or some other constant.
- Replacing missing values with an expected value, e.g. area weighted mean

"Missing Data" Approaches



Used in Arc-SDM

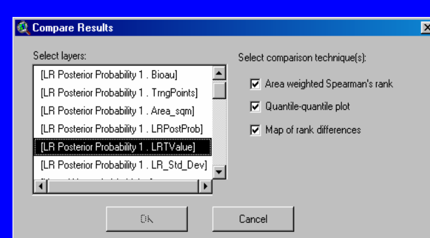
Can then compare the results from weights of evidence to logistic regression

This is then a check on the effect of conditional dependence on the results of weights of evidence, although if missing data and multi-class categorical evidential themes have been used, then one cannot be absolutely sure what effect the recoding in logistic regression has on the results.

Compare Results

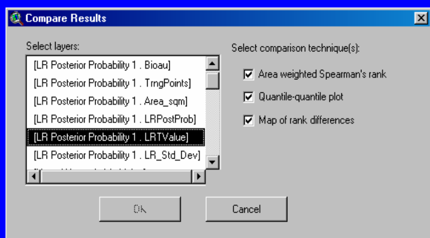
- ArcSDM includes three techniques for comparing the results of different techniques:
  - Spearman's Area Weighted Rank Correlation
  - Quantile-quantile plot
  - Map of rank differences

Compare Results



Compare Results

- Possible inputs:
  - integer grid theme with numeric field(s)
  - floating point grid theme



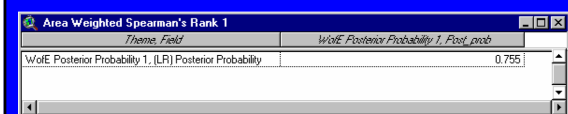
Bonham-Carter, 1999

Compare Results
Spearman's Rank Correlation and Rank Mapping

- Arcview 3 - Classifies both variables into 20 quantiles (ranks). ArcGIS – user specifies number of ranks
- Spearman's Area Weighted Rank Correlation is calculated and written to a dBase file
- Map of rank differences generates a difference map, classifies and symbolizes it to show where the two input evidential themes are similar or dissimilar

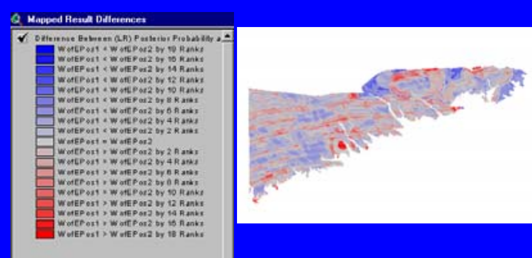
Bonham-Carter, 1999

Compare Results
Spearman's Rank Correlation



Bonham-Carter, 1999

Compare Results
Map of Rank Differences



ArcGIS – User has to symbolize and specify number of classes

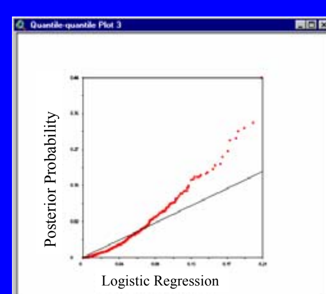
Bonham-Carter, 1999

Compare Results
Quantile-quantile plot

- Sorts the values in each field or theme in ascending order
- if one variable has more observations than the other (for Arcview3), its values are interpolated so that there are equal number of values. ArcGIS: specify number of classes
- values are plotted as x and y coordinates

Bonham-Carter, 1999

Compare Results
Quantile-quantile Plot

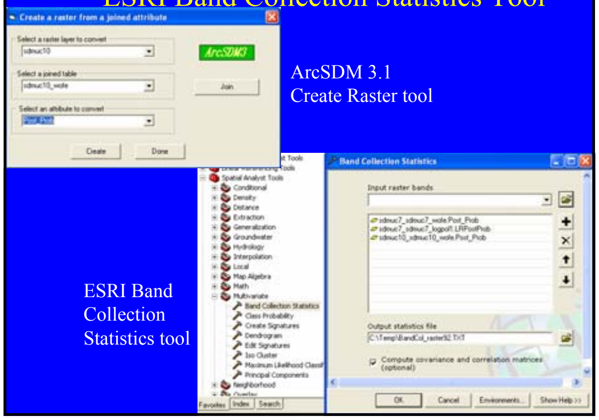


Bonham-Carter, 1999

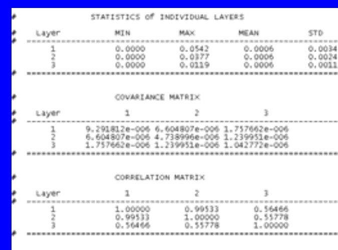
ESRI Band Collection Statistics Tool

ArcSDM 3.1
Create Raster tool

ESRI Band Collection Statistics tool



Band Collection Statistics Report



STATISTICS of INDIVIDUAL LAYERS

Layer	MIN	MAX	MEAN	STD
1	0.0000	0.0542	0.0006	0.0034
2	0.0000	0.0377	0.0006	0.0024
3	0.0000	0.0119	0.0006	0.0011

COVARIANCE MATRIX

Layer	1	2	3
1	9.291812e-006	6.604807e-006	1.757662e-006
2	6.604807e-006	4.738996e-006	1.239951e-006
3	1.757662e-006	1.239951e-006	1.042772e-006

CORRELATION MATRIX

Layer	1	2	3
1	1.00000	0.99533	0.56466
2	0.99533	1.00000	0.55778
3	0.56466	0.55778	1.00000

1. SDMUC7 WofE
2. SDMUC7 LR
3. SDMUC10 WofE

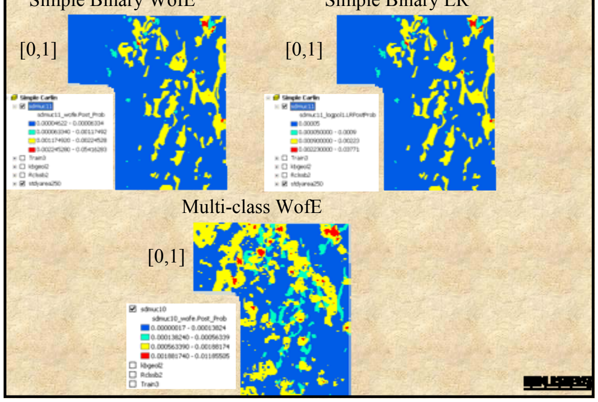
Pearson's Correlation Coefficient

Evolution of Models

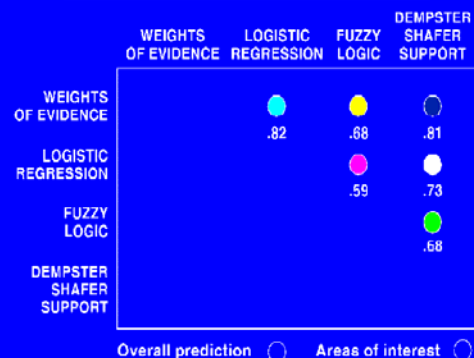
Simple Binary WofE

Simple Binary LR

Multi-class WofE



COMPARISON OF PREDICTIONS



From Wright, 1996

SUMMARY

- Logistic regression can be compared to weights of evidence to check CI assumption
- The total expected number of deposits is usually slightly underestimated by LR (rounding?)
- In general the results of the two methods are similar in terms of ranks, except the WofE probabilities are usually higher than LR probabilities because of CI

Bonham-Carter, 1999

SUMMARY (2)

- ArcSDM will generate LR automatically (expanding the UC table for categorical themes and substituting area-weighted mean values for missing data) at the same time as running WofE, if desired
- Tools for comparing maps are provided in ArcSDM Post Processing and ArcGIS geoprocessing tools.

Bonham-Carter, 1999

Multiple Maps
Fuzzy Logic

Modified from Graeme Bonham-Carter

OUTLINE

- Crisp vs. fuzzy logic
- Fuzzy membership functions
- Fuzzy combination operators
- Application

Crisp Logic

- Membership of crisp set defined as either 1 or 0, True or False
  - (1) Truth(This location is close to a lineament) = 1
  - (2) Truth(This location is on a geochemical anomaly)= 0
- Combination of (1) and (2) by AND, OR, NOT Boolean operators.
  - Truth(1 AND 2) = 0
  - Truth(1 OR 2) = 1

Fuzzy logic

- Fuzzy membership defined in the range [0,1], allowing for gradational membership
  - (1) Truth(This location is close to a lineament) = 0.6
  - (2) Truth(This location is on a soil geochemical anomaly) = 0.9
- Fuzzy operators
  - fuzzy AND, fuzzy OR, fuzzy algebraic SUM, fuzzy algebraic PRODUCT, fuzzy GAMMA, etc
  - Truth(1 Fuzzy Or 2) = 0.9
  - Truth(1 Fuzzy And 2) = 0.6

Fuzzy Membership Functions

- Membership defined by a functional relationship, or by a table of ordered pairs
- Membership reflects degree of truth of some proposition or hypothesis (often a linguistic statement)

Non-spatial example

- Truth of proposition (Person X is Tall)
- Degree of tallness depends on height
- Need a fuzzy membership function relating height to degree of tallness
- In range [0,1], similar to probability, but not satisfying probability laws
- Sometimes termed “possibility”

Tallness

Person	Height	Tallness
Fred	3'2"	0.00
Mike	5'5"	0.21
Sally	5'9"	0.38
Marg	5'10"	0.42
John	6'1"	0.54
Sue	7'2"	1.00

Tallness = 0 if height < 5',
Tallness = (height-5)/2;
if 5 < =height<=7'; or
Tallness = 1 if height > 7'

Truth(Marg is tall) = 0.42

Bonham-Carter, Oct. 1999

Oldness

Person	Age	Oldness
Sally	27	0.21
Mike	30	0.29
Marg	32	0.33
John	41	0.54
Sue	45	0.64
Fred	65	1

Oldness = 0 if age < 18;
Oldness = (age-18)/42
if 18 <= age <= 60; or
Oldness = 1 if age > 60

Truth(Fred is old)=1.00

Bonham-Carter, Oct. 1999

Fuzzy Combination of Tallness and Oldness

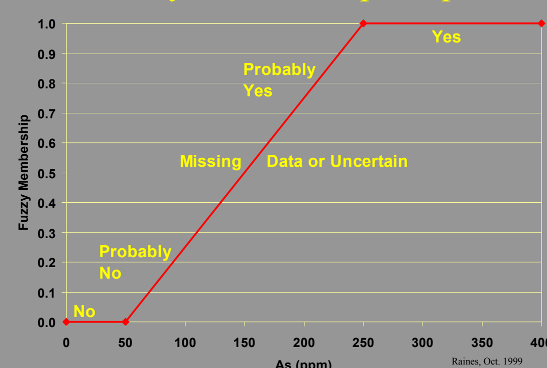
Person	Height	Tallness	Age	Oldness	Tall and old	Tall or old
Fred	3'2"	0.00	65	1.00	0.00	1.00
Mike	5'5"	0.21	30	0.29	0.21	0.29
Sally	5'9"	0.38	27	0.21	0.21	0.38
Marg	5'10"	0.42	32	0.33	0.33	0.42
John	6'1"	0.54	41	0.54	0.54	0.54
Sue	7'2"	1.00	45	0.64	0.64	1.00

Truth(Sally is tall AND old) = min(0.38, 0.21) = 0.21

Truth(John is tall OR old) = max(0.54, 0.54) = 0.54

Bonham-Carter, Oct. 1999

Fuzzy Membership Graph



Raines, Oct. 1999

Fuzzy Membership Function

$$\mu(x) = 0 \text{ if } x < 50$$

$$\mu(x) = \frac{(x-50)}{250} \text{ if } 50 < x < 250$$

$$\mu(x) = 1 \text{ if } x > 250$$

Where μ(x) is the membership value for x

Modified from Bonham-Carter, 2002

Fuzzy Membership Table

Class	Membership	Source Intervals
1	0.8	'142 - 166 ppm As'
2	0.7	'112 - 142 ppm As'
3	0.3	'28 - 52 ppm As'
5	0.2	'17 - 28 ppm As'
6	0.2	'12 - 17 ppm As'
7	0.2	'7 - 12 ppm As'
8	0.2	'2-7 ppm As'
9	0.2	'No data'

Bonham-Carter, Oct. 1999

Fuzzification Functions

- Functions
  - Large
  - Small
  - Near
  - Gaussian
  - MS Small
  - MS Large
  - Linear
- Hedges (square root and squared)
  - Somewhat
  - Very

Fuzzification Functions

Large

$$\mu(x) = \frac{1}{1 + \left(\frac{1}{f2}\right)^{-f1}}$$

Small

$$\mu(x) = \frac{1}{1 + \left(\frac{1}{f2}\right)^{f1}}$$

Near

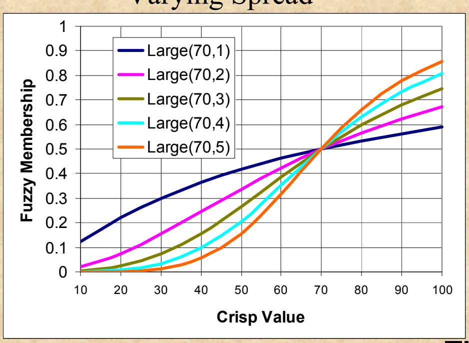
$$\mu(x) = \frac{1}{1 + f1 * (x - f2)^2}$$

Gaussian

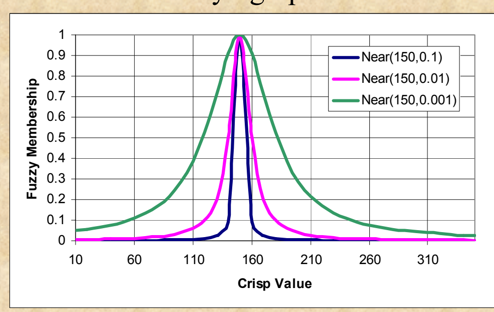
$$\mu(x) = e^{-f1*(x-f2)^2}$$

where f2 = mid point and f1 = spread
μ(x) is the membership value for x

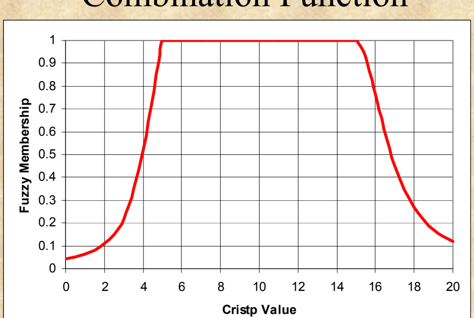
Large Function
Varying Spread



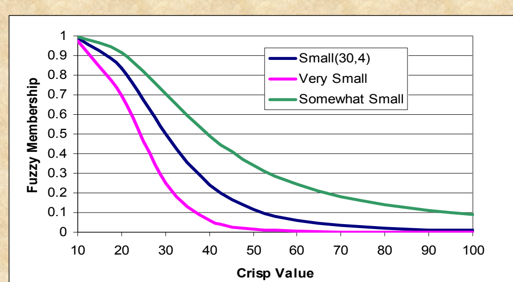
Near Function
Varying Spread



Combination Function



Hedge Applied to Small



Very Small = small squared

Somewhat small = square root of small

Mean-Standard Deviation (MS) Small and Large

MS Large

$$\mu(x) = 1 - \frac{bs}{x - am + bs} \text{ if } x > am$$

Otherwise $\mu(x) = 0$

MS Small

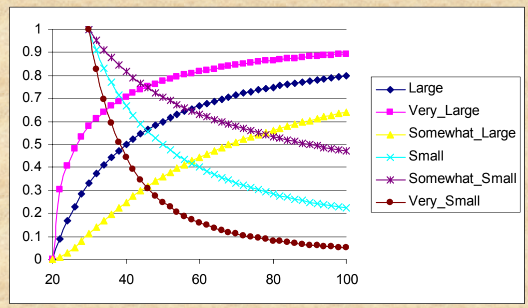
$$\mu(x) = \frac{bs}{x - am + bs} \text{ if } x > am$$

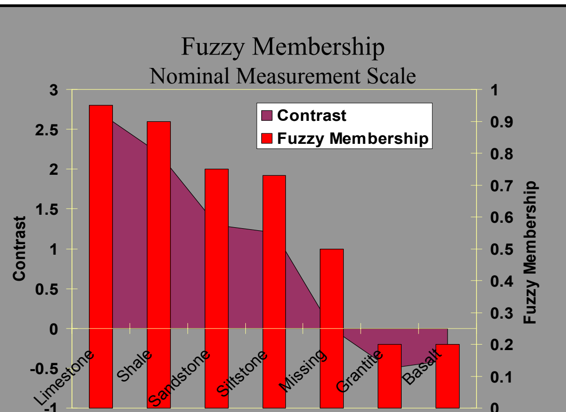
Otherwise $\mu(x) = 0$

Where: m = mean, s = standard deviation

a and b are user input parameters, defaults 1 and 1

Examples of MS Small and Large





Raines, Oct. 1999

Fuzzy Membership in ArcSDM2

Operators
Fuzzy Or and Fuzzy And

Fuzzy Or

$$\mu_{\text{Combination}} = Max(\mu_a, \mu_b, \mu_c, \ldots)$$

Fuzzy And

$$\mu_{\text{Combination}} = Min(\mu_a, \mu_b, \mu_c, \ldots)$$

Combination Operators

$$\text{Fuzzy Product} = \mu_{\text{Combination}} = \prod_{i=1}^{n}(\mu_i)$$

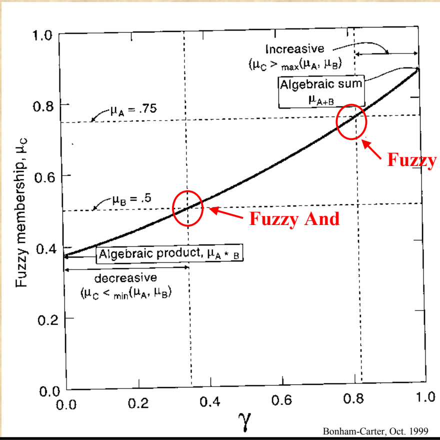
$$\text{Fuzzy Sum} = \mu_{\text{Combination}} = 1 - \prod_{i=1}^{n}(1 - \mu_i)$$

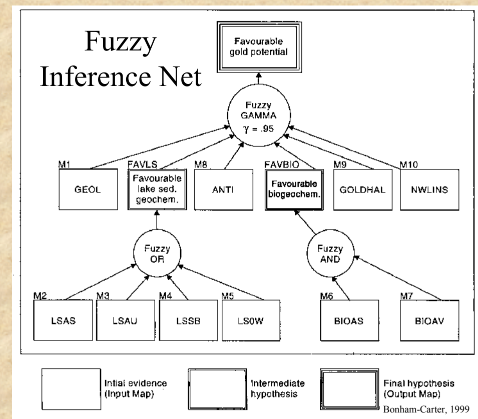
Gamma Operator

$\mu_{\text{Combination}} =$

$(\textit{Fuzzy Sum})^{\gamma} * (\text{Fuzzy Product})^{1-\gamma}$

Where γ = Gamma specified by user





Decisions for Fuzzy Logic

- Fuzzy Memberships
  - Thresholds can be gradational, potentially many values to assign
  - Named classes can be fuzzy, potentially a value for each class
- How the criteria (maps) interact
  - Fuzzy AND, OR, and GAMMA
  - Fuzzy SUM and PRODUCT - not used often
  - Gamma value to define fuzzy relationships of criteria

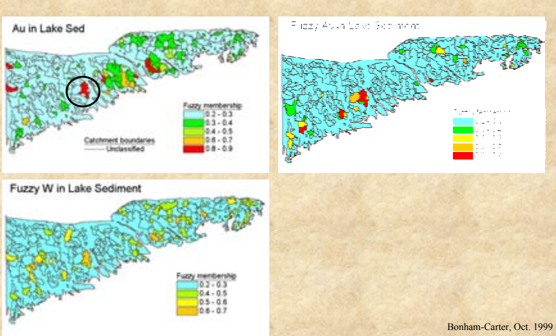
USGS

Fuzzy Logic Summary

- Advantages
  - Flexibility of assigning fuzzy memberships
  - Choice of combination operators
  - Mimic decision making by expert
  - Can deal with “maybe”
  - Not limited to binary criteria
  - Easy to understand
- Disadvantages
  - Problem of missing data
  - Confusion between fuzzy membership and uncertainty
  - Potentially many fuzzy membership values to assign

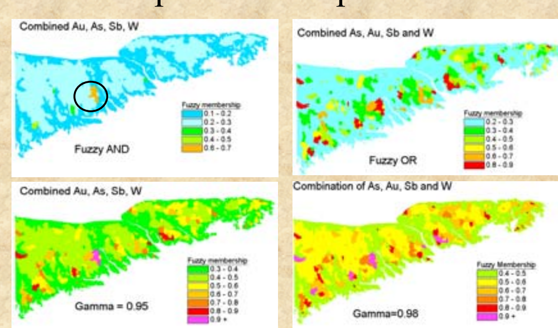
Modified from Bonham-Carter, Oct. 1999; Wright, 1996

Comparison of Fuzzy Evidence

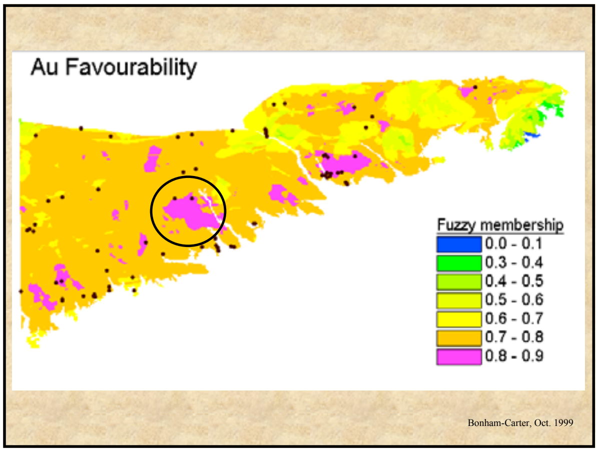


Bonham-Carter, Oct. 1999

Operator Comparison



Bonham-Carter, Oct. 1999


Au Favourability
Fuzzy membership
0.0 - 0.1
0.3 - 0.4
0.4 - 0.5
0.5 - 0.6
0.6 - 0.7
0.7 - 0.8
0.8 - 0.9
Bonham-Carter, Oct. 1999

Neural Networks

Fuzzy Clustering (Unsupervised)
Radial Basis Functional Link Net (Supervised)

Modified from
Carl G. Looney, Prof. of Computer Science
Computer Science and Engineering/171, UNR

1. Intro. to Classification

- Humans accumulate knowledge by grouping observed objects into classes
- This saves the effort of storing every object as a unique item with its own special list of properties
- Classification allows knowledge to be built and organized efficiently

Looney, 2004

1. Intro. to Classification

- Given a population of objects and the goal of classifying them, we must first find measurable properties they all share that

  - distinguish them to some extent
  - allow multiple individuals to be alike

- We call such measurable properties *features*

Looney, 2004

1. Intro. to Classification

- We represent the objects in the population by their *feature vectors*
- It is the set of feature vectors that we classify
- To classify, we must partition, or *cluster*, the feature vectors into groups with similarity within groups, and dissimilarity between groups

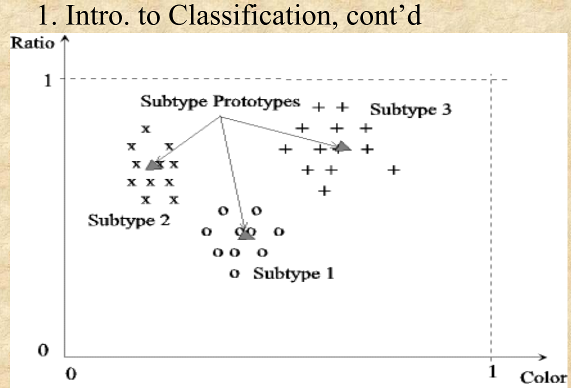
Looney, 2004

1. Intro. to Classification

- Suppose there are 3 types of beetles
- Let us measure the green color intensity x and the height-to-width ratio y
- Then the feature vector for a beetle is (x,y)

Looney, 2004

1. Intro. to Classification, cont'd



Looney, 2004

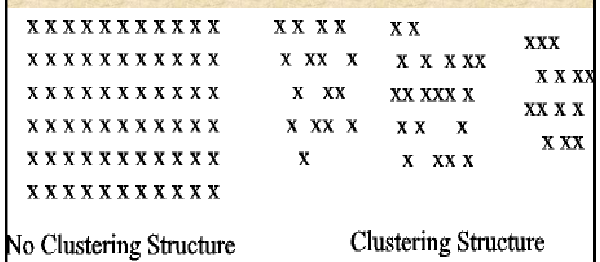
1. Intro. to Classification

- After the clustering into clusters, a vector is used to represent each cluster (called *prototypes* or *centers*) [a cluster is also called a *class*]
- When a new feature vector from that population is to be recognized, it is compared with the prototypes in the various clusters
- It is recognized as belonging to the class that has a prototype most similar to it

Looney, 2004

1. Intro. to Classification

- Data may or may not have clustering structure



Looney, 2004

2. Recognition

- ***Classification***: self-organizing, or unsupervised learning, of classes by a system [e.g., clustering]
- ***Recognition***: supervised learning, or training, of a system to determine which class an input feature vector belongs to [e.g., neural networks]

Looney, 2004

2. Recognition

\+ After classification, we desire to train an on-line automatic recognizer that recognizes the class of any new input vector from the same population.

\+ We use the set of *labeled* feature vectors to ***train*** a

-- fuzzy neural network
-- fuzzy recognizer

Looney, 2004

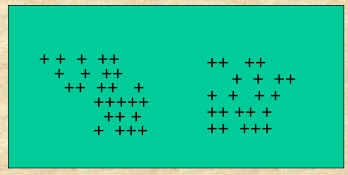
Fuzzy Clustering

Unsupervised Method
No Training Sites Needed

Unique Conditions Table VAT

- Each row can be thought of as a feature vector, **x** = $(x_1, x_2, \ldots x_n)$ where each x_n is the value or attribute of the feature.
  - There are N attributes for any object in a population of objects.
- There are Q rows or feature vectors
- Goal is to partition the population of feature vectors in classes of objects by partitioning the feature vectors.

Classification



Each + represents vectors (s) in the plane, includes error and measurement noise, but on average they fall into two subpopulations (classes).

Weighted Fuzzy Expected Value

WFEV

i) initialize with sample average

$$\mu = \frac{s_1 + s_2 + \ldots + s_P}{P}$$

ii) compute fuzzy weights

$$w_p = \exp[-(s_p - \mu)^2 / (2\sigma^2)] \text{ for } p = [1,\ldots,P]$$

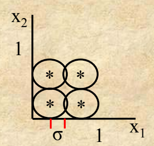
$$W_p = w_p / \sum_{r=1}^{P} w_r$$

iii) compute WFEV

$$\mu = \sum_{p=1}^{P} W_p \bullet s_p$$

iv) if (stop_criteria) then stop

Estimation of Variance



Xs are standardized between [0,1]

N = 2 = number of evidential layers

M = 4 = number of clusters, experience indicates if want 2 final clusters start with M = 10

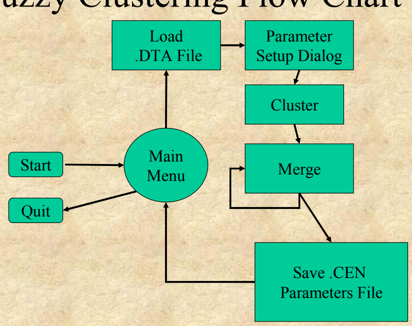
Where M is large enough, then can initially estimate the variance by

$$\sigma = \frac{1}{4} * \left(\frac{1}{M}\right)^{\frac{1}{n}} = \frac{1}{4} * \left(\frac{1}{4}\right)^{\frac{1}{2}} = \left(\frac{1}{8}\right) = 0.125$$

Fuzzy Clustering Algorithm

- Input a number K of classes that is larger than the expected number of classes
- Assign first K of the Q vectors as cluster centers $z^{(1)}, \ldots, z^{(K)}$
- For q = 1 to Q
  - Assign $x^{(q)}$ to closest $z^{(K)}$ by c[q]=k
  - Find WFEV for each cluster to obtain a new center $\{z^{(K)}\}$
  - If(any center changes more than ε) start over
  - Else Compute weighted fuzzy variance for each cluster and WFEV d_{WFEV} of distances between centers
- for k =1 to K-1
  - for kk = k+1 to K
    - if distance($z^{(k)}$,$z^{(kk)}$)<ßd_{WFEV} then merge (k,kk)

Fuzzy Clustering Flow Chart



Calibration of Fuzzy NN Models

- Cluster validity – make as small as possible

Radial Basis Function Link Net

Supervised

Training Sites Required

Radial Basis Functional Link Nets

- A radial basis functional link net (neural network, NN) transforms each N-dimensional input ***feature vector*** into an output ***target vector***
  - $\mathbf{x} = (x_1,\ldots,x_n) \rightarrow \mathbf{NN} \rightarrow \mathbf{t} = (t_1,\ldots,t_n)$
- **Target vector t** is a code word that represents a class. This is called ***supervised learning*** because the network must be told the class for each input feature vector x.
- NNs have a relatively large number of parameters that can be thought of as dials. The parameters are also known as ***weights***.
- During training a set of feature vectors are presented to the network and the dials are adjusted until each feature vector is mapped to its known target vector
  - These feature vectors are called ***training vectors*** when used to train the network.

Diagram of Process

$$x = (x_1,\ldots,x_n) \rightarrow NN \rightarrow z = (z_1,\ldots,z_n) \rightarrow e \leftarrow t = (t_1,\ldots,t_n)$$

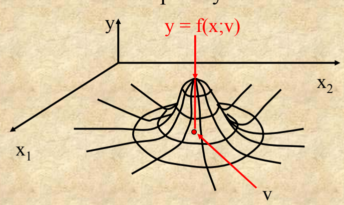
The error to be minimized over all Q input feature vectors is

$$\mathrm{E} = \sum_{q=1}^{Q}\sum_{j=1}^{J}\left(t_j^q - z_j^q\right)^2$$

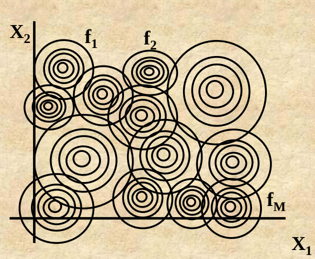
In our case j = 1 because only one target value.

Radial Basis Function

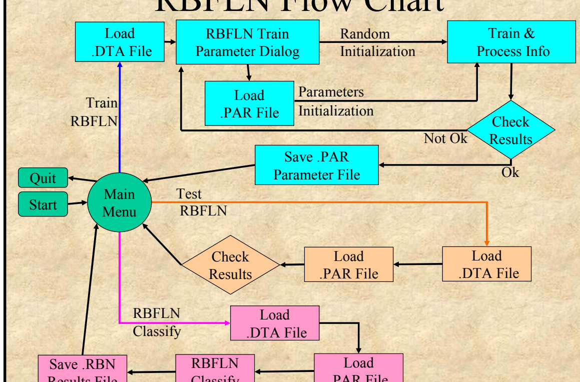
- RBF is a Gaussian function. It has a center vector v and processes any input vector x via $y = f(x;v) = \exp[(x-v)^2/(2\sigma^2)]$ $(0<y\leq 1)$
- Each middle-layer node in RBFN or RBFLN contains a RBF whose output fans out to each node in the output layer.



RBF Contour Curves in the Plane



RBFLN Flow Chart



Input Data Format

N M J Q

$\mathbf{g_1^{(1)}, g_2^{(1)}, g_3^{(1)}, x_1^{(1)}, x_2^{(1)}, \ldots, x_N^{(1)}, t_1^{(1)}}$

$\mathbf{g_1^{(2)}, g_2^{(2)}, g_3^{(2)}, x_1^{(2)}, x_2^{(2)}, \ldots, x_N^{(2)}, t_1^{(2)}}$

……………………

……………………

$\mathbf{g_1^{(Q)}, g_2^{(Q)}, g_3^{(Q)}, x_1^{(Q)}, x_2^{(Q)}, \ldots, x_N^{(Q)}, t_1^{(Q)}}$

The t_1 values are the training-set fuzzy-membership output values. This allows for ranking of training sets.

- N = number of evidential layers
- M = number of nodes (RBFs) in middle layer
- J = number of output classes = 1
- Q = number of feature vector/target vector pairs, that is number of unique conditions
- $g_1^{(1)}, g_2^{(1)}, g_3^{(1)}, x_1^{(1)}, x_2^{(1)}, \ldots, x_N^{(1)}$ = first input feature vector, $g_1^{(1)}$ is the key field to join with unique conditions table.
- $t_1^{(1)}$ = first target output value in [0,1], where
  - **1 = yes and 0 = no**
  - **0.9 = strong indication of yes**
  - **0.1 = strong indication of no**
  - **Can use to say "kind of like" a training site!**

Output Results File

$\mathbf{g_1^{(1)}, c^{(1)}, f_1^{(1)}, f_2^{(1)}, \ldots, f_K^{(1)}}$

$\mathbf{g_1^{(2)}, c^{(2)}, f_1^{(2)}, f_2^{(2)}, \ldots, x_K^{(2)}}$

……………………

……………………

$\mathbf{g_1^{(Q)}, c^{(Q)}, f_1^{(Q)}, x_2^{(Q)}, \ldots, x_K^{(Q)}}$

- $g_1^{(q)}$ is the key field to join with unique conditions table
- $c^{(q)}$ is the fuzzy class number
- $f_1^{(q)}$ fuzzy membership values, respectively for input vector q belonging to class k = 1, …,K.

Decision with Neural Networks

- Transform evidential values into range [0,1]
  - Can use fuzzy membership values as inputs
  - Possibly can use value field
- Ranking of training sites
- Evaluation of reported measures of classification

Calibration Measures of RBFLN

- Minimize number of clusters, M.
- Small number of iterations
  - Over fitting
- Calibration measures.

PNN

- What say?
- Uses circular functions in space.
- Measures
  - Nearness of cluster
  - Measures
  - Overfitting

Summary

- Advantages
  - Can rank training sites
  - Non-linear mathematics
  - Unsupervised and Supervised method
- Disadvantages
  - Model parameters are difficult to understand
  - Need training sites for occurrence and non-occurrence
  - Approaches to ranking of training sites not well understood
  - Overall use is poorly understood

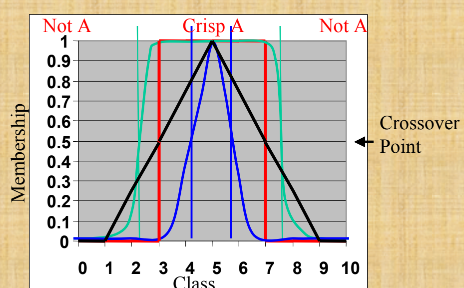
Miscellany

Fuzzy Membership
Nature of Evidence
Semantic Classification of Response
Testing of Predictions

Fuzzy Membership

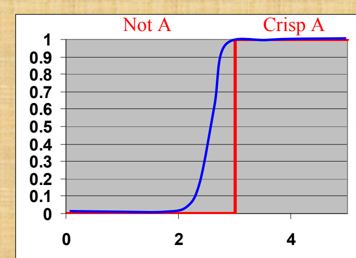
Semantic Approach

Membership Functions



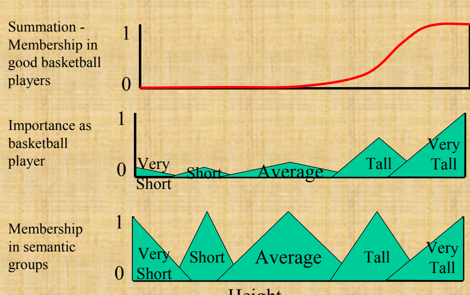
Crossover Point

Membership Function

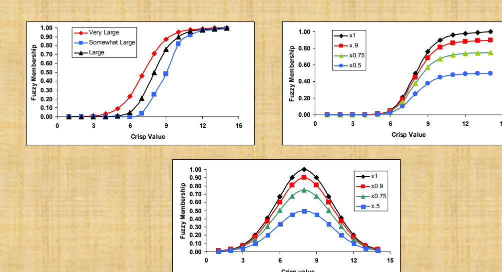


Crossover Point

Semantic Summation



Ranking Fuzzy Membership



Speculations on the Nature of Evidence

Generalization

Positive and Negative Evidence

Categorical Generalization Simple

Categorical Class	Area	#Points	Contrast	
Sand	3000	35	2.0	Inside
Lime mud	2000	3	-1.0	Outside
Shelly sand	75	0	Null or 0	

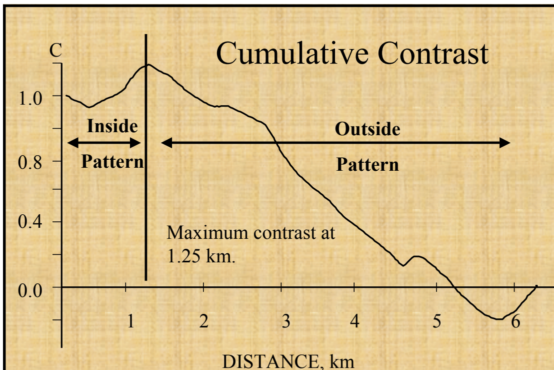
- **Assume both have significant Studentized Contrast**
- **Always check categorical generalization by calculating weights of the generalization**
- **Does Shelly sand belong with Lime mud?**

Categorical Generalization Expert Interpretation

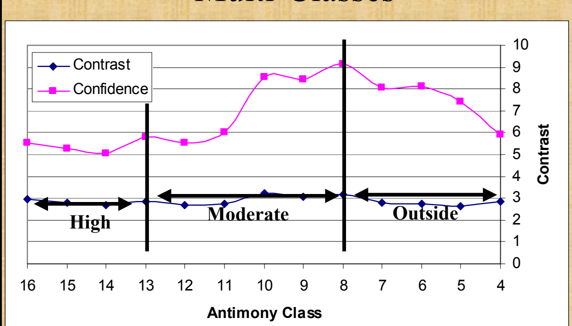
Categorical Class	Area	#Points	Contrast	
Sand	3000	35	2.0	Inside
Lime mud	2000	3	-1.0	Outside
Shelly sand	75	0	Null or 0	Inside

- **For percolation of water through sediments, Sand and Shelly sand are more alike than Lime mud!**
- **Always check categorical generalization by calculating weights of the generalization**

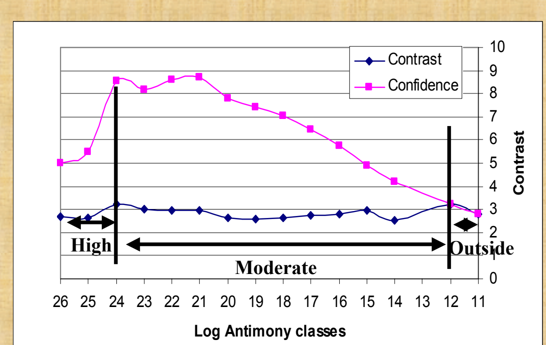
Cumulative Contrast



Multi-Classes

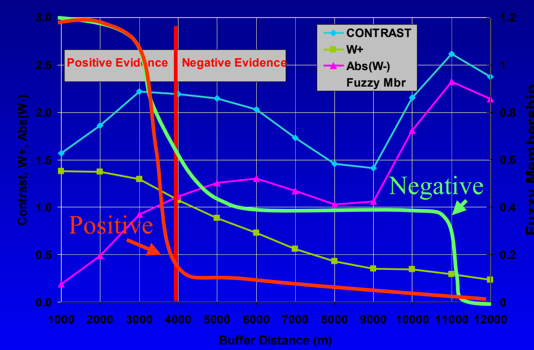


Another Multi-Class



Positive and Negative Evidence

Fuzzy Membership - Ordered Variables



What are the rules of Generalization?

- A model should follow a consistent rule of generalization.
- Rules might define how to consistently derive specific types of models.
- Models are always wrong but sometimes useful!
- Not all models are equal.
- Is there a best model or simply a collection of better models?

Semantic Classification of Response

- What is the significance of a particular posterior probability or fuzzy membership value?
- Have to interpret in context of the model.
  - Number of training sites: Do you have a large or small sample of the possible training sites?
  - State of knowledge about process being modeled: How good is the scientific understanding of the process?
  - Quality of the evidence: Consider accuracy and precision of the values and the location.

Solution

- Analyst has the best understanding of the significance of the response value.
  - Highest posterior probability may not be a high or large value. Might be quite low.
  - Consider the meaning of the prior probability
- Assign names to intervals of response values.
  - Carefully consider the meaning or implication of the selected terms.

Testing of Predictions

- How well does the Response value predict the training sites?
  - ArcSDM2: Associate Responses with Point Theme
- Experimental Design
  - Hold back training sites to test the model
  - ArcSDM2: Associate Responses with Point Theme
- Field studies

Chapter 9

Summary

Comments on Exercise 1

Guidelines for Modeling

- Formal statement of the problem.
- Define the user of the model.
- Specification - preprocess the data to provide useful information, that is evidence.
  - Data exploration
  - Data transformation, filtering, and scaling
  - Reduce the dimensionality by eliminating redundant or correlated information
  - Use the minimum information necessary
- Prediction - combine the evidence to create the model.
- Testing - evaluate the model and it's properties.

Purpose of GIS Projects

- Combine data from diverse sources
- To describe and analyze interactions
- To make predictions, that is models
- To provide support for decision makers

Properties of Evidence

- Selected attributes must discriminate between one or more classes of objects.
- Selected attributes must not be correlated with other attributes to any moderately strong extent.
- Selected attributes must have meaning for humans.

Types of Models

- Prescriptive or Deterministic
  - Application of good technical practices
  - Process: Boolean rules, Equations
  - Output: Binary (yes or no), Index overlay (score)
- Predictive
  - Application of mathematics to represent how people think about the evidence but cannot represent as equations.
  - Process: weighting of evidence and combination of weights
  - Output: Favorability, probability, or fuzzy map [0 to 1]

Knowledge Driven Methods

- **Boolean Logic** - True/False representation of maps with all maps rated equally. Simple method with True/False answer.
- **Index Overlay with Binary Maps** - Maps are given different weights. Linear combination of maps.
- **Index Overlay with Multi-Class Maps** - Maps are given different weights as well as the classes of the maps are given different weights. Linear combination of maps.
- **Fuzzy Logic** - More flexible weighting of maps and map classes. Nonlinear combination of maps.
- **Expert Weights of Evidence** – Weighting of evidence easily understood. Log linear combination of maps.

Data Driven Methods

- Weights of Evidence
  - log linear combination of binary or multi-class maps.
  - Classifies areas by probability or favorability of occurrence of a training site.
  - Model parameters easy to understand.
- Logistic Regression
  - log regression combination of binary maps
  - Classifies areas by probability of occurrence of a training site.
  - Model parameters complex.
- Neural networks
  - Experimental, nonlinear combination of fuzzy or map classes
  - Classifies areas by fuzzy membership in training set.
  - Can also be self organizing to produce fuzzy membership.
  - Model parameters complex.

Recognition of a Pattern

- Task - Determine what the appropriate level of aggregation and simplification is for the problem at hand, a problem of reclassification.
  - Aggregation and simplification are tied to scale of observation.
  - The is no single scale at which to view a system.
  - Does not mean that all scales serve equally well or there are not scaling laws.
- Description of patterns is the starting point.
- Spatial models start with an assemblage of patterns and associated processes.

Examples of Measurement Scales

Scale Type	Examples	Operations	Means
Nominal	Rock type	=	Mode
Ordinal	Relative age	><	Median
Interval	Temperature	+-*/	Mean
Ratio	Distance	+-*/	Mean

Buffer Resolution

Threshold Weighting Reclassification

Map Scale	Map Resolution	Geologic Resolution	Buffer Resolution
1:2,500,000	1250	2500	5000
1:500,000	250	500	1000
1:250,000	125	250	500
1:100,000	50	100	200

Units - Meters

Map Resolution = (Scale denominator)/2000

Testing Data-driven Methods

- Evaluate classification of training points
  - **Associate Points with Response**
  - **Efficiency of Classification**
- Use points not included in training set to test the model
  - Implementation - use a random subset of training set to develop the weights and use the remainder to evaluate the model. (**Efficiency of Prediction**)
  - Problem - for many models there may only be a small number of training points to start with.
  - Field Studies

Spatial Analysis in GIS
Map Pairs

- Overlay
- Map Correlation

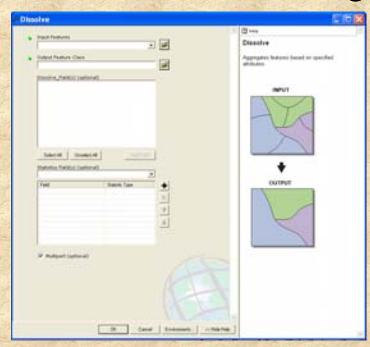
Guidelines for Modeling

- Formal statement of the problem.
- Define the user of the model.
- Specification - preprocess the data to provide useful information, that is evidence.
  - Data exploration
    - Reclassification, filtering, transformation, and scaling
  - Reduce the dimensionality by eliminating redundant or correlated information
  - Use the minimum information necessary
- Prediction - combine the evidence to create the model.
- Testing - evaluate the model and it's properties.

Spatial Joins

- Define relationships between features in layers.
- Apply to points, lines, or polygons
- Nature of the join changes as a function of what type of spatial layers are being joined

Dissolve Management



Useful for hierarchical data

Example

Sedimentary Rock

Carbonate

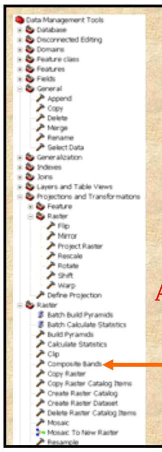
Limestone

Dolomite

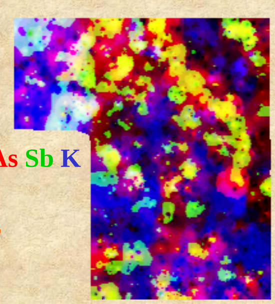
Combining Grids

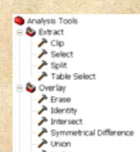
- Zonal Statistics - summarize one grid for zones in another grid or shape file
- Map Calculator - some sort of map algebra
  - Combine
  - Con
- Multivariate Statistics (Scaling Issues)
  - Maximum Likelihood Classification
  - Principal Components
- Merge grids
  - Unique polygons
  - Unique conditions

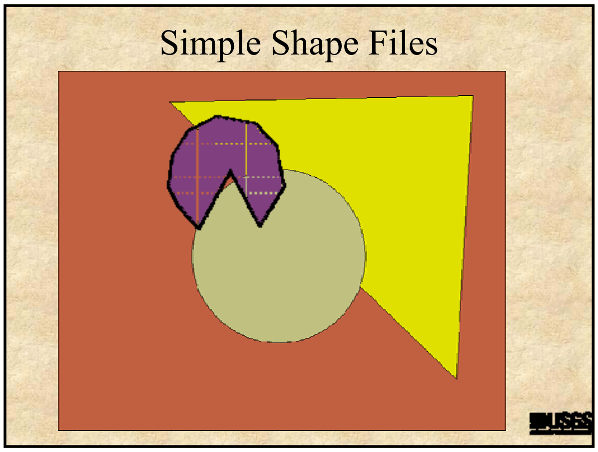
Spatial and Other Transforms of Rasters

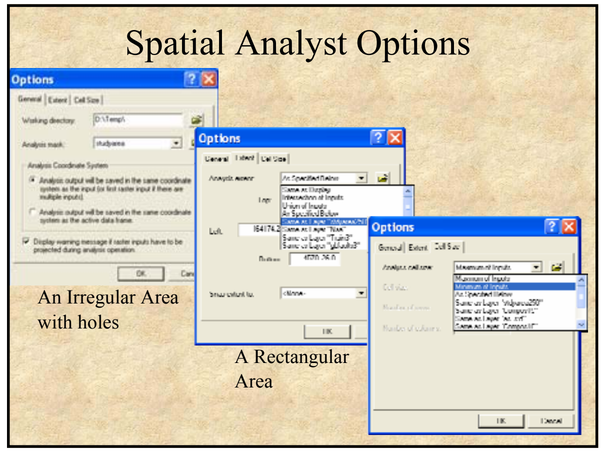


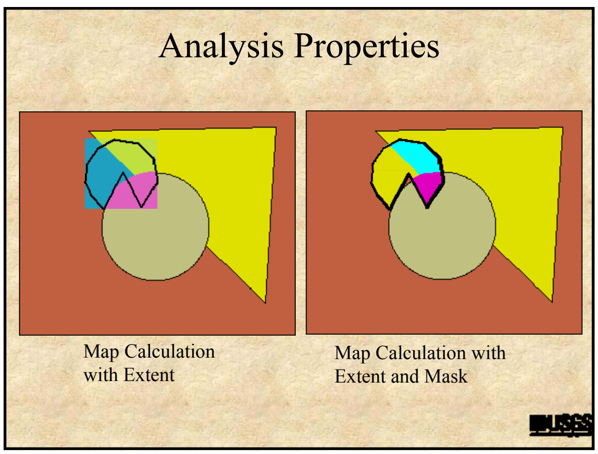
As Sb K

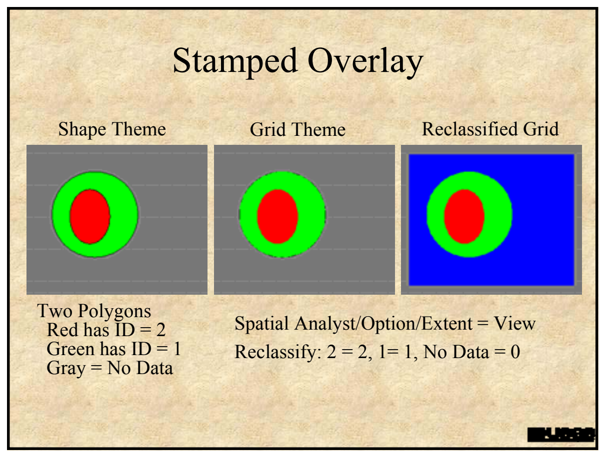


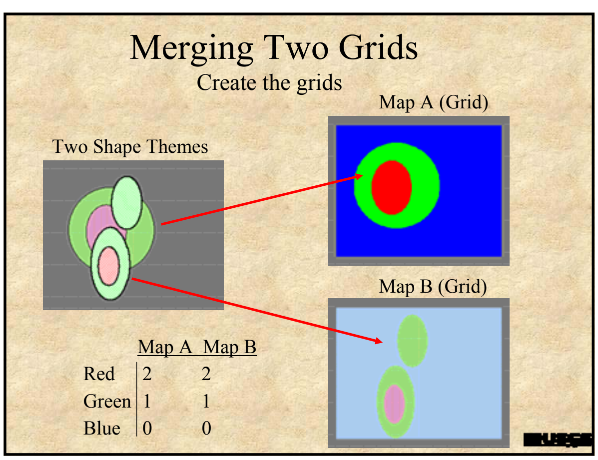


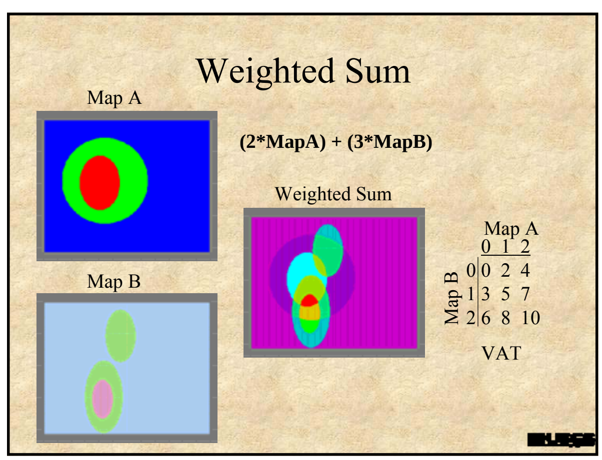

Simple Shape Files


Spatial Analyst Options
Options
An Irregular Area
with holes
A Rectangular
Area


Analysis Properties
Map Calculation
with Extent
Map Calculation with
Extent and Mask

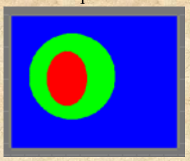

Stamped Overlay
Shape Theme
Grid Theme
Reclassified Grid
Two Polygons
Red has ID = 2
Green has ID = 1
Gray = No Data
Spatial Analyst/Option/Extent = View
Reclassify: 2 = 2, 1= 1, No Data = 0


Merging Two Grids
Create the grids
Map A (Grid)
Two Shape Themes
Map B (Grid)
Map A Map B
Red 2 2
Green 1 1
Blue 0 0

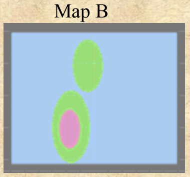

Weighted Sum
Map A
(2*MapA) + (3*MapB)
Weighted Sum
Map B
Map A
0 1 2
0 0 2 4
1 3 5 7
2 6 8 10
Map B
VAT

Conditional Overlay
(Join overlay in text)

Map A

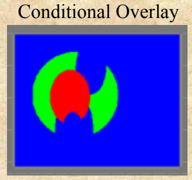


Map B



Con(MapA > MapB, MapA,0)

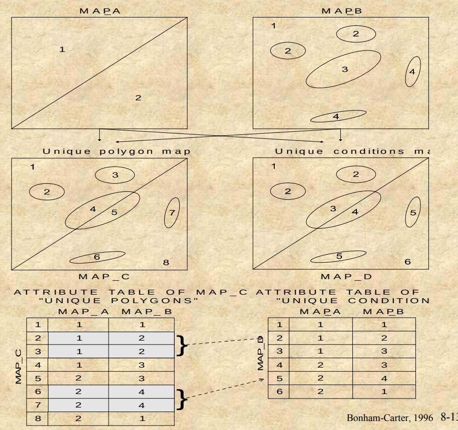
Conditional Overlay



		Map A		
		0	1	2
Map B	0	0	1	2
	1	0	0	2
	2	0	0	0

VAT values

Stamped, Joined, Compare



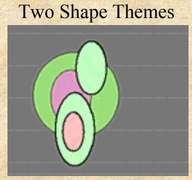
MAP_C	MAP_A	MAP_B
1	1	1
2	1	2
3	1	2
4	1	3
5	2	3
6	2	4
7	2	4
8	2	1

MAP_D	MAPA	MAPB
1	1	1
2	1	2
3	1	3
4	2	3
5	2	4
6	2	1

Bonham-Carter, 1996 8-13

Unique Polygons vs. Unique Conditions

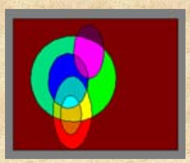
Two Shape Themes



Unique Polygons

14 polygons

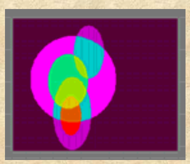
Shape file or Grid



Unique Conditions

9 Classes

Grid theme



Transform Grid/Combine

sptrnfrm.avx

Problem with the VAT

Unique Conditions

9 Classes

Grid theme

VALUE	COUNT	Map B	Map A
1	53517	0	0
2	3291	1	0
3	9356	0	1
4	4139	1	1
5	2971	0	2
6	2642	1	2
7	545	2	2
8	1071	2	1
9	718	2	0

Value not sorted with regards to Map B and Map A values.

Frequency

VAT with Case added

VALUE	COUNT	Map B	Map A	CASE
1	53517	0	0	1
3	9356	0	1	2
5	2971	0	2	3
2	3291	1	0	4
4	4139	1	1	5
6	2642	1	2	6
9	718	2	0	7
8	1071	2	1	8
7	545	2	2	9

Consistent numbering of the matrix or VAT.

Do this for Shapefile in ArcGIS 9.1.

Can do in ArcGIS 9.2

ArcMap: ArcToolBox Analysis Tools/Statistics/Frequency

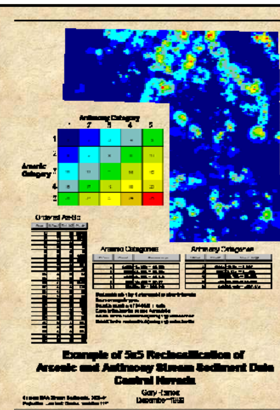
Sort and add Case in Excel

Or

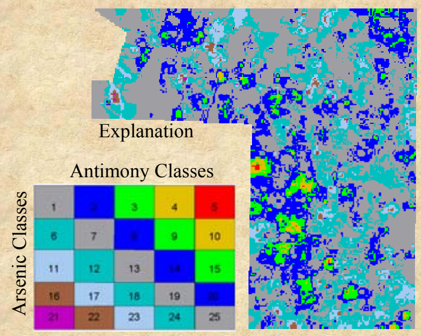
ArcMap: Symbolize by multiple attributes

Grid Overlay Application

- Reclassification rasters
- Convert to Shapefiles
- Union Shapefiles
- Symbolize by Multiple Attributes.
  - Add Sorting Attribute [1-25]
  - Convert to raster on Sorting Attribute.



Correlation Analysis



Summary

- Shape files - several tools
  - Computations can be slow
- Grid overlay offers great flexibility
  - Numerical and logical combinations
  - Ordered VAT or table of combinations opens the door for many types of modeling
  - Unique conditions table shortens the ordered matrix and simplifies programming in modeling
  - Computations are very fast

Spatial Analysis in GIS
Map Pairs

•Map Correlation

Probability

Put 3 red balls and 7 blue balls in a bag.

What is the probability of drawing a blue ball from the bag?

What is the probability of drawing a red ball from the bag?

Probability of drawing a blue ball is 7/10 = 0.7 = P_b

Probability of drawing a red ball is 3/10 = 0.3 = P_r = 1- P_b

Probability

Probability	Put the following balls in a bag:	
0.3	3 red-blue balls	----Red-Blue (RB)
0.2	2 red-green balls	----Red-Not Blue (RG)
0.1	1 blue-green ball	----Blue-Not Red (BG)
0.4	4 green balls	----Not Red-Not Blue (G)

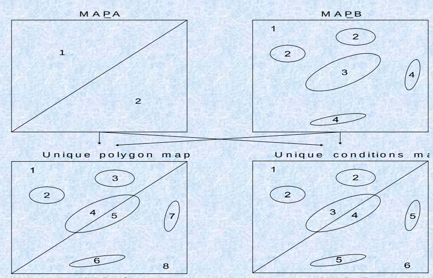
Area Tabulation Table

	R	BG	
B	$3=T_{11}$ $0.3=P_{11}$	$1=T_{12}$ $0.1=P_{12}$	$4=T_{1.}$ $0.4=P_{1.}$
RG	$2=T_{21}$ $0.2=P_{21}$	$4=T_{22}$ $0.4=P_{22}$	$6=T_{2.}$ $0.6=P_{2.}$
	$5=T_{.1}$ $0.5=P_{.1}$	$5=T_{.2}$ $0.5=P_{.2}$	$10=T_{..}$

What is the probability of drawing each type?

What is probability of drawing a blue ball?

Marginal Probability of a blue ball = 0.4



ATTRIBUTE TABLE OF MAP_C "UNIQUE POLYGONS"

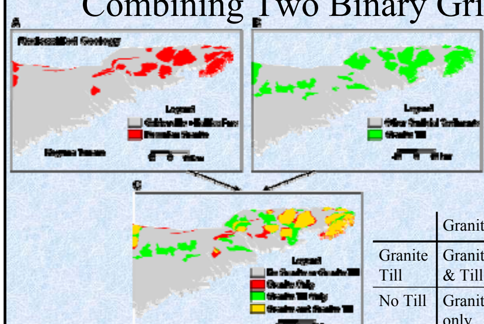
MAP_C	MAP_A	MAP_B
1	1	1
2	1	2
3	1	2
4	1	3
5	2	3
6	2	4
7	2	4
8	2	1

ATTRIBUTE TABLE OF "UNIQUE CONDITION

MAP_D	MAPA	MAPB
1	1	1
2	1	2
3	1	3
4	2	3
5	2	4
6	2	1

Bonham-Carter, 1996 Fig. 8-13

Combining Two Binary Grids



	Granite	No Granite
Granite Till	Granite & Till	Granite Till only
No Till	Granite only	No Granite or Till

Bonham-Carter, Fig. 8-14

Tabulate Areas or Unique Conditions

Reclassified Antimony

Reclassified Arsenic

VALUE	VALUE_1	VALUE_2	VALUE_3	VALUE_4	VALUE_5
1	881207640.750	411230232.350	2098113.430	0.000	0.000
2	3212211661.900	26041783897.000	354581169.730	0.000	0.000
3	39864155.177	9426823642.600	2979321071.100	10490567.152	0.000
4	0.000	117494352.100	448996274.100	46158495.468	0.000
5	0.000	0.000	23079247.734	4196226.861	2098113.430
	Units = Area			Correlation Coefficient	
					0.803

VALUE	VALUE_1	VALUE_2	VALUE_3	VALUE_4	VALUE_5
1	2.00	0.93	0.00	0.00	0.00
2	7.30	59.18	0.81	0.00	0.00
3	0.09	21.42	6.77	0.02	0.00
4	0.00	0.27	1.02	0.10	0.00
5	0.00	0.00	0.05	0.01	0.00
	Units = Percent of Area				Agreement
					68.07

Agreement = 100*(Sum of Diagonal (gray cells)/Total).

Also called area cross tabulation or confusion matrix.

Nominal Scale Data

Information Statistic

Area-Proportions Cross-Tabulation Table

Map B \ Map A					
	P_{11}	P_{12}	P_{13}	…	$P_{1.}$
	P_{21}	P_{22}	P_{23}	…	$P_{2.}$
	P_{31}	P_{32}	P_{32}	…	$P_{3.}$
	…	…	…	…	…
	$P_{.1}$	$P_{.2}$	$P_{.3}$	…	$P_{..}$

Where
$P_{ij} = T_{ij}/T_{..}$
$P_{i.} = T_{i.}/T_{..}$
$P_{.j} = T_{.j}/T_{..}$

Information Statistics

$$H(A) = -\sum_{j=1}^{m} p_{\bullet i} \ln p_{\bullet j}$$

$$H(B) = -\sum_{j=1}^{n} p_{i\bullet} \ln p_{i\bullet}$$

Joint Entropy

$$H(A,B) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \ln p_{ij}$$

Joint Information Uncertainty

$$U(A,B) = 2\left[\frac{H(A)+H(B)-H(A,B)}{H(A)+H(B)}\right]$$

Nominal Scale Data

Coefficient of Agreement, kappa

Area-Proportions Cross-Tabulation Table

Map B \ Map A					
	P_{11}	P_{12}	P_{13}	…	$P_{1.}$
	P_{21}	P_{22}	P_{23}	…	$P_{2.}$
	P_{31}	P_{32}	P_{32}	…	$P_{3.}$
	…	…	…	…	…
	$P_{.1}$	$P_{.2}$	$P_{.3}$	…	$P_{..}$

Where
$P_{ij} = T_{ij}/T_{..}$
$P_{i.} = T_{i.}/T_{..}$
$P_{.j} = T_{.j}/T_{..}$

$$\kappa = \frac{\sum_{i=1}^{n} p_{ii} - \sum_{i=1}^{n} q_{ii}}{1 - \sum_{i=1}^{n} q_{ii}}$$

where
$q_{ij} = p_{i\bullet} * p_{\bullet j}$
n = number of classes, which is the same in both maps.
Conditional kappa for the i-th class

$$\kappa_i = \frac{p_{ii} - q_{ii}}{p_{i.} - q_{ii}}$$

Weighted Pearson's Correlation Coefficient

Modified for Cross-Tabulation Table

$$r_s = \frac{\sum_{i=1}^{columns}\sum_{j=1}^{rows} T_{ij} * \left(X_i - \overline{X}\right)\left(Y_j - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{columns}\sum_{j=1}^{rows} T_i * (X_i - \overline{X})^2 \sum_{i=1}^{columns}\sum_{j=1}^{rows} T_i * (Y_j - \overline{Y})^2}}$$

where
X_i and Y_j = Values of Map X and Map Y respectively
$\overline{X}$ and $\overline{Y}$ = Area-weighted mean respectively
T_{ij} = area or count in the i^{th} - j^{th} pair or cell of the cross-tabulation matrix.

Same as in Excel!

Band Collection Statistics: Pearson

STATISTICS of INDIVIDUAL LAYERS

Layer	MIN	MAX	MEAN	STD
1	2.0604	621.9418	25.9557	25.7315
2	0.1713	1475.7814	6.7429	17.8302

COVARIANCE MATRIX

Layer	1	2
1	535.13788	120.38564
2	120.38564	256.94983

CORRELATION MATRIX

Layer	1	2
1	1.00000	0.32465
2	0.32465	1.00000

Layer 1 = As

Layer 2 = Sb

Both IDW defaults as real-valued rasters

Same as in Excel!

Ordinal Data

Weighted Spearman's Rank Correlation

$$r_s = 1 - \frac{6 * \sum_{i=1}^{rows}\sum_{j=1}^{columns} T_{ij} * \left(R_{xi} - R_{yj}\right)^2}{n\left(n^2 - 1\right)}$$

where
R_x and R_y are ranks for Maps X and Y
n = sum of cells in cross-tabulation matrix

Use this formula:

where ranks are numbered 1, 2, 3, …,n and

where there are no ties.

If have ties, then ranks are given average of the ranks!

Bonham-Carter's Modification

Weighted Spearman's Correlation Coefficient

$$r_{sw} = \frac{\sum_{i=1}^{columns}\sum_{j=1}^{rows} T_{ij} * \left(R_{xi} - \overline{R}_x\right)\left(R_{yj} - \overline{R}_y\right)}{\sqrt{\sum_{i=1}^{columns}\sum_{j=1}^{rows} T_{ij} * (R_{xi} - \overline{R}_x)^2 \sum_{i=1}^{columns}\sum_{j=1}^{rows} T_{ij} * (R_{yj} - \overline{R}_y)^2}}$$

where
R_{xi} and R_{yj} = ranks of Map X and Y respectively
$\overline{R}_x$ and $\overline{R}_y$ = Area-weighted average rank respectively
T_{ij} = area or count in the i^{th} - j^{th} pair.

If the area-weighted average rank is simply the normal area-weights average this is the same as Pearson Correlation Coefficient. This is what is done in ArcSDM 3.1

Spearman's - Ordinal Data

Table 8-9B

Tabulate Area Table

T_{ij}

	Map A (x)					Sum	Cum.	Rank Y
Map B (y)	80	0	18	0	0	98	98	49
	37	71	0	0	0	108	206	152
	0	0	0	10	0	10	216	211
	0	0	0	22	30	52	268	242
	0	0	17	3	51	71	339	303.5
Sum	117	71	35	35	81		169.5	
Cum.	117	188	223	258	339	169.5		
Rank X	58.5	152.5	205.5	240.5	298.5			

Sum = row or column sum

Cum. = Cumulative row or column sum

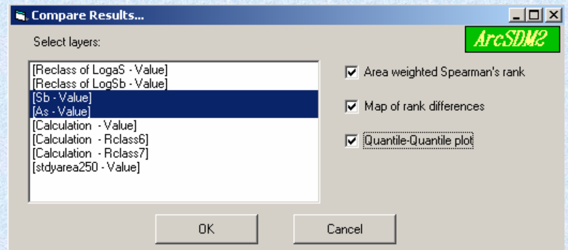
$$r_{swG} = 0.826$$

$$r_{sw} = 0.874$$

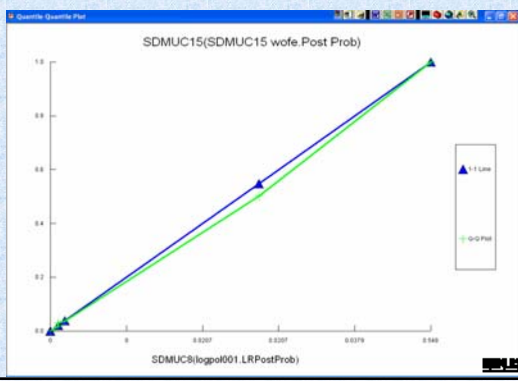
$$RankX_i = CumX_{i-1} + \frac{SumX_i}{2} = 223 + \frac{35}{2} = 240.5$$

$$\overline{RankX} = \frac{Max(CumX_i)}{2} = \frac{339}{2} = 169.5$$

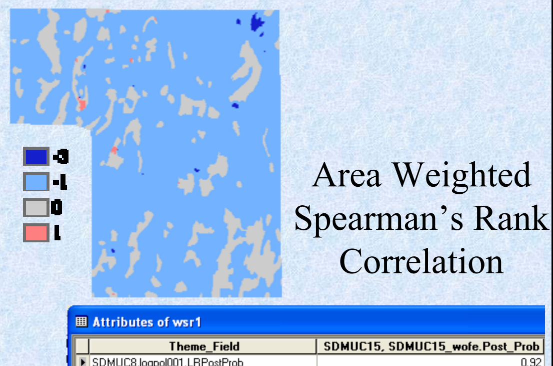
ArcSDM2 Compare Results



Quantile-Quantile Plot



Rank Difference

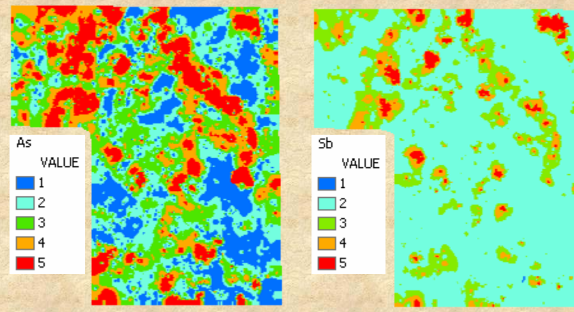


Area Weighted Spearman's Rank Correlation

Spatial Analysis in GIS
Map Pairs

- Map Correlation

Input As and Sb Rasters
Reclassified into 5 classes by Quantiles



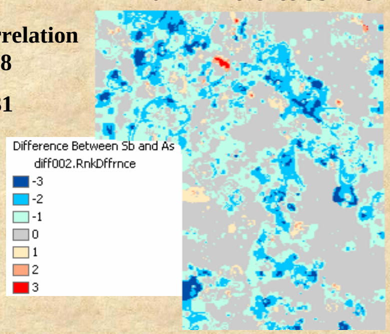
Rasters created by IDW

Spatial Correlation

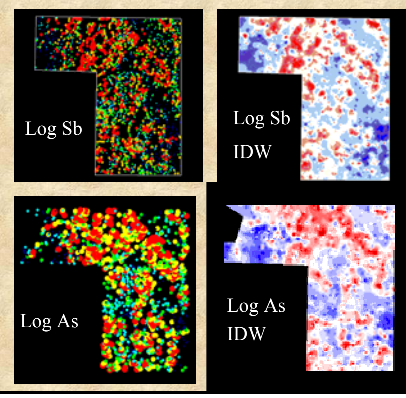
Area-Weighted Spearman's Correlation Coefficient = 0.28

Agreement = 0.31

Rank Difference Sb - As



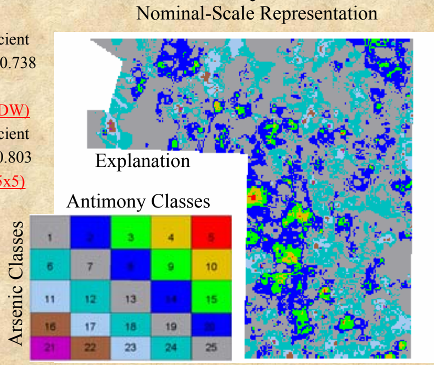
Source Ratio Data



Correlation Analysis

Points (ratio scale)
Correlation Coefficient
Log As:Log Sb = 0.738
As:Sb = 0.3000
Grid (ratio scale, IDW)
Correlation Coefficient
Log As:Log Sb = 0.803
Grid (reclassified 5x5)
Agreement = 68%

Nominal-Scale Representation



Interval and Ratio Scale

- Pearson's product moment correlation coefficient - measure of linear correlation
  - Varies from -1 to 1
    - -1 - perfect negative correlation
    - 0 - no correlation
    - 1 - perfect positive correlation
  - Use for ratio and interval measurement scales.
  - Not appropriate for nominal and ordinal measurement scales.

Pearson's Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

where
X_i and Y_j = values of Map X and Y respectively
$\overline{X}$ and $\overline{Y}$ = Average of Map X and Map Y respectively
n = number of cells in intersection of two Maps.

Nominal Scale Data
Chi-square statistic

Area Cross-Tabulation Table

Map A (columns), **Map B** (rows)

T_{11}	T_{12}	T_{13}	…	$T_{1.}$
T_{21}	T_{22}	T_{23}	…	$T_{2.}$
T_{31}	T_{32}	T_{32}	…	$T_{3.}$
…	…	…	…	…
$T_{.1}$	$T_{.2}$	$T_{.3}$	…	$T_{..}$

$$T_{ij}^{*} = \frac{T_{i\bullet} * T_{\bullet j}}{T_{\bullet\bullet}}$$

$$\chi^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\left(T_{ij} - T_{ij}^{*}\right)^2}{T_{ij}^{*}}$$

Where
T_{ij}, where there are I = 1, 2, 3, …, N classes on Map B (rows of the table) and j = 1, 2, 3, …, M classes on Map A (columns of the table).
Ti. is the sum of the i^{th} row,
T.j is the sum of the j^{th} column, and
T.. is grand sum over rows and columns.

Proportional-Area Table

	A	$\overline{A}$	
B	0.117 (p_{11})	0.130 (p_{12})	0.247 $(p_{1\bullet})$
$\overline{B}$	0.049 (p_{21})	0.705 (p_{22})	0.753 $(p_{2\bullet})$
	0.116 $(p_{\bullet 1})$	0.830 $(p_{\bullet 2})$	$(p_{\bullet\bullet})$

Probability Tabulation

$$p_{ij} = \frac{T_{ij}}{T_{\bullet\bullet}}$$

$$P\{A\} = p_{\bullet 1}$$

$$P\{B\} = p_{1\bullet}$$

$$P\{A \cap B\} = p_{11}$$

$$P\{A \cap \overline{B}\} = p_{21}$$

$$P\{\overline{A} \cap B\} = p_{21}$$

$$P\{\overline{A} \cap \overline{B}\} = p_{22}$$

Conditional Probability

$$P\{B \mid A\} = \frac{P\{B \cap A\}}{P\{A\}} = \frac{p_{11}}{p_{\bullet 1}} = \frac{T_{11}}{T_{\bullet 1}}$$

$$P\{\text{Granite Till} \mid \text{Granite}\} = \frac{345}{486} = 0.7098$$

$$P\{\text{Granite Till}\} = p_{1\bullet} = \frac{T_{1\bullet}}{T_{\bullet\bullet}} = 0.247$$

Probability and Odds

P = probability

O = odds

$$O = \frac{P}{1-P}$$

P	O	lnO
0.0	0	-∞
.1	1/9	-2.20
.2	1/4	-1.39
.4	2/3	-0.41
.5	1/1	0.00
.6	3/2	0.41
.8	4/1	1.39
.9	9/1	2.20
1.0	∞	∞

Conditional Odds

$$O\{B\} = \frac{P\{B\}}{1-P\{B\}} = \frac{\frac{T_{1\bullet}}{T_{\bullet\bullet}}}{1-\frac{T_{1\bullet}}{T_{\bullet\bullet}}} = \frac{T_{1\bullet}}{T_{\bullet\bullet} - T_{1\bullet}}$$

$$O\{B \mid A\} = \frac{P\{B \mid A\}}{1-P\{B \mid A\}} = \frac{P\{B \mid A\}}{P\{\overline{B} \mid A\}}$$

$$O\{B \mid A\} = \frac{p_{11}/p_{\bullet 1}}{p_{21}/p_{\bullet 1}} = \frac{p_{11}}{p_{21}} = \frac{T_{11}}{T_{21}}$$

Conditional Odds Example

$$O\{GraniteTill\} = \frac{727}{2945-727} = 0.328$$

or 3 to 10

$$O\{GraniteTill \mid Granite\} = \frac{345}{141} = 2.45$$

or 25 to 10

If Granite is present, then the odds of Granite Till also being present is 25 to 10

Odds Ratio - Binary Maps

$$O_R = \frac{O\{B \mid A\}}{O\{B \mid \bar{A}\}} = \frac{T_{11}T_{22}}{T_{12}T_{21}}$$

$$O_R = \frac{345 * 2077}{382 * 141} = 13.3$$

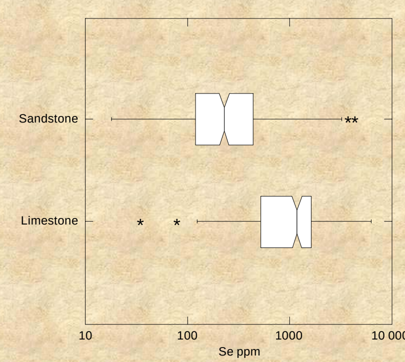
$$O_R = \frac{\text{Measure of Agreement}}{\text{Measure of Disagreement}}$$

Contrast

$$C_W = \ln(O_R)$$

$$C_W = \ln(13.3) = 2.59$$

Mixed Scales - Box Plots



Use Excel chart types Stock and Combination charts to get similar plots.

Bonham-Carter, 1996 Fig. 8-16

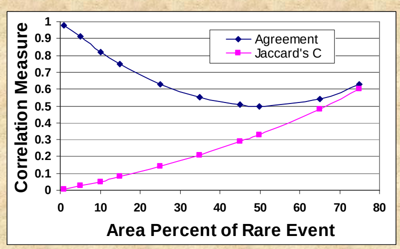
Mixed Scales

	Nominal	Ordinal	Interval/ Ratio
Nominal	Chi-square, O_r, C_w, etc.	Median by nominal class	Mean by nominal class
Ordinal		Rank correlation coefficient	Rank correlation coefficient
Interval/ Ratio			Covariance Correlation coefficient

Categorical Correlation Summary

α and κ	Useful, nice results between –1 and +1.
κ	Where number of classes match, useful for binary and multi-class maps.
O_R and C_W	Useful, comparable results to κ and α and are easy to compute.
C_j	Useful test if positive agreement is more important than negative agreement.
χ^2, C, and U	Avoid for binary maps. Does not distinguish large interactions due to agreement or disagreement.
C_A	Use with care because does not account for chance associations.
Qualification	Choice of counting region (study area) influences the correlation measured.

Correlation of Rare Events



Kappa and Contrast are zero for all cases!

Cross Tabulation

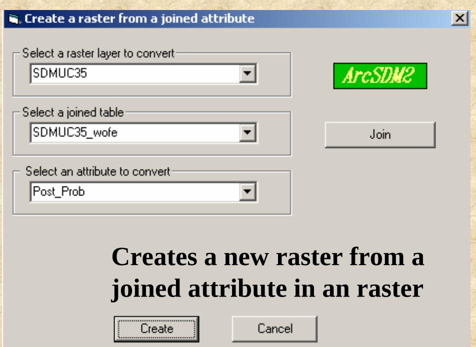
T_{11}	T_{12}	T_{1*}
T_{21}	T_{22}	T_{2*}
T_{*1}	T_{*2}	T_{**}

Agreement ($\frac{(T_{11}+T_{22})}{T_{**}}$) and Jaccard's C ($\frac{T_{22}}{T_{12}+T_{22}+T_{21}}$) from cross tabulation of two random binary grids.

Arc/Info Statistical Tools

- Grid: Autocorrelation tools
  - Correlation - calculates cross correlation
  - Geary and Moran spatial autocorrelation index
- Grid: Multi-variant clustering
  - Isocluster( ) - natural clustering of attributes in attribute space
  - Mlclassify( ) - maximum-likelihood classification in attribute space
  - Princomp ( ) - principal components classification in attribute space
  - Regression - linear or logistic regression coefficients
- Stackstats - standard statistics for a stack of grids

ArcSDM3 Create Raster



Creates a new raster from a joined attribute in an raster

Summary

- Ratio and Interval
  - Pearson's correlation coefficient
- Ordinal
  - Spearman's rank correlation coefficient
- Categorical
  - Several measures. Kappa is very useful as long as have same number of classes.
- Problems when dealing with rare events.

Summary

- Quantitative comparison between two maps can be done several ways!
  - Chap. 8 provides a brief overview and a starting point for further investigation.
- Area tabulation or cross-tabulation table is a fundamental input to most of the correlation measures.

Spatial Analysis in GIS
Single Maps

- Modeling - Pattern Recognition
- Reclassification
- Filtering

Guidelines for Modeling

- Formal statement of the problem.
- Define the user of the model.
- Specification - preprocess the data to provide useful information, that is evidence.
  - Data exploration
    - Reclassification, filtering, transformation, and scaling
  - Reduce the dimensionality by eliminating redundant or correlated information
  - Use the minimum information necessary
- Prediction - combine the evidence to create the model.
- Testing - evaluate the model and it's properties.

Data Exploration

- Process of seeking patterns on maps that help predict spatial phenomena.
  - Visualization leads to recognition of a pattern and the association of the pattern with something of interest.
  - A model is proposed that describes the association.

Data Exploration

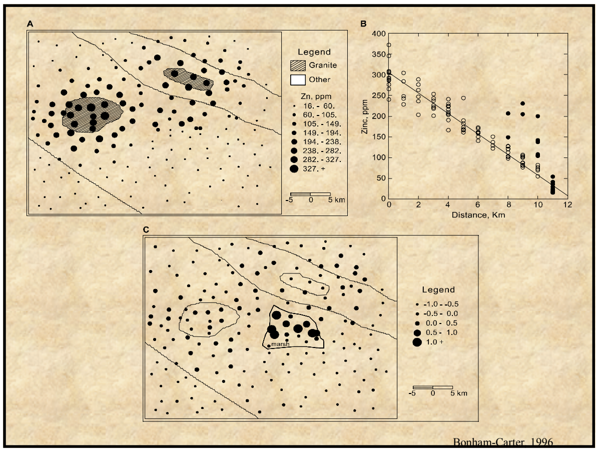
- Seeking patterns involves:
  - Measurement
  - Statistical Summary
  - Visualization
  - Description
  - Understanding of processes causing pattern
- Foundation is data model.

Pattern

- An area having a consistent, recognizable characteristics associated with some object or process.
  - A pattern is something that deviates from the norm.
  - A pattern is associated with a particular scale of observation!
  - It is a primitive.
- Association of patterns and their causes are the bricks of scientific knowledge.

Types of Recognition

- Classification is the process of grouping objects together in classes according to perceived similarities.
- Identification is the recognition of an individual object as a unique singleton class.
- Discrimination is the recognition that an individual object as different from a class.



Bonham-Carter, 1996

Recognition of a Pattern

- Task - Determine what the appropriate level of aggregation and simplification is for the problem at hand, a problem of reclassification.
  - Aggregation and simplification are tied to scale of observation.
  - There is no single scale at which to view a system.
  - Does not mean that all scales serve equally well or there are not scaling laws.
- Description of patterns is the starting point.
- Spatial models start with an assemblage of patterns and associated processes.

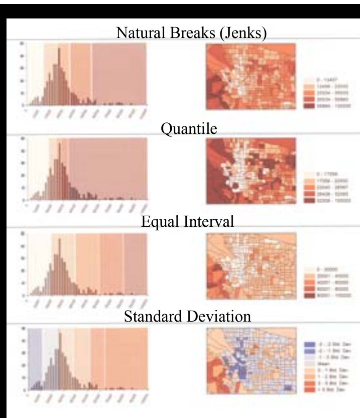
Measurement Scales

- Nominal (Categorical)
  - An unordered label of categories or classes.
- Ordinal (Rank)
  - Measurements ordered (ranked) according to relative position on a scale with unequal intervals between classes.
- Interval
  - Measurements that can be labeled and ordered with an equal interval between classes but without a true zero.
- Ratio
  - Measurements that can be labeled and ordered, with an equal interval between classes, and with a true zero.

Reclassification

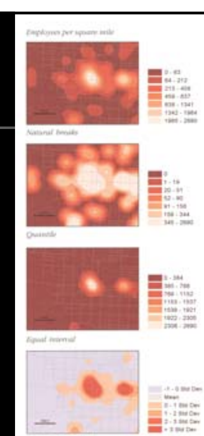
- Reclassification Methods - Continuous measurement scales - definitions
  - Natural breaks
  - Quantile, Equal area
  - Equal intervals
  - Standard deviation
- Semantic Reclassification - Categorical measurement scales

Reclassification in Arcview



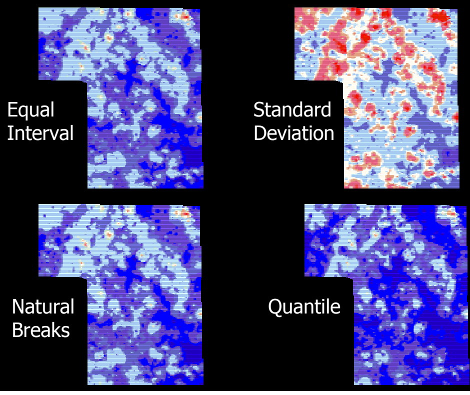
Mitchell, 1999

Reclassification Another Example



Mitchell, 1999

Reclassification – Normal Distribution



Data Transformations

Transform to common range

$$X_i^* = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}$$

Standard Normal Distribution

$$X_i^* = \frac{X_i - \bar{X}}{STD(X)}$$

Skewed Distributions

$$X_i^* = \log(X_i)$$

Floating to Integer Transform

$X_i^* = (X_i + 0.5.AsGrid).\mathrm{int}$ AV3

$X_i^* = \mathrm{int}(X_i + 0.5)$ ArcMap

where

X_i^* is an integer value

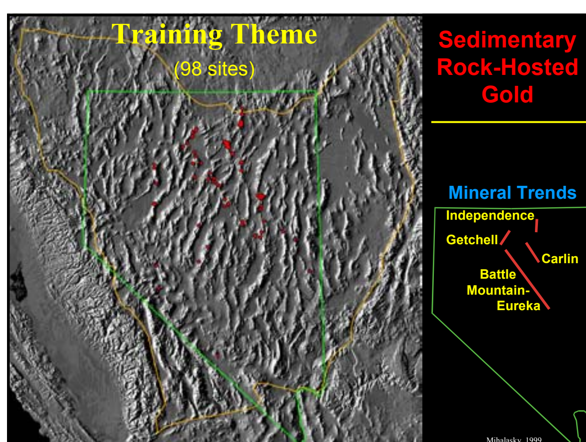
X_i is a floating value

Guidelines Continuous Measurements

- Histograms are essential.
- Quantile, Natural Breaks, and Equal Intervals are least sensitive to frequency distribution.
- Standard deviation are very sensitive to frequency distribution.
- If interested in tails, use standard deviation.
- If interested in middle, use quantile.
- If interested in minimizing class variance, use natural breaks.

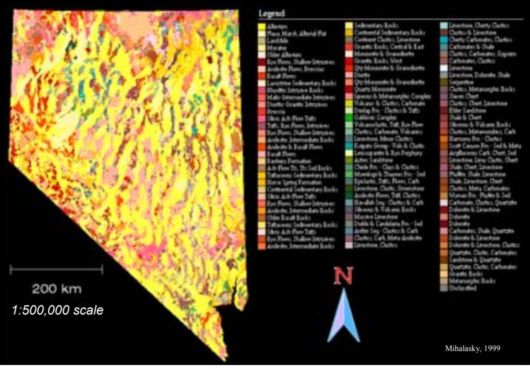
Semantic Reclassification Categorical Measurements

- This is an important problem!
- Expert Systems
  - GeoGen - http://geology.usgs.gov/dm/
- Spatial Association - How to define?
  - Expert decision
  - Measurement such as ArcSDM Contrast



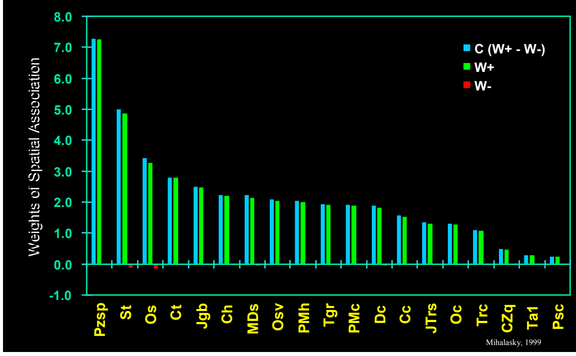
Lithology Evidence Theme

Stewart & Carlson (1978) - 101 Lithologic Units



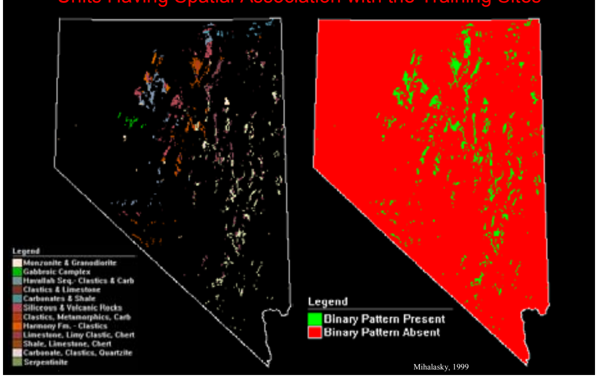
Lithology Evidence Theme

All Units Having a Positive Contrast

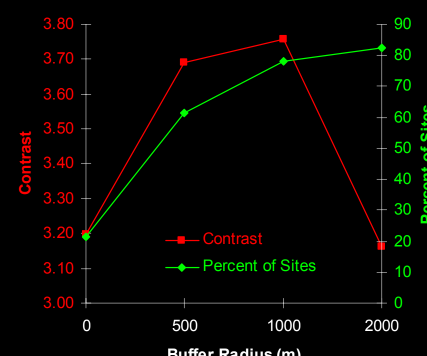


Lithology Predictor Pattern

Units Having Spatial Association with the Training Sites



Near - Proximity To Alteration



USGS

Guidelines and Reclassification Summary

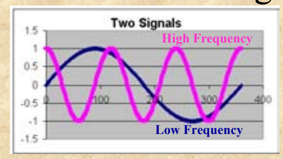
- Concept of a pattern.
- Reclassification of continuous measurement scales.
  - Many tools
- Reclassification of categorical measurement scales.
  - Few tools - current research
  - Expert decision guided by statistics, Contrast and Studentized Contrast

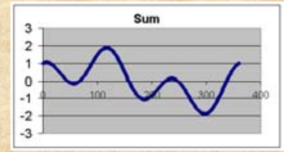
USGS

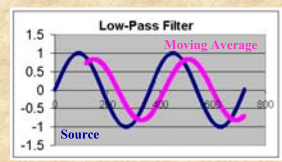
Spatial Analysis in GIS
Single Maps

- Modeling - Pattern Recognition
- Reclassification
- Filtering

Filtering Overview







Moving Average

$$F_{(t+1)} = 1/N \sum_{j=1}^{N} A_{t-j+1}$$

N = Number of prior periods to include in average

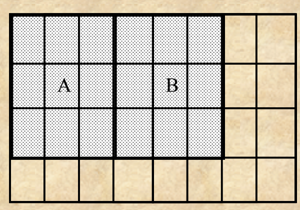
A_j = Actual value at time j

F_j = Forcasted value at time j

Filters

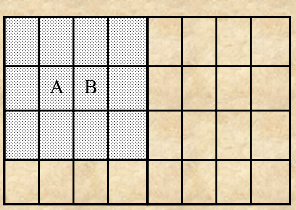
- Interpolate a surface
  - Inverse distance weighting (IDW)
  - Spline
  - Kriging (Geostatistics extension in ArcMap)
- Block statistics and Focal statistics
  - Neighborhood Statistics
- Zonal Statistics
- Hillshade, slope, and aspect
- Convolution Filters

3x3 BlockStats Function



- No overlap of neighborhoods
- All cells in neighborhood receive same value
- A way to decrease the resolution

3x3 FocalStats Function



- Overlapping neighborhoods
- Only the central value receives the new value
- Loose the outside of the theme.

Types of Neighborhoods or Filters

Interval and Ratio Scales

*Mean (Low Pass)

Standard Deviation

Ordinal Scales

*Median (Low Pass)

Nominal Scales

*Majority (Low Pass)

*Variety (Diversity)

Maximum (High Pass?)

Minimum (Low Pass?)

Kernal Properties

Height and Width - 3x3

Type of neighborhood

Weights

1/9	1/9	1/9	
1/9	1/9	1/9	Mean Filter weights
1/9	1/9	1/9	

Others

Minority?

Sum- Program other filters

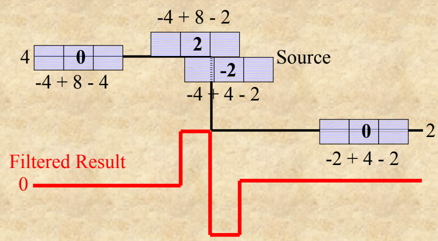
Convolution Filters

0 -1 0 -1 4 -1 0 -1 0	Laplacian weights	
-1 -1 -1 -1 9 -1 -1 -1 -1	High Frequency	
1 1 -1 1 -2 -1 1 1 -1	Directional West	
0.25 0.50 0.25 0.50 1.00 0.50 0.25 0.50 0.25	High Frequency	
-1 -1 -1 1 -2 1 1 1 1	Directional South	

There are a large number of other filters for many applications.

Available as Arcview 3 extensions with problems.

Laplacian Filter



One-Dimensional Laplacian weigths: -1 2 -1

Cascade Programming in AV3

Problem: How do you define the weights?
Neighborhoods can only be defined as including or not including a cell (0,1).

Fragment of Cascading Avenue Code

```
firstLine = {0,1,0}
secondLine = {1,0,1}
thirdLine = {0,1,0}
theKernal = {firstLine,secondLine,thirdLine}
aNbrHood = NbrHood.MakeIrregurlar (theKernal)
theResult = sourceGrid*4.AsGrid -
   sourceGrid.FocalStats(#GRID_STATYPE_SUM,
   theNbrHood, True)
```

Cascade Programming in ArcMap

Problem: How do you define the weights?
Neighborhoods can be defined by two methods:

SetIrregular (weights 0 and 1)

SetWeights (any real or integer value)

SetWeights is most useful to weight individual cells in the filter.

Fragment of VB Code to Define and Apply Filter

```
Dim kernel As Variant
kernel = MakeIrregularKernel
  ' pHood.SetIrregular 3, 3, kernel
pHood.SetWeight 3, 3, kernel
   ' Perform Spatial operation
Dim pOutRaster As IRaster
Set pOutRaster = _
pNeigbOp.FocalStatistics(pGeoDs, _
esriGeoAnalysisStatsSum, pHood, True)
```

VB Code to fill Kernel

```
Private Function MakeIrregularKernel() As Variant
    Dim OutArray() As Long
    Dim X As Long, Y As Long
    X = 3
    Y = 3
    ReDim OutArray(X * Y)
    OutArray(0) = 0
    OutArray(1) = -1
    OutArray(2) = 0
    OutArray(3) = -1
    OutArray(4) = 4
    OutArray(5) = -1
    OutArray(6) = 0
    OutArray(7) = -1
    OutArray(8) = 0
    MakeIrregularKernel = OutArray
End Function
```

Cascade Neighborhoods in AV3

Laplacian 3x3

```
0 1 0
1 0 1
0 1 0
```

Laplacian 9x9

```
0 0 0 1 1 1 0 0 0
0 0 0 1 1 1 0 0 0
0 0 0 1 1 1 0 0 0
1 1 1 0 0 0 1 1 1
1 1 1 0 0 0 1 1 1
1 1 1 0 0 0 1 1 1
0 0 0 1 1 1 0 0 0
0 0 0 1 1 1 0 0 0
0 0 0 1 1 1 0 0 0
```

Odd number of rows and columns!

Cascade Neighborhoods in ArcMap SetWeight Method

Laplacian 3x3

```
0 -1 0
-1 4 -1
0 -1 0
```

Laplacian 9x9

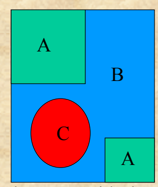
```
0 0 0 -1 -1 -1 0 0 0
0 0 0 -1 -1 -1 0 0 0
0 0 0 -1 -1 -1 0 0 0
-1 -1 -1 4 4 4 -1 -1 -1
-1 -1 -1 4 4 4 -1 -1 -1
-1 -1 -1 4 4 4 -1 -1 -1
0 0 0 -1 -1 -1 0 0 0
0 0 0 -1 -1 -1 0 0 0
0 0 0 -1 -1 -1 0 0 0
```

Odd number of rows and columns!

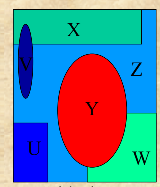
Recursive Filtering

- Often necessary to filter the filtered grid to remove artifacts.
  - For example on the Laplacian, may only want the high and not the low.
  - May wish to eliminate isolated cells.
- Often human interpretation necessary to remove various types of artifacts.

ZonalStats



Shape or Grid Theme
The Zones

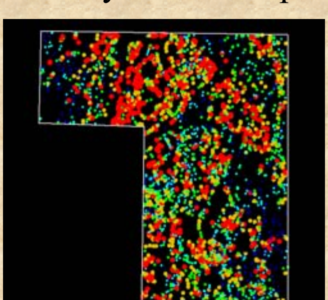


Grid Theme
Measurement to Summarize

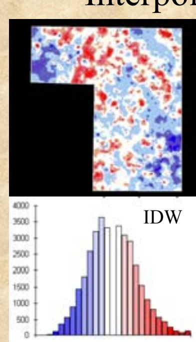
Table from ZonalStats

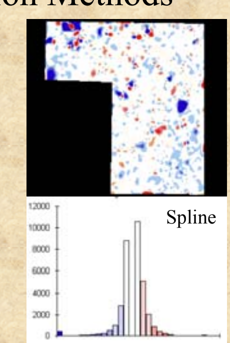
	Mean	STD	Min	Max
A|	n1	n2	n3	n4
B|	n5	n6	n7	n8
C|	n9	n10	n11	n12

Antimony Point Samples



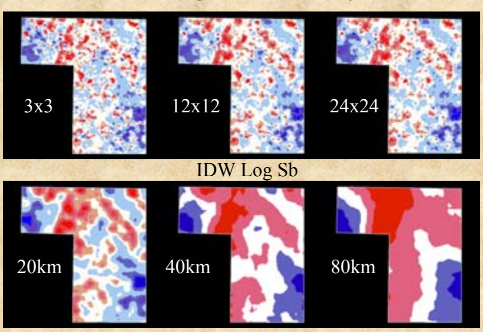
Interpolation Methods





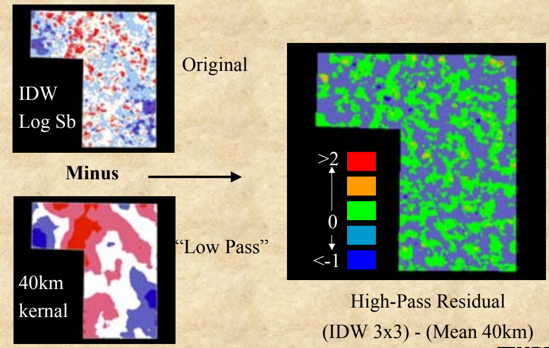
Symbolized by 1/4 standard deviations

Filtering Antimony



Neighborhood mean from 3x3 surface at various kernal sizes

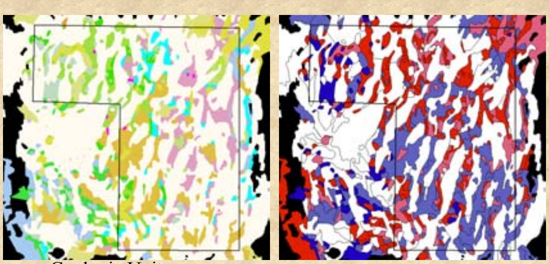
A pattern?



High-Pass Residual

(IDW 3x3) - (Mean 40km)

Zonal Statistics

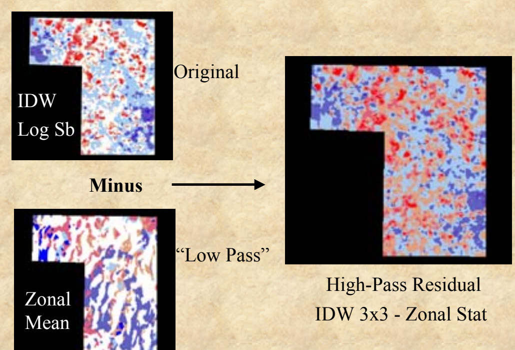


Geologic Units

Central Nevada

Mean Antimony

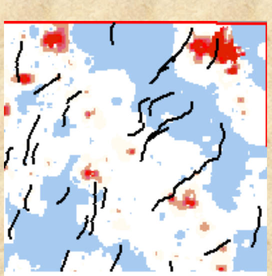
Another Pattern?



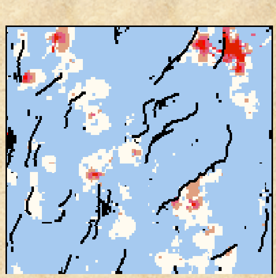
High-Pass Residual

IDW 3x3 - Zonal Stat

Surface with Barriers

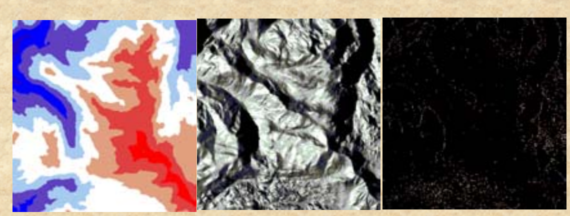


Sb Surface without barriers applied



Sb Surface with barriers applied

3 x 3 Laplacian Filter



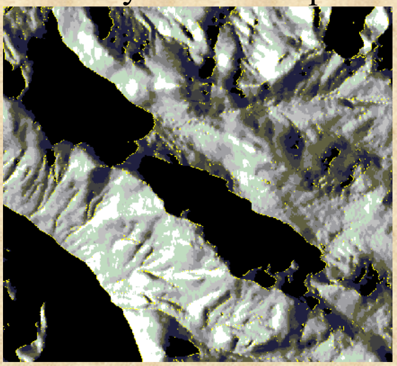
Source: DEM

Shaded Relief

HillShade

Laplacian of Shaded Relief

Recursively Filtered Laplacian



Filtering Summary

- Objective of filtering is to define a pattern that may not be obvious in the original data.
  – Edges of homogeneous areas are often important.
- Filtering is an art!
  – May require recursive filtering or interpretation to remove artifacts.
- Powerful tool for data exploration!

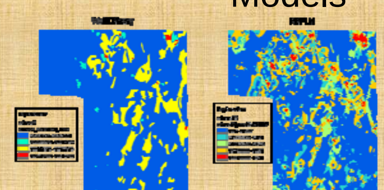
Model Testing

Measures to compare, describe, and validate models
Simple Carlin Model Example

Measures of Models

- Correlation measures to compare models
  - Kappa for ranked models
  - Pearson's for raw models
- Fragstats: Measure the texture or appearance of the model. Does the model look geologic?
- Efficiency of Classification
  - Training sites
  - Not-Training sites: What should "Nots" be?
- Efficiency of Prediction (Validation)
  - Sites not used for training

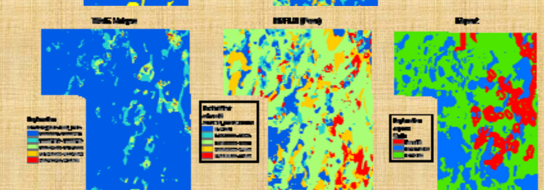
Models



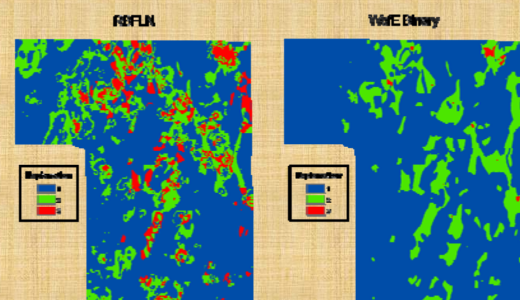
All models using King and Beikman geology and Nure Sb

"Nots" all other mineral deposits in the area.

All models symbolized by natural breaks.



3-Class Models



Correlation between Models

	WofE Binary	*RBFLN (Poor)*	*RBFLN*	*WofE Unique*
WofE Binary	1	0.068	0.29	0.706
RBFLN (Poor)	0.005	1	-0.006	0.132
RBFLN	0.343	0.08	1	0.29
WofE Unique	0.159	-0.01	0.063	1

Pearson's above and Kappa below the diagonal

Kappa: 3 classes by natural breaks

Fragstats: RBFLN-WofE

Index Name	RBFLN	WofE
Number of Patches	593	253
Patch Density	0.0104	0.0045
Largest Patch Index	54.7	62.3
Total Edge	14,398,750	6,117,000
Edge Density	2.5	1.1
Landscape Shape Index	16.2	6.7
Shape Index Mean	1.5115	1.5602
Fractal Dimension Index Mean	1.0404	1.0419
Perimeter-Area Fractal Dimension	1.3236	1.2699
Patch Area Mean	9,570.6	22,442.3
Patch Area-Weighted Mean	1,774,070	2,443,288
Patch Area Median	287.5	300
Patch Area Standard Deviation	129,952	233,087
Patch Area Coefficient of Variation	1357.8	1038.6
Shannon's Diversity Index	0.8	0.5
Simpson's Diversity Index	0.5	0.3
Shannon's Evenness Index	0.7	0.4
Simpson's Evenness Index	0.7	0.4
Largest Patch Index	54.7	80.7
Interspersion Juxtapostion Index	81.0	18.7
Patch Cohesion Index	99.7	99.9
Aggregation Index	96.9	99.0

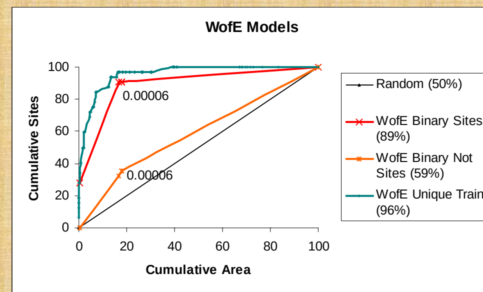
- 3 classes by Natural Breaks.
- RBFLN has larger number of smaller patches (Higher density of patches)
- RBFLN patches have more complex shapes.
- RBFLN patches are more mixed.

PRC: Efficiency of Prediction
SRC: Efficiency of Classification

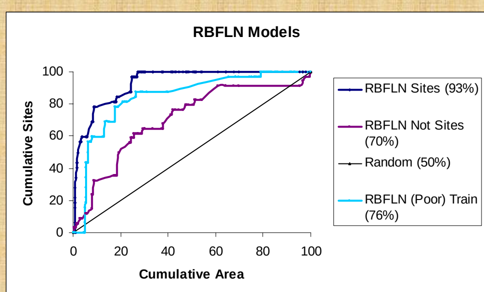
- Intersect points with response grid.
- Frequency of points.
- Join frequency of points with counts in response grid.
- Summation
  - Sort response value descending
  - Cumulative area from high to low response value.
  - Cumulative number of points from high to low response value.
- Plot Cumulative area versus cumulative number of points
- Calculate area under the curve.
  - Area under the curve for sites should be greater than 50% of total area, then have a positive association with points.
  - Area under the curve for "Not" sites should be less than 50% of total area, then have a positive association with points
  - If area under the curve, then have a random association with the evidence. Evidence provides no better information than guessing.
- Point in curve where goes from steep slope to flat slope is an optimal break between predicted sites and not sites.

Chung and Fabbri, 2003, Validation of spatial prediction models for landslide hazard mapping: Natural Hazards, v. 20, p.451-472

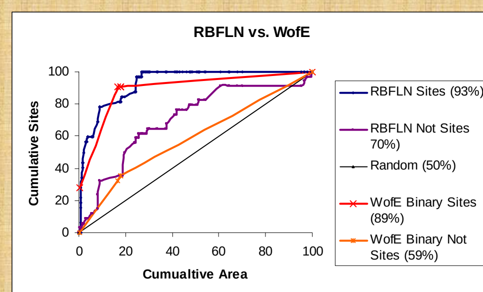
Efficiency of Classification of WofE



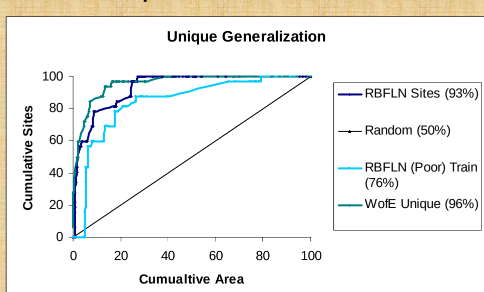
Efficiency of Classification of RBFLN



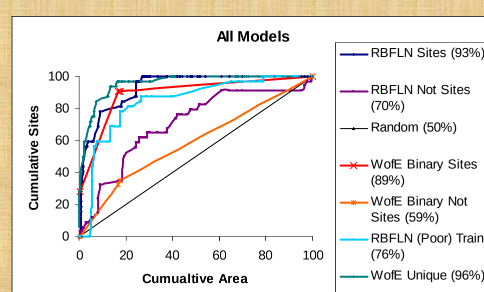
Efficiency of Classification
RBFLN - WofE



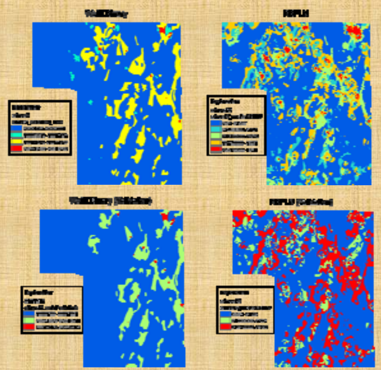
Efficiency of Classification
Unique Generalization



Efficiency of Classification



Validation Models



All models using King and Beikman geology and Nure Sb

"Nots" all other mineral deposits in the area.

All models symbolized by natural breaks.

Used 50% of sites for training and remaining 50% for validation.

Efficiency of Classification Summary

Model	Area Under Curve	Patch Density	Number of Patches	Shannon Diversity Index
WofE Binary Sites	0.89	0.0045	253	0.5
WofE Binary Not Sites	0.59			
WofE Unique Train	0.96			
RBFLN Sites	0.93	0.0104	559	0.8
RBFLN Not Sites	0.70			
RBFLN (Poor) Sites	0.76			
Expert Sites	0.77	0.0062	354	1
Expert Not Sites	0.45			

Correlation: Validation

	WofE Binary	RBFLN	WofE Binary Validation	RBFLN Validation
WofE Binary	1	0.290	0.760	0.186
RBFLN	0.290	1	0.300	0.754
WofE Binary Validation	0.760	0.300	1	0.238
RBFLN Validation	0.186	0.754	0.238	1

ROC Terminology

Processing Steps

- Intersect points with response grid to get probability at points.
- Frequency of points.
- Summations with data sorted from highest to lowest response values.

	Positive	Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Sensitivity = TP/(TP + FN)

TP + FN = Total number of sites

1- Sensitivity = Type II errors (Errors of Omission)

Specificity = TN/(TN+FP)

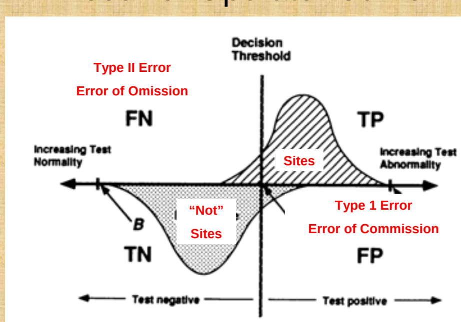
TN + FP = Total number of "Not" or negative sites

1- Specificity = Type I errors (Errors of Commission)

Measures are free from prevalence (rare events) and thresholds.

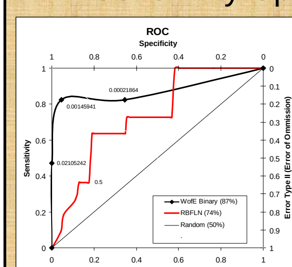
How to define the negative sites ("Nots")?

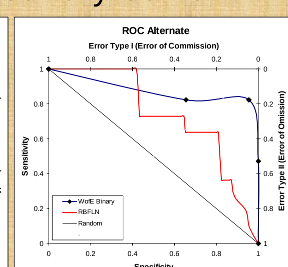
Validation: ROC Receiver Operator Curve



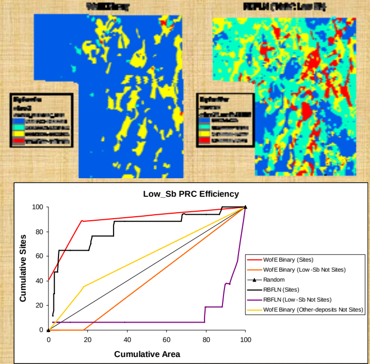
Brismar, 1991, American Roentgen Ray Society: v 157, p. 1119-1121.

Validation Measures: ROC Sensitivity-Specificity Curve

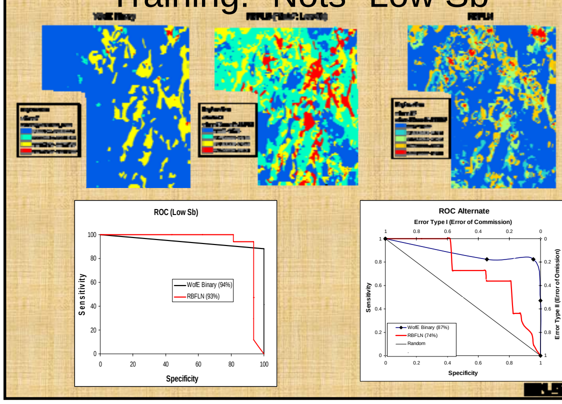




Training: “Nots” Low Sb



Training: “Nots” Low Sb



Validation Summary
WofE versus RBFLN

- Correlation: WofE-validation model (76%) correlates slightly better with WofE than RBFLN-validation (75%) with RBFLN.
  - Insignificant differences. Measure inconclusive.
- FRAGSTATS: WofE is a simpler map.
  - Is this a significant measure?
  - What should these statistics be for a good model?
- SRC & PRC: RBFLN has a higher efficiency of classification of training sites (SRC: 93% to 89%) than WofE and a greater PRC for “Not” sites (PRC: 70% to 59%).
  - A small difference for sites and a significant difference with “Not” sites; suggestive of significance. RBFLN explicitly considered the “Nots” in training.
  - Were the appropriate sites used for “Nots”?
- ROC: WofE has greater efficiencies of prediction (Other-Deposits ROC: 87% to 74%).
  - Question of ROC test because Other-Deposit “Not” sites have PRC values greater than 50% with regards to evidence.
  - Using low Sb as “Nots” gives almost the same validation models (Pearson’s correlation 75-76%) and WofE Validation is the same as the RBFLN Validation (ROC: 94% to 93%).
- Conclusions
  - “Nots” were used to train the RBFLN. This issue with the “Nots” raise questions about the meaning of the RBFLN model?
  - The “Nots” simply further qualifies the meaning of the model. So this RBFLN model may be different than this WofE model.
  - Therefore, if question the “Nots”, then cannot compare these models.
  - Alternatively accept the “Nots”, therefore the WofE model is slightly better than this RBFLN based on multiple ROC curves.
  - Definition of the “Nots” would seem to be a critical consideration to understand and validate a model.

Spatial Modeling in GIS

Summary

Examples of Measurement Scales

Scale Type	Examples	Operations	Means
Nominal	Rock type	=	Mode
Ordinal	Relative age	><	Median
Interval	Temperature	+-*/	Mean
Ratio	Distance	+-*/	Mean

Guidelines for Modeling

- Formal statement of the problem.
- Define the user of the model.
- Specification - preprocess the data to provide useful information, that is evidence.
  - Data exploration
  - Data transformation, filtering, and scaling
  - Reduce the dimensionality by eliminating redundant or correlated information
  - Use the minimum information necessary
- Prediction - combine the evidence to create the model.
- Testing - evaluate the model and it's properties.

Properties of Evidence

- Selected attributes must discriminate between one or more classes of objects.
- Selected attributes should not be correlated with other attributes to any moderately strong extent.
- Selected attributes must have meaning for humans.

Types of Models

- Prescriptive or Deterministic
  - Application of good technical practices
  - Process: Boolean rules, Equations
  - Output: Binary (yes or no), Index overlay (score)
- Predictive
  - Application of mathematics to represent how people think about the evidence but cannot represent as equations.
  - Process: weighting of evidence and combination of weights
  - Output: Favorability, probability, or fuzzy map [0 to 1]

Knowledge Driven Methods

- **Boolean Logic** - True/False representation of maps with all maps rated equally. Simple method with True/False answer.
- **Index Overlay with Binary Maps** - Maps are given different weights. Linear combination of maps.
- **Index Overlay with Multi-Class Maps** - Maps are given different weights as well as the classes of the maps are given different weights. Linear combination of maps.
- **Fuzzy Logic** - More flexible weighting of maps and map classes. Nonlinear combination of maps.
- **Expert Weights of Evidence** – Weighting of evidence easily understood. Log linear combination of maps.

Data Driven Methods

- Weights of Evidence
  - log linear combination of binary or multi-class maps.
  - Classifies areas by probability or favorability of occurrence of a training site.
  - Model parameters easy to understand.
- Logistic Regression
  - log regression combination of binary maps
  - Classifies areas by probability of occurrence of a training site.
  - Model parameters complex.
- Neural networks
  - Experimental, nonlinear combination of fuzzy or map classes
  - Classifies areas by fuzzy membership in training set.
  - Can also be self organizing to produce fuzzy membership.
  - Model parameters complex.

Recognition of a Pattern

- Task - Determine what the appropriate level of aggregation and simplification is for the problem at hand, a problem of reclassification.
  - Aggregation and simplification are tied to scale of observation.
  - There is no single scale at which to view a system.
  - Does not mean that all scales serve equally well or there are not scaling laws.
- Description of patterns is the starting point.
- Spatial models start with an assemblage of patterns and associated processes.

Buffer Resolution

Threshold Weighting Reclassification

Map Scale	Map Resolution	Geologic Resolution	Buffer Resolution
1:2,500,000	1250	2500	5000
1:500,000	250	500	1000
1:250,000	125	250	500
1:100,000	50	100	200

Units - Meters

Map Resolution = (Scale denominator)/2000

Reclassification Summary

- Concept of a pattern.
- Reclassification of continuous measurement scales.
  - Many tools
- Reclassification of categorical measurement scales.
  - Few tools - current research
  - Expert decision

Filtering Summary

- Objective of filtering is to define a pattern that may not be obvious in the original data.
  - For example, edges of homogeneous areas can be important.
- Filtering is an art!
  - May require recursive filtering or interpretation to remove artifacts.
- Powerful tool for data exploration!

Correlation Summary

- Ratio and Interval
  - Pearson’s correlation coefficient
  - Independent of thresholds (reclassification).
- Ordinal
  - Spearman’s rank correlation coefficient
  - Sensitive to thresholds (reclassification).
- Others
  - Kappa for correlation involving rare events
  - Sensitive to thresholds (reclassification).

Testing
Data-driven Methods

- Evaluate classification of training points
- Use points not included in training set to test the model
  - Implementation - use a random subset of training set to develop the weights and use the remainder to evaluate the model.
  - Problem - for many models there may only be a small number of training points to start with.
    - Make a validation model from a subset of training sites and test that validation model is same as model from all training sites.
- Conclusion of testing is often identification of some deficiency in the evidence.
- Field testing of the model.

KISS - Keep It Simple

- Quickly make a simple model based on binary generalization of existing evidence or a neural network model without generalization of existing evidence.
- Test this model to determine what is right and what is wrong with this simple model.
- If the model is reasonably acceptable, refine the model within the time available.
  - Add new evidence
  - Improve evidence: new field work or present in a different way (filtering, reclassification, Boolean or Fuzzy combination of several evidence layers)
  - Rethink the binary generalizations
  - Multi-class generalization

Weights of Evidence – Rules of Thumb

- What is the significance of conditional independence – the big issue in Bayesian methods?
  - If only interested in ranks, not an important issue. Ignore conditional dependency.
  - Can use combination of generalized evidence as a new evidence factor.
  - Can use fuzzy models to combine conditionally dependent evidence as a new evidence factor.
- Binary generalization based on maximum contrast or maximum confidence with acceptable confidence.
- Multiclass generalization based on categorical weights using contrast with acceptable confidence.
- What about generalization based on maximum Studentized contrast or equal weights? Area of on-going research.
- Symbolization by natural breaks gives similar breaks points to breaks on cumulative area vs. posterior probability or efficiency of classification.
- Posterior Probability should be thought of as a measure of favorability of occurrence, a relative ranking.
  - Prior probability is generally taken as defining the neutral point between favorable and unfavorable.

USGS

Neural Networks – Rules of Thumb

- Literature suggests equal number of deposits and not deposits produces better training
  - Basis as a general rule not well tested. May not apply to RBFLN.
- Can always decrease SSE by compressing the evidence, that is fewer unique conditions.
  - Excessive number of unique conditions can lead to noisy response.
- Should be unaffected by conditional dependency. Not proven.
- Fuzzy memberships of training points improves the classification.
- RBFLN
  - Seems unaffected by mix of deposits and not deposits.
  - Training most sensitive to number of RBF.
    - Make small adjustments in number of RBF, parameter M.
    - Then adjust number of iterations. Increasing number of iterations will always decrease SSE and might decrease MSE. Trick is to optimize training so get optimal classification
    - Can test for over fitting by
      - A testing set: Complex to do because of design of software. Maybe a weak test if testing set only tests unique conditions used in models.
      - Optimize the training by finding the optimal classification. Optimal means minimum MSE and SSE
    - Influence of the "Not" sites is not well understood.
  - Symbolize response by natural breaks.

USGS

Neural Networks – Rules of Thumb (Continued)

- PNN
  - Sensitivity to "Not" sites is not clear. May be insensitive to "Nots".
  - Training
    - Adjust distance parameter by small amounts, for example by 0.1.
    - Start by decreasing and if SSE does not decrease then increase parameter.
    - Not terribly sensitive to distance parameter
  - Symbolize the response by quantiles.
- Fuzzy Neural Network
  - Adjust distance parameter by small amounts, for example by 0.1.
  - Two outputs, clusters and membership in clusters.
  - More clusters may represent subtypes, for example of deposits.
  - Symbolize patterns by natural breaks.
- Response themes are all fuzzy membership in favorability of occurrence, a relative ranking.
  - There seem to be scaling problems so neutral point between favorable and unfavorable is not always defined.
  - Fuzzy membership of 0.5 may be good threshold in RBFLN and Fuzzy Neural Network
  - Cannot use fuzzy membership of 0.5 in PNN as a threshold. Considering rare events, might use some small area of high fuzzy membership.

USGS

Fuzzy Logic – Rules of Thumb

- Conditional independence is a consideration for Fuzzy sum, product, and gamma.
  - Best to use conditionally dependent evidence to create a fuzzy factor that utilizes the Fuzzy Or or And.
- A sigma-shaped fuzzification seems to be how people think about evidence.
- Can weight evidence by a multiplier, which must be [0,1].
  - Weighting reflects the importance of the weights.
  - Try to adjust things so 0.5 is neutral.
  - Can use training to define weights (Luo and Dimitrakopoulos, 2003)
- Combining factors is an aggregation process where the combination of factors is more favorable than individual factors.
  - Fuzzy Gamma and Sum are appropriate operators.
    - Optimize Gamma so neutral response is fuzzy membership of 0.5.
- Response themes are all fuzzy membership in favorability of occurrence, a relative ranking.
  - Easier to utilize if tuned so fuzzy membership of 0.5 is neutral between favorable and unfavorable. Tune fuzzification and/or weights.
  - Can symbolize by equal intervals between 0 and 1.

USGS

Evaluation of Models – Rules of Thumb

- Use efficiency measures (SRC, PRC, and ROC) to evaluate models
  - ROC is a stronger test than PRC.
  - PRC is simpler to use because does not require "Nots".
- Symbolization: The big question is how many classes.
  - Breaks in Area vs. Posterior Probability
  - Breaks in slope of efficiency of classification.
  - Backward first derivative defines ranked break points.
- Absolute measures
  - How well classifies the training points, SRC and PRC.
    - A weak test but often all that can be done.
    - Use Brown's probability measure
  - How well classifies points not used in training, ROC.
    - A strong test that can be made with existing data if have appropriate "Not" sites..
- All measures are relative, that is for comparison of different models of the same study area.
  - Rank differences
  - Correlation measures: Spearman's, Pearson's, and Kappa
  - FRAGSTAT – appearance of the response map
  - Efficiency measures

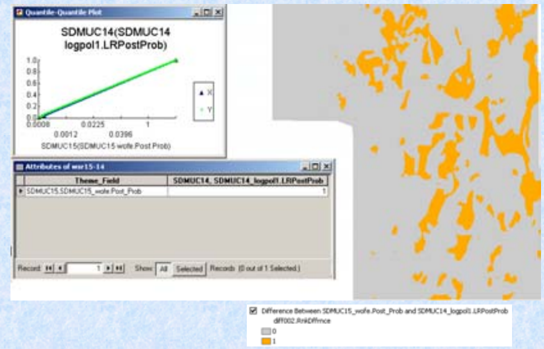
Which Method? – Rules of Thumb

- Have adequate training
  - WofE: Need an understanding of physical process
    - LR: Dealing with conditional dependency problems
      - Can also help define conditionally dependent evidence or highly correlated by zero coefficient
  - RBFLN or PNN Neural Networks: Quick answer
    - Nonlinear classification problems.
- Lack adequate training
  - Fuzzy Logic: Based on how experts think about the problem
    - Address conditional dependency in WofE model.
  - Fuzzy Neural Network: Quick answer
    - Nonlinear classification problems
  - Expert WofE: Model expert thinking in a WofE context.
    - Apply WofE model from one location in another location.
    - Adjust the prior probability to define number of undiscovered deposits? A controversial approach.

Correlation Exercise

Compare Results

Rank Difference



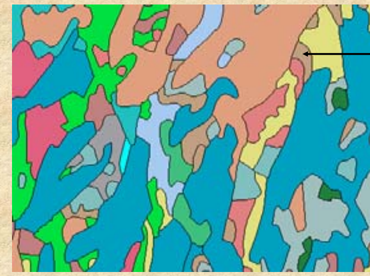
Problem with Reclassify from Real Values

- Seems to be a problem with Reclassify for posterior probability rasters when reclassify by quantile and 3 classes.
- Seems to give more reasonable results if first use Raster Calculator to calculate to an integer and then Reclassify the integer raster.

Overview
of
Fragstats in ArcGIS

http://www.umass.edu/landeco/research/fragstats/fragstats.html
See FRAGSTATS.PDF in reading for an example.

Terminology



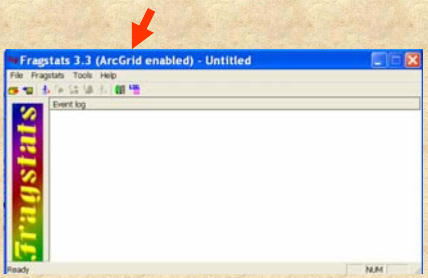
Definitions

Patch – an individual polygon.

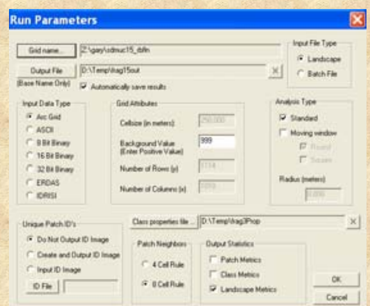
Class – a group of related polygons, such as the green areas.

Landscape – the whole area.

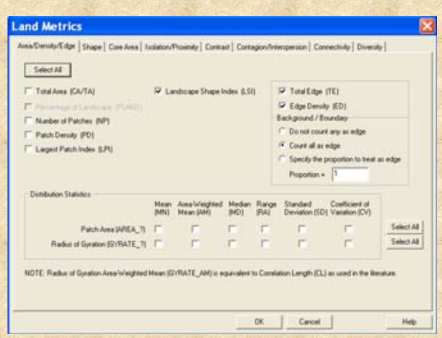
FRAGSTATS Metrics Menu



FRAGSTATS Parameters Menu

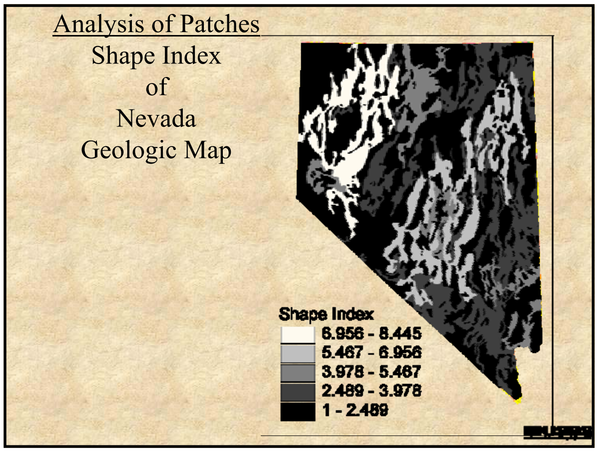


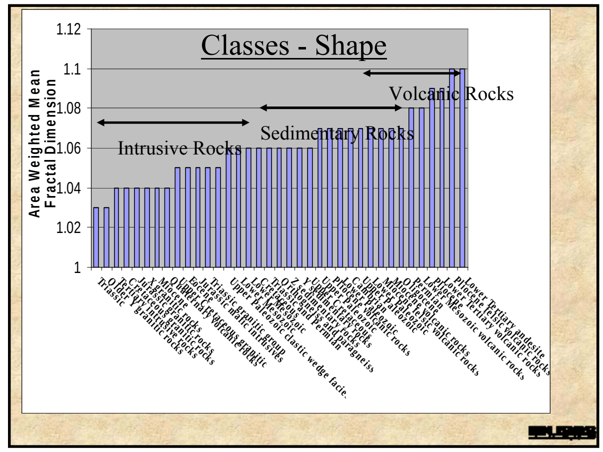
FRAGSTATS Metrics Menu

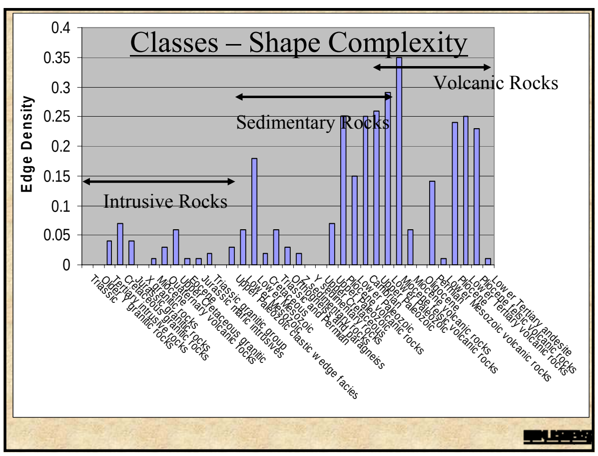


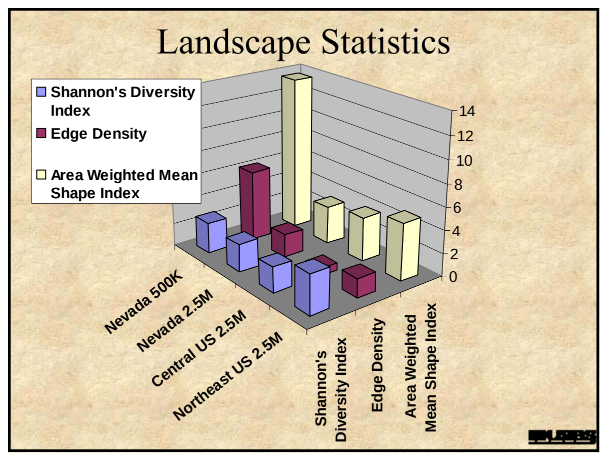
Patch and Class Metrics are subsets of Landscape Metrics.

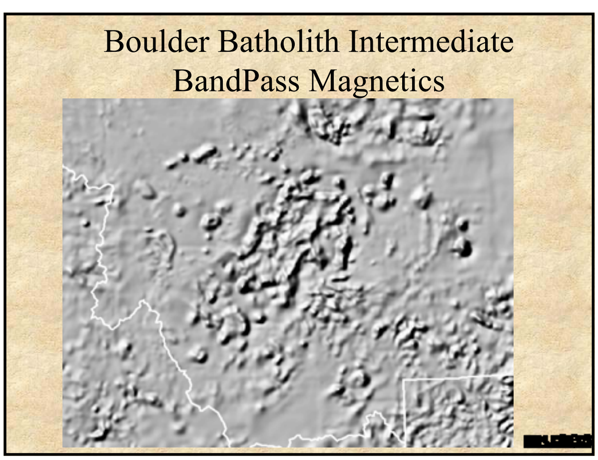
FRAGSTATS Metric	Short Definition
Total Landscape Area	Sum of areas of all patches in the landscape
Number of Patches	Total number of patches in the landscape
Mean Patch Size	Average patch size
Patch Size Standard Deviation	Standard deviation of patch area
Patch Size Coefficient of Variation	Coefficient of variation of patch areas, that is patch size standard deviation divided by mean path size
Total Edge	Sum of perimeter of patches
Edge Density	Amount of edge relative to the landscape area
Mean Shape Index	Shape complexity, equals 1 when all patches are circular (polygons) or square (grids).
Area Weighted Mean Shape Index	Shape complexity weighted by the area of patches.
Mean Patch Fractal Dimension	Shape complexity, equals 1 for shapes with simple perimeters and approaches 2 when shapes are more complex.
Area Weighted Mean Fractal Dimension	Shape complexity weighted by the area of patches
Interspersion Juxtaposition Index	Measure of patch adjacency
Shannon's Diversity Index	Measure of relative patch richness.
Shannon's Evenness Index	Measure of patch distribution and abundance
Total Core Area	Sum of all core areas in the landscape
Core Area Density	Measure of relative distribution of core area (hectares).
Mean Core Area	Average area of disjunct core patches
Core Area Standard Deviation	The standard deviation of disjunct core areas (hectares).
Core Area Coefficient of Variation	The relative number of disjunct core patches relative to the landscape area.
Total Core Area Index	Proportion of core area in the landscape.

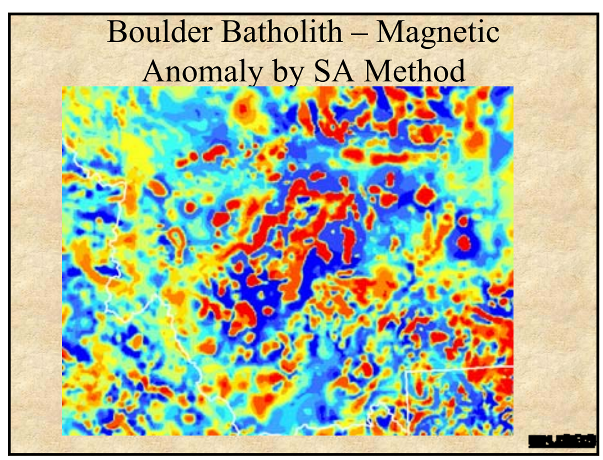

Analysis of Patches
Shape Index
of
Nevada
Geologic Map
Shape Index
6.956 - 8.445
5.467 - 6.956
3.978 - 5.467
2.489 - 3.978
1 - 2.489


Classes - Shape
Area Weighted Mean Fractal Dimension
1.12
1.1
1.08
1.06
1.04
1.02
1
Intrusive Rocks
Sedimentary Rocks
Volcanic Rocks

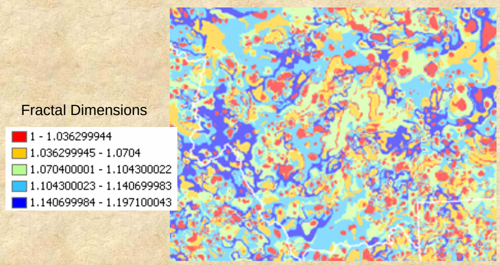

Classes – Shape Complexity
Edge Density
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Intrusive Rocks
Sedimentary Rocks
Volcanic Rocks


Landscape Statistics
Shannon's Diversity Index
Edge Density
Area Weighted Mean Shape Index
14
12
10
8
6
4
2
0
Nevada 500K
Nevada 2.5M
Central US 2.5M
Northeast US 2.5M
Shannon's Diversity Index
Edge Density
Area Weighted Mean Shape Index


Boulder Batholith Intermediate BandPass Magnetics


Boulder Batholith – Magnetic Anomaly by SA Method

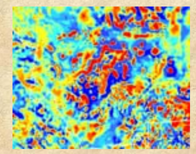
Boulder Batholith – Fractal Dimensions of Magnetic Anomalies



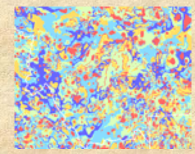
If(Mag Anomaly and Simple Shape) Then Pluton Else Not Pluton

- con([mag_anom] >= 1.879 & [NbrMajor2 of magam15id8_FragMagAM15.patch.Frac2] <= 1.0519,1,0)

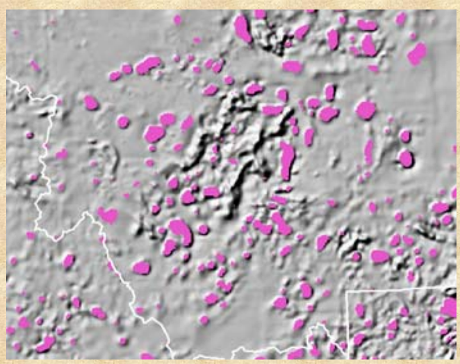
Magnetic Anomaly



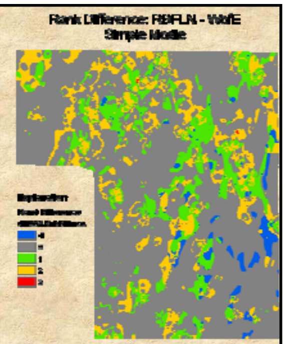
Patch Fractal Dimensions



Boulder Batholith, Plutons



Comparison Carlin RBFLN – WofE 3 classes



0.922 - Spearman 3 Classes, Natural Breaks

0.659 - Band Collection 3 Classes, Natural Breaks

0.290 - Band Collection unclassified

Fragstats: RBFLN-WofE

Index Name	RBFLN	WofE
Number of Patches	593	253
Patch Density	0.0104	0.0045
Largest Patch Index	54.7	62.3
Total Edge	14,398,750	6,117,000
Edge Density	2.5	1.1
Landscape Shape Index	16.2	6.7
Shape Index Mean	1.5115	1.5602
Fractal Dimension Index Mean	1.0404	1.0419
Perimeter-Area Fractal Dimension	1.3236	1.2699
Patch Area Mean	9,570.6	22,442.3
Patch Area-Weighted Mean	1,774,070	2,443,288
Patch Area Median	287.5	300
Patch Area Standard Deviation	129,952	233,087
Patch Area Coefficient of Variation	1357.8	1038.6
Shannon's Diversity Index	0.8	0.5
Simpson's Diversity Index	0.5	0.3
Shannon's Evenness Index	0.7	0.4
Simpson's Evenness Index	0.7	0.4
Largest Patch Index	54.7	80.7
Interspersion Juxtapostion Index	81.0	18.7
Patch Cohesion Index	99.7	99.9
Aggregation Index	96.9	99.0

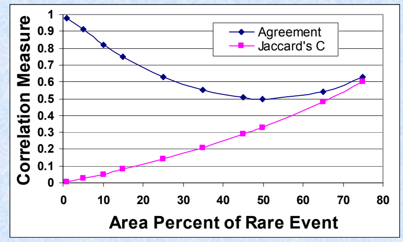
- 3 classes by Natural Breaks.
- RBFLN has larger number of smaller patches (Higher density of patches)
- RBFLN patches have more complex shapes.
- RBFLN patches are more mixed.

Fragstats Summary

- Descriptive tool: Quantifies the texture of a map at various scales.
  - Patches: The pieces of the map
  - Classes; The groupings of the pieces of the map.
  - Landscape: The whole map.
- Analytical tool: Texture measures can give a new presentation of aspects of information in a map.
  - Maps of shape index, etc.

Rare Events

Correlation of Rare Events



Kappa and Contrast are zero for all cases!

Cross Tabulation

T_{11}	T_{12}	T_{1*}
T_{21}	T_{22}	T_{2*}
T_{*1}	T_{*2}	T_{**}

Agreement ($\frac{(T_{11}+T_{22})}{T_{**}}$) and Jaccard's C ($\frac{T_{11}}{T_{12}+T_{11}+T_{21}}$) from cross tabulation of two random binary grids.

Nominal Scale Data

Coefficient of Agreement, kappa

Area-Proportions Cross-Tabulation Table

Map A (columns), **Map B** (rows)

P_{11}	P_{12}	P_{13}	…	$P_{1.}$
P_{21}	P_{22}	P_{23}	…	$P_{2.}$
P_{31}	P_{32}	P_{32}	…	$P_{3.}$
…	…	…	…	…
$P_{.1}$	$P_{.2}$	$P_{.3}$	…	$P_{..}$

Where
$P_{ij} = T_{ij}/T_{..}$
$P_{i.} = T_{i.}/T_{..}$
$P_{.j} = T_{.j}/T_{..}$

$$\kappa = \frac{\sum_{i=1}^{n} p_{ii} - \sum_{i=1}^{n} q_{ii}}{1 - \sum_{i=1}^{n} q_{ii}}$$

where
$q_{ij} = p_{i\bullet} * p_{\bullet j}$
$n =$ number of classes, which is the same in both maps.
Conditional kappa for the i-th class

$$\kappa_i = \frac{p_{ii} - q_{ii}}{p_{i.} - q_{ii}}$$

Calculation for Rare Events

Counts

	A	Not A	
B	0	0	0
Not B	10	990	1000
	10	990	1000

Agreement 0.99
Jaccard's C 0

Proportion

	A	Not A	
B	0	0	0
Not B	0.01	0.99	1
	0.01	0.99	1

Expected Values (Product of marginals)

	Not A	A
B	0	
Not B		0.99

Kappa 0

Counts

	Not A	A	
Not B	990	10	1000
B	0	0	0
	990	10	1000

Agreement 0.99
Jaccard's C 0

Proportion

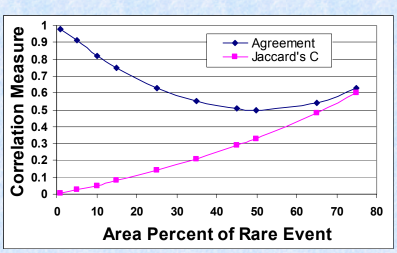
	Not A	A	
Not B	0.99	0.01	1
B	0	0	0
	0.99	0.01	1

Expected Values (Product of marginals)

	Not A	A
Not B	0.99	
B		0

Kappa 0

Correlation of Rare Events



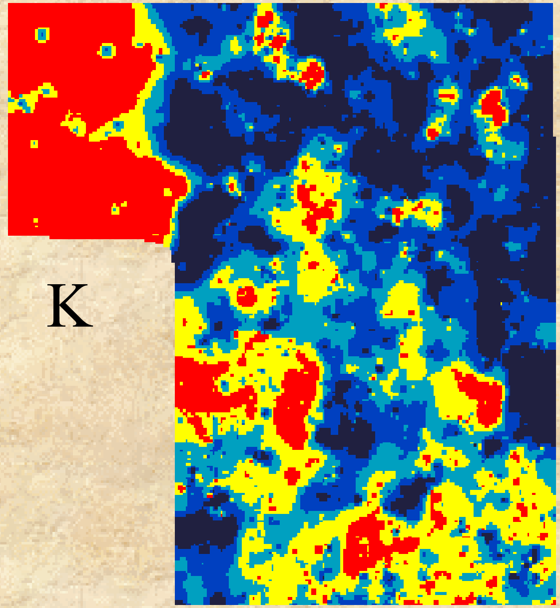
Kappa and Contrast are zero for all cases!

Cross Tabulation

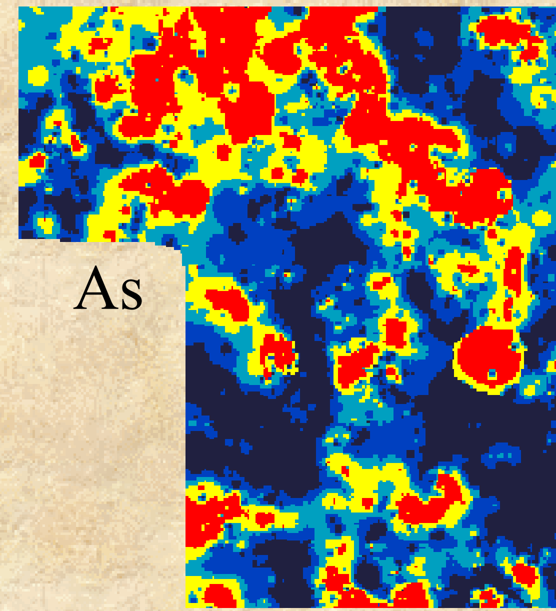
T_{11}	T_{12}	T_{1*}
T_{21}	T_{22}	T_{2*}
T_{*1}	T_{*2}	T_{**}

Agreement ($\frac{(T_{11}+T_{22})}{T_{**}}$) and Jaccard's C ($\frac{T_{11}}{T_{12}+T_{11}+T_{21}}$) from cross tabulation of two random binary grids.

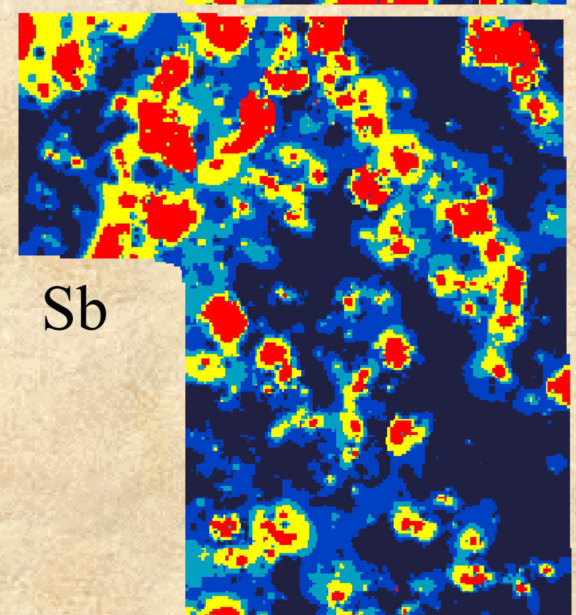
Source Ratio Data



K

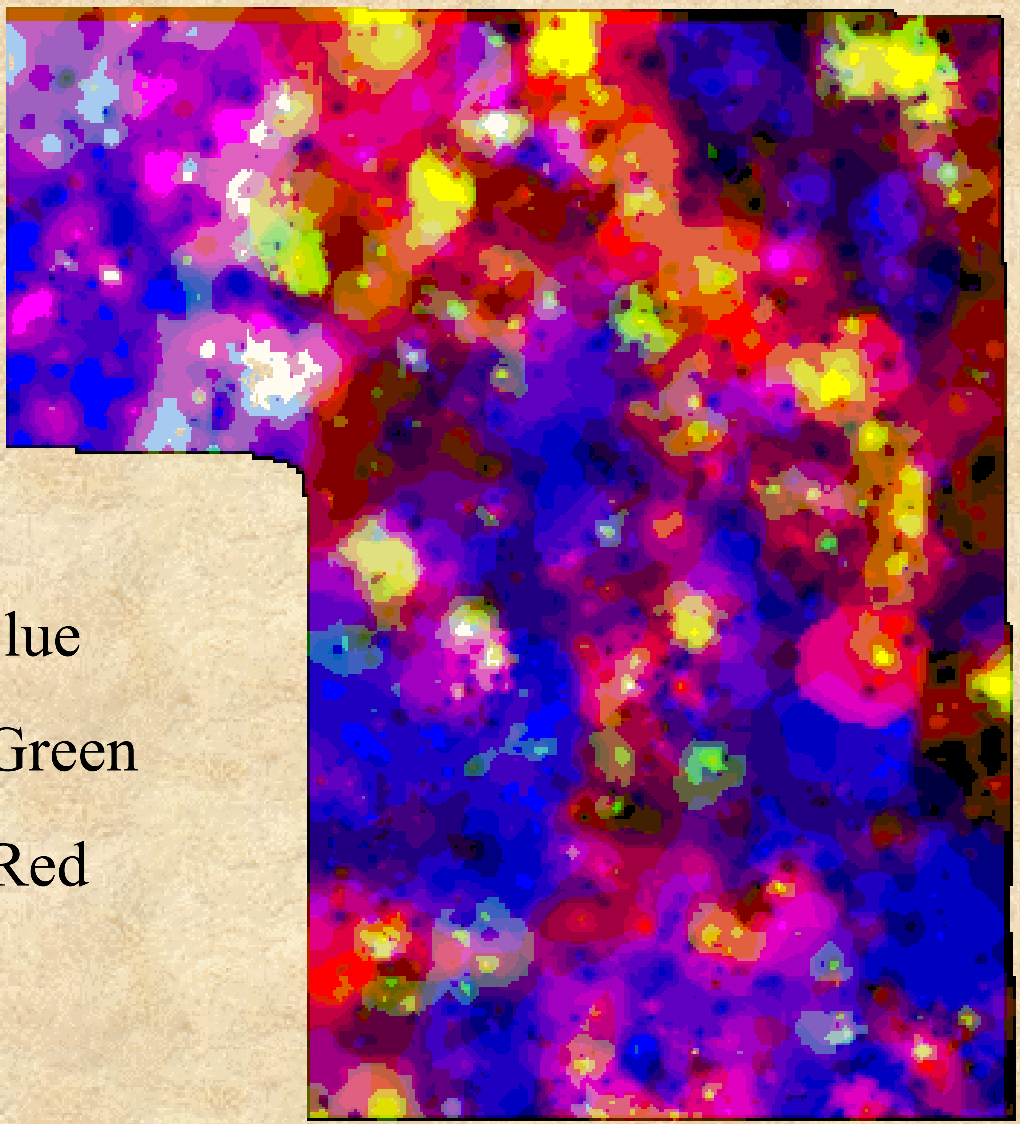


As



Sb

Color Composite



K – Blue

Sb – Green

As – Red

Short Course Version of Lectures

GIS – Insights Through Integration

Graeme Bonham-Carter (GSC)

Gary Raines (USGS)

Mineral Potential Mapping

- Light table origins
- Overlap of anomalies from difference evidence
- Multivariate statistical approach started in the 1960s.
  - Very tedious process to get data in formats that could be used by specially written software.

GIS Catalyst

- Sparked a revolution is spatial data
  - Availability of digital data
  - General purpose software for spatial data analysis
- Ability to deal with
  - High resolution grids
  - Spatial objects in vector form
  - Complex and simple attributes
- Potential for linkage to specialized analytical tools

GIS Preprocessing
"Extraction of Spatial Evidence

- Surfaces from point data
- Extract texture, diversity, derivatives, and other measures
- Reclassify complex data, such as geologic maps, with simple or complex attributes
- Derive contact relationships
- Derive proximity relationships (Buffering)
- Subset of spatial objects (linears by orientation, deposits by types, etc.) using queries of attributes

Modeling Continuum
King and Kramer (1993)

Engineering ↔ Science ↔ Public Policy Decision Making

How to build a bridge at location X?

Which location is best?

Do we need a bridge somewhere?

Deterministic ↔ Probabilistic, Fuzzy, ?

Version 2, Sept. 2001

Models –Simplification of Reality

- Modeling involves application of artificial constructs at many stages
  - The geological map is a model
  - Interpolated surfaces are models
  - The notion of combining evidence from multiple sources using a weighting scheme involves a model (statistical or subjective)

Philosophy of Modeling Data Exploration

- Models must be used but must never be believed. As T.C. Chamberlain said “Science is holding of multiple working hypotheses” (Attributed to Tukey in The Practice of Data Analysis: Essays in Honor of John W. Tukey)
- … models are not destructive; at worst they are ineffectual, and at best, they help to strengthen the quality of the decision making process. (King and Kramer, 1993)

Why Model?

King and Kramer (1993)

- Models are most useful when the right answer is not clear.
- Modeling clarifies the issues of debate in evaluation of an answer.
- Modeling enforces a discipline of analysis, discourse, and consistency.
- Models provide a powerful form of “advice”, that is not “truth”, but a refined result of a particular viewpoint.

Version 1, January 2000

Refined Viewpoint No “Right or “Wrong” Answers

- The models we use can change:
  - Different selections of training points
  - Different choice of evidence
  - Different generalizations of evidence
  - Different weightings and combination method
- We learn by experimenting with the data and investigating spatial associations

Steps in Typical Study

- Assemble digital data
- Extract spatial evidence (preprocessing)
- Prediction analysis by weighting
  - Data Driven
  - Knowledge Driven
- Interpretation and Evaluation

Approaches to Quantitative Mineral Prospectivity Mapping

- Statistical approach (“data driven”)
  - use measured associations between evidential themes and known mineral deposits
    - e.g. regression, neural networks, weights of evidence
- Expert system approach (“knowledge-driven”)
  - use expert exploration knowledge
    - e.g. fuzzy logic, Dempster-Shafer belief functions

Program

<u>Second Day</u>

- Morning
  - Overview
  - Case Studies
  - Weights of Evidence
  - Introduction to ArcSDM (Hands on)
- Afternoon
  - Demonstration of Carlin Model
  - Carlin Model (Hands on)

<u>Third Day</u>

- Morning
  - Carlin Model (Hands on)
  - Demo: Logistic Regression
- Afternoon
  - Fuzzy Logic

Program

Fourth Day

- Morning
  - Demo: Neural Networks
  - Carlin Model (Hands on)
- Afternoon
  - Carlin Model or Your data (Hands on)

Fifth Day

- Morning
  - Hands on wrap up
- Afternoon
  - ArcSDM as geoprocessing tools

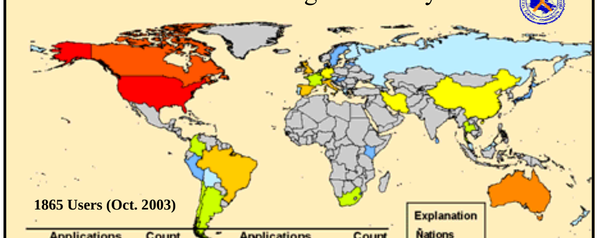
CD-ROM

- ArcSDM3 – Software and documentation
- Handouts – PDF files of handouts for lectures
- Reprints – Useful papers on spatial modeling in PDF format
- Training materials
  - Carlin – Project with data for the Carlin exercise and PDF training file

Spatial Modeling Case Studies

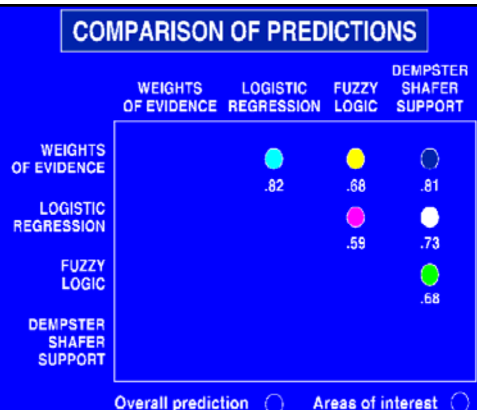
Gary Raines and Graeme Bonham-Carter

ArcSDM Usage Summary 2003



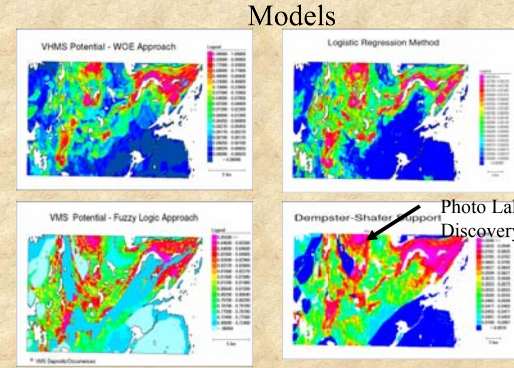
1865 Users (Oct. 2003)

Applications	Count	Applications	Count
Anthropology	3	Hazards/landslides/earthquakes	67
Archeology	133	Hydrology	71
Business applications	15	Land Management	91
Crime prevention	11	Marine Research	16
Ecology/Habitat studies	254	Medical Geography	12
Education	190	Mineral Exploration	236
Epidemiology	46	Mineral Potential Mapping	111
Forestry	46	Not Specified - other	298
Geochemistry	18	Other Geological Application	92
Geology	166	Public health	16
Geophysics	25	Soils	25



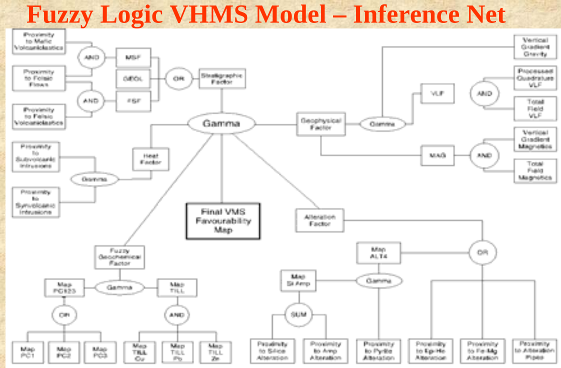
From Wright, 1996

Volcanic-hosted Massive Sulfide Deposit Models



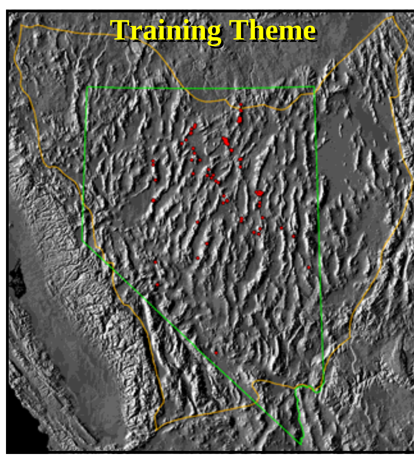
From Wright, 1996

Fuzzy Logic VHMS Model – Inference Net



From Wright, 1996

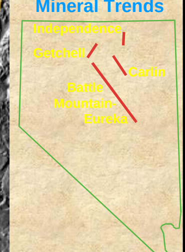
Training Theme



Sedimentary Rock-Hosted Gold

Mihalasky, 1999

Mineral Trends



Evidence Themes

Examined for Potential Inclusion in Modeling

Lithologic
- Lithologic units
- Diversity of lithologic units
- Cenozoic igneous rock unit distance buffers
- Cenozoic igneous rock time-slices
- Cenozoic igneous rock composition-slices
- Mesozoic pluton distance buffers
- Mesozoic pluton density
- Clastic and carbonate rock units
- Clastic and carbonate rock unit distance buffers

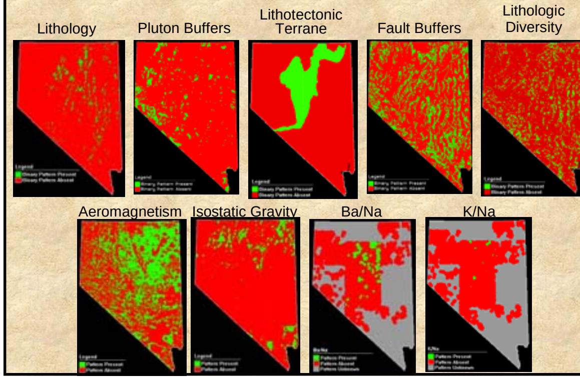
Geochemical (and related)
- K/Na, Ba/Na, U/Th, Fe/Al, La/K, Sc/Fe, Sc/V, K, Al, As (NURE data)
- Igneous rock major element analyses (PETROS)
- Igneous rock radiometric age dates (RADB data)
- Mineralization (metallic deposits) radiometric age dates

Geographic
- Topographic elevation
- Shaded relief of topography
- Degree of topographic slope

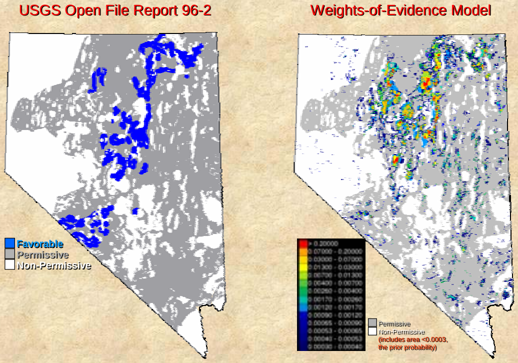
Structural/tectonic
- Cenozoic fault distance buffers
- Cenozoic fault density
- Thrust-front distance buffers
- Strike-slip fault distance buffers
- LANDSAT linear-features distance buffers
- LANDSAT linear-features density
- Highly extended upper crustal terrain
- Deep-seated basement fracture system buffers
- Lithotectonic terrane
- Crustal thickness
- Tertiary rock dip angle and direction
- Late-Paleozoic–early Mesozoic paleothermal anomaly

Geophysical
- Bouguer gravity anomaly
- Isostatic residual gravity anomaly
- Total residual field geomagnetic anomaly
- Geothermal conductivity (gradient)
- Crustal heat flow
- Geothermal wells and springs temperature

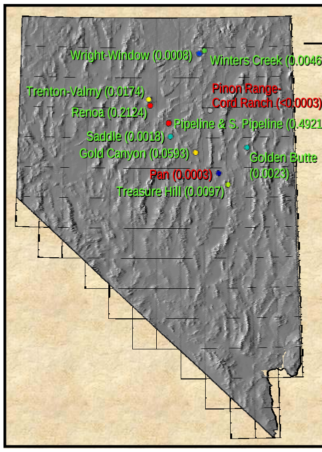
Evidence Themes



Comparison to Expert-Defined Favorable Areas



Agreement with Favorable Area = 92.8%



Blind Test

How well does the model predict *12* newly discovered sedimentary rock-hosted deposits that were not part of the training sites?

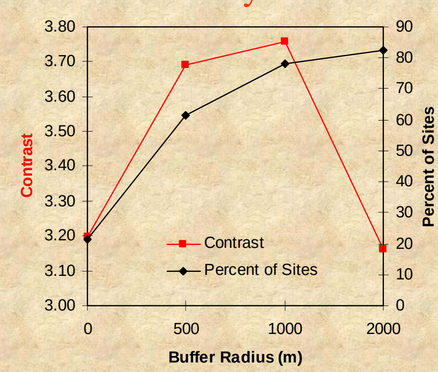
10 of *12* deposits were estimated to have posterior probabilities higher than the prior probability (**0.0003**), where four had relatively high posterior probabilities

Training Sites – Epithermal Gold

Deposit Types	Percentage
Hot Springs Au (25A)	3.2
Comstock Epithermal (25C)	60.6
Volcanogenic U. (25F)	9.3
Epithermal Mn (25G)	5
Hot Spring Hg (27A)	21.9
Summation	100

Raines, 1999

Proximity To Alteration



Evidence

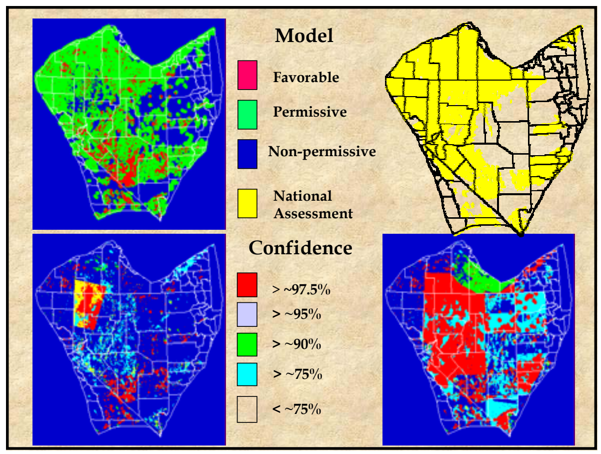
Pattern	*Criteria*	*Source*
Volcanic Rock Proximity	Within 8 km of volcanic rocks less than 34 ma	Nevada state geologic map
Alteration Proximity	Within 1 km of hydrothermal alteration	Western Mining Corp. data
Placer Proximity	Within 1 km of known placer workings	MRDS sites classification
Vent Proximity	Within 2 km of Tertiary vents, shallow intrusives, dome complexes, and other units indicating a volcanic rock source area.	Nevada state geologic map
Fault Proximity	Within 4 km of faults trending NNW to NNE and NW to W	Nevada state geologic map
Anomalous Uranium	NURE equivalent uranium greater than 2 eu	NURE data, Duval's national compilation
Linear Feature Proximity	Within 0.5 km of linear features trending NS and NE to E	Landsat MSS interpretation, Offield, Sawatzky, & Raines
Anomalous Aeromagnetics	NURE aeromagnetics greater than 0 gammas	NURE data, Hildenbrand compilation
Anomalous Geochemistry	Theisen polygons with Ag > 2ppm or As > 5ppm or Mn > 2000ppm or Se > 1.9ppm	NURE stream sediment data, Raines's Great Basin compilation

Summary of Contrast

	Contrast	Studentized Contrast
• Volcanic Rock Proximity	4.901	3.65
• Alteration Proximity	3.756	8.27
• Placer Proximity	3.012	8.375
• Vent Proximity	1.42	10.418
• Fault Proximity	1.317	8.446
• Anomalous uranium	1.253	3.864
• Linear Feature Proximity	1.149	6.453
• Anomalous Aeromagnetics	0.762	6.556
• Anomalous Geochemistry	0.721	5.672

Summary of Weights

	W+	W-
• Volcanic Rock Proximity	0.204	-4.697
• Alteration Proximity	2.331	-1.425
• Placer Proximity	2.989	-0.024
• Vent Proximity	1.247	-0.172
• Fault Proximity	0.338	-0.979
• Anomalous Uranium	0.072	-1.181
• Linear Feature Proximity	1.072	-0.077
• Anomalous Aeromagnetics	0.334	-0.428
• Anomalous Geochemistry	0.346	-0.375



Posters

- Weights of Evidence Solution to Spatial Modeling
- The Problem of Training in Weights of Evidence Compared to Neural Networks
- Demonstration of a Method of Regional Small-Scale Mineral Assessment Based on Geology
- Prediction of Northwest Goshawk Habitat Using Weights of Evidence

What have we learned?

- Results comparable or acceptable to expert's assessment
  - Comparison with US National Assessment
  - Spokane - Epithermal gold and Mississippi Valley deposits
  - Humboldt Assessment
- New discoveries
  - Massive sulfide deposit (Wright and Bonham-Carter)
  - Deposits not in training set are in areas of high posterior probability (Raines and Mihalasky)
  - Packrat model (Mensing and others)
  - Gold deposit in Finland (Nykanen)
  - $10M new investment in exploration in New Zealand based on WofE models (Partington)
- Results are not dependent on mathematics used
- Proximity analysis is powerful data exploration tool
- Conditional Independence problems are most severe in mineral-exploration applications

Weights-of-Evidence Method

- Originally developed as a medical diagnosis system
  - relationships between symptoms and disease evaluated from a large patient database
  - each symptom either present/absent
  - weight for present/weight for absent (W+/W-)
- Apply weighting scheme to new patient
  - add the weights together to get result

Weights of Evidence - WofE

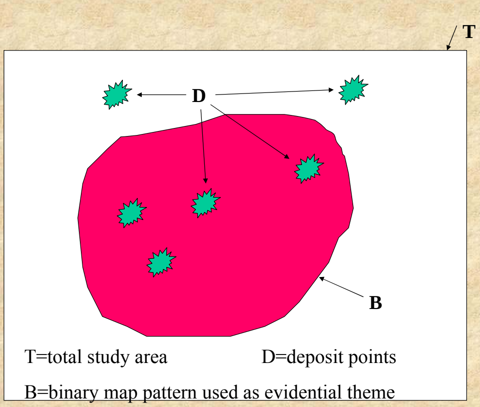
- Data driven technique
  - Requires training sites
- Statistical calculations are used to derive the weights based upon training sites.
- Evidence (maps) are generally reclassified into binary patterns.

Weights-of-Evidence Terms

- Weights for patterns
  - W+ - weight for inside the pattern
  - W- - Weight for outside the pattern
  - 0 - Weights for areas of no data
- Contrast - a measure of the spatial association of pattern with sites
- Studentized Contrast - a measure of the significance of the contrast

Weights of Evidence

- Binary maps to define favorable areas
  - Can use multi-layer patterns
- Measurements
  - Area of study
  - Area of Pattern
  - Number of training sites
  - Number of training sites inside the pattern



Bonham-Carter, personal comm. 2002

Preprocessing

Nominal Measurement Scale

- For example - Geological map
  - select particular stratigraphic units or class
  - generalize by reclassification
  - extract and buffer boundaries between units

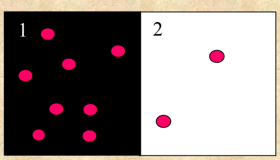
Preprocessing
Continuous Measurement Scale

- Histogram transformations
- Physical properties processing
- Filter
  - separate anomaly/background
- Spatial interpolation (e.g. surfaces, krige)
- Logical combinations (merging, boolean, fuzzy logic)
- Summarize by zonal statistics
  - separate anomaly/background
  - define a residual
  - multivariate analysis
    - principal components analysis and others

Overlay combination

- In vector
  - create polygon overlay and associated PAT
  - create unique conditions overlay and associated PAT
  - Topological selections
- In raster
  - superimpose grids

Application to Binary Evidence

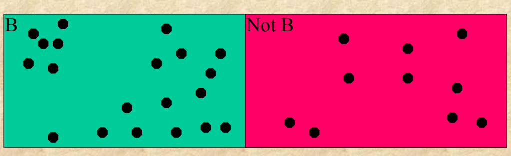


Class	Area	#sites	Relative density	Weight
1	50	8	0.8/0.5=1.6	ln(1.6)= + 0.47
2	50	2	0.2/0.5=0.4	ln(0.4)= - 0.92
Total	100	10		

Expected Values of Weights

- If sites occur randomly,
  - Relative density (RD)=1.0
  - Weight (W) = ln(RD) =0.0
- If sites occur more frequently than chance
  - RD > 1.0, W is positive
- If sites occur less frequently than chance
  - RD < 0.0, W is negative

Example – More Points Than Chance



N(T) = 1000 unit cells (area of study region)

N(B) = 500 unit cells (area of theme B present)

N(B&D) = 20 (count of number of training points on B)

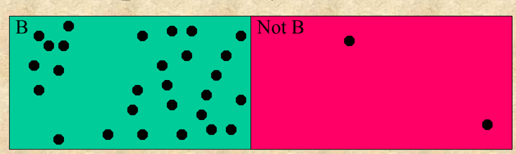
N(D) = 30 (count of total number of training points)

$W^+ = 0.2980$ $W^- = -0.4157$ C = 0.7138

More points on theme than would be expected due to chance

Bonham-Carter, personal comm. 2002

Example – Many More Points



N(T) = 1000 unit cells (area of study region)

N(B) = 500 unit cells (area of theme B present)

N(B&D) = 28 (count of number of training points on B)

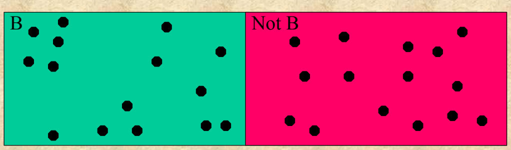
N(D) = 30 (count of total number of training points)

$W^+ = 0.6513$ $W^- = -2.0414$ C = 2.6927

Many more points on theme than would be expected due to chance

Bonham-Carter, personal comm. 2002

Example – Equal Pattern and Points



N(T) = 1000 unit cells (area of study region)

N(B) = 500 unit cells (area of theme B present)

N(B&D) = 15 (count of number of training points on B)

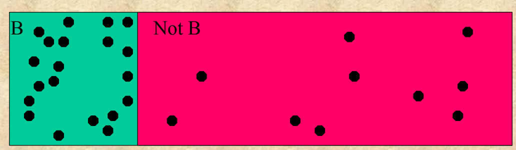
N(D) = 30 (count of total number of training points)

W^+ = 0.0 W^- = -0.0 C = 0.0

Number of points on theme equals that expected due to chance

Bonham-Carter, personal comm. 2002

Example – Small Pattern and Many Points



N(T) = 1000 unit cells (area of study region)

N(B) = 250 unit cells (area of theme B present)

N(B&D) = 20 (count of number of training points on B)

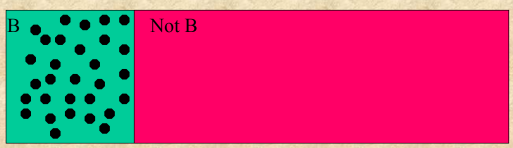
N(D) = 30 (count of total number of training points)

W^+ = 1.0338 W^- = -0.8280 C = 1.8617

Many more points on theme than would be expected due to chance

Bonham-Carter, personal comm. 2002

Example - Weights Undefined



N(T) = 1000 unit cells (area of study region)

N(B) = 250 unit cells (area of theme B present)

N(B&D) = 30 (count of number of training points on B)

N(D) = 30 (count of total number of training points)

W^+ = inf W^- = -inf C = inf

Undefined: practical solution--assign fraction of point to (not B)

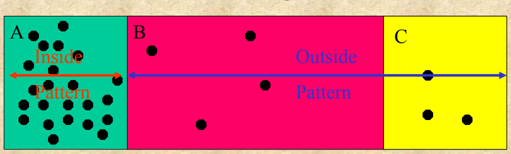
Bonham-Carter, personal comm. 2002

Multi-class Themes

- Maps (themes) with unordered classes (categorical) e.g. geological map. Calculate weights for each class and then group classes (reclassify) as needed.
- Maps (themes) with ordered classes (contour maps e.g. geochemical or geophysical field variables). Usually calculate weights based on successive contour levels, cumulatively. Then reclassify.

Bonham-Carter, personal comm. 2002

Multi-class – Categorical Classes



N(T) = 1000 unit cells (area of study region)

N(A) = 250, N(B) = 500, N(C) = 250,

N(A&D) = 23, N(B&D) = 4, N(C&D) = 3,

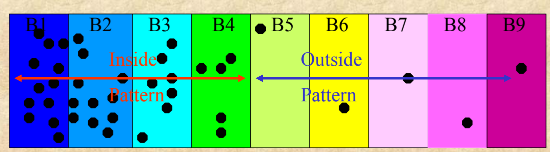
N(D) = 30 (count of total number of training points)

W1 = 1.1866 W2 = -1.3442 W3 =-0.9347 C_{max} =2.5308

Three classes, e.g. rock types (categorical scale of measurement)

Bonham-Carter, personal comm. 2002

Ordered Classes - Cumulative

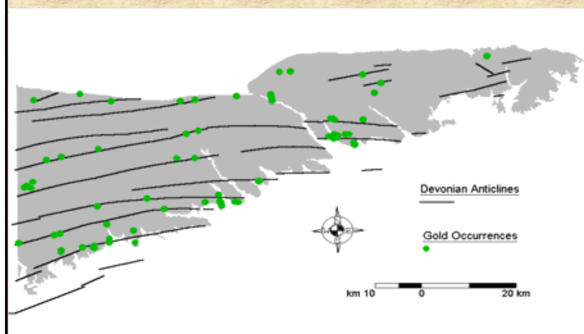


$N(B_i)$	100	100	100	100	100	100	100	100	100
Cum	100	200	300	400	500	600	700	800	900
N(D)	12	11	7	5	1	1	1	1	1
Cum	12	23	30	35	36	37	38	39	40
W+	1.08	1.03	0.87	0.72	0.51	0.35	0.21	0.10	--
W-	-0.25	-0.63	-1.01	-1.53	-1.53	-1.53	-1.53	-1.53	--
C	1.33	1.66	1.88	2.25	2.04	1.88	1.74	1.64	--

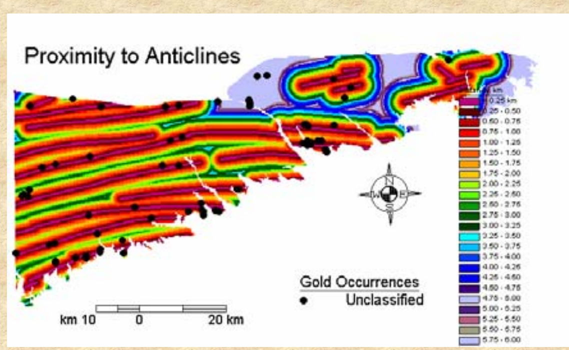
Bonham-Carter, personal comm. 2002

Weights Calculations

- Choose a small unit cell – affects the prior probability but only a little on the weights
- Can have multi-class maps but often not enough training points to get stable weights.
  - Use Studentized contrast to evaluate stability of weights.
- Contrast can be used to define optimal thresholds.
  - Use Studentized contrast to evaluate stability of contrast.
- See Bonham-Carter, Agterberg, and Wright (1988) for equations (WofE_NovaScotia.pdf)

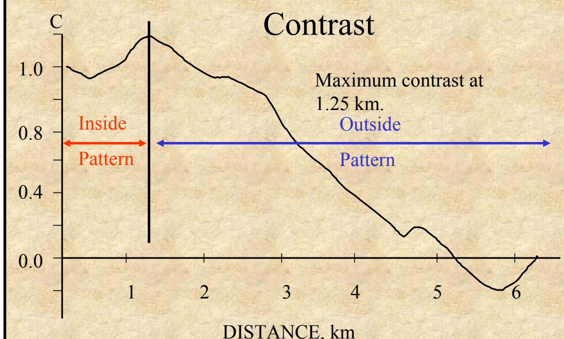


Bonham-Carter, 1996

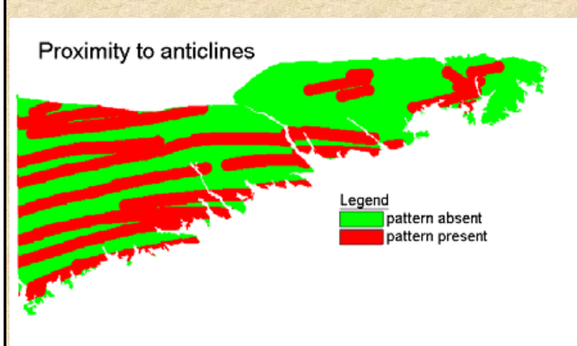


Bonham-Carter, 1996

Contrast

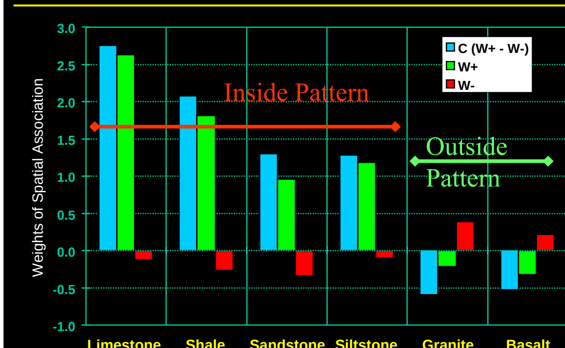


Bonham-Carter, 1996



Bonham-Carter, 1996

Categorical-Weighting Reclassification



Modified from Mihalasky, 1999

Handling Uncertainty

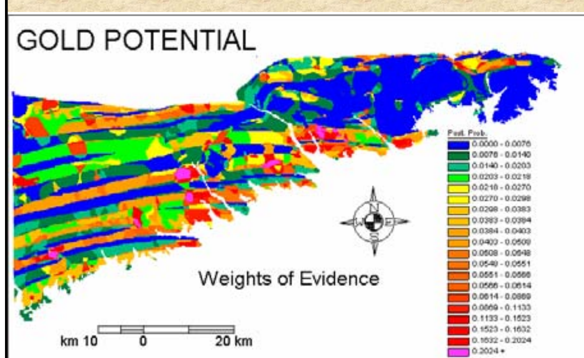
- Uncertainty due to weights – variance of weights.
- Uncertainty due to missing data – estimate of variance due to missing data
- Other measures of uncertainty?
- For Response Map can combine the various uncertainty measures to obtain a total variance.
- Studentized posterior probability (PP/s(PP)) can provide a useful measure of confidence.

More Evidence

- Most gold deposits occur close to anticlines
- Generate map showing distance to anticlines
- How many intervals? The robustness of weight estimates inversely proportional to number of intervals
- Can explore relationship of contrast for binary interval and "optimize" cutoff

Summary of Weights

		WEIGHTS		C
• Geology	0.31	-1.24	-1.74	2.05
• Lake sed geochem		1.42	-0.38	1.80
• Anticlines		0.56	-0.83	1.39
• Au in vegetation		0.84	-0.29	1.13
• Geol contact(1)		0.37	-0.27	0.64
• Geol contact (2)		0.22	-0.04	0.26
• NW lineaments		0.04	-0.01	0.05



Bonham-Carter, 1996

Decisions for Weights of Evidence

- Define the study area
- Define the training set
- Select confidence level for contrast
- Select the evidential maps
  - Use Contrast and Studentized Contrast to evaluate.
  - Reclassification (Binary or Multi-class)
  - Thresholds maximum, minimum, or grouping of nominal classes
- These decisions define objective, binary reclassification
  - Needed measurements: Area of study, Area of the pattern, Number of training sites, Number of training sites inside the pattern

Weights of Evidence

- Advantages
  - Objective assignment of weights
  - Multiple patterns combined simply
  - Reclassification to optimize contrast gives insights into spatial relationships
  - Deals with missing data
  - Measures aspects of uncertainty that can be mapped
- Disadvantages
  - Assumption of conditional independence
  - Requires a training set of sufficient size.

Logistic Regression Method

Graeme Bonham-Carter

Bonham-Carter, 1999

Introduction

- "Data-driven" method applicable where training set of mineral sites is available
- The response variable is dichotomous (binary), e.g. presence/absence of mineral site
- The explanatory variables (evidential themes) are ordered or dichotomous (not multi-class categorical).

Bonham-Carter, 1999

In ordinary regression, the response variable is continuous, unbounded and measured on an interval or ratio scale

In situations where the response variable is binary (present/absent) this causes a problem, because the predicted response must be in the interval [0,1].

The response variable can be assumed to be $P(Y=1)$, from which we also know $P(Y=0)=1-P(Y=1)$

Bonham-Carter, 1999

The solution to the problem of forcing the response variable to be in the range [0,1] is to use the logit transform.

Logits = natural logs of odds

Odds = Probability/(1-Probability)

$$\text{Logit}(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

Where the b's are unknown coefficients and the X's are the explanatory variables

Bonham-Carter, 1999

Logistic Regression Vs. Weights of Evidence

$$\text{Logit}(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

(simultaneous solution of b's)

$$\text{Logit}(Y) = \text{Prior Logit} + W_1 + W_2 + W_3 + \ldots + W_k$$

(solution for W's theme by theme, not simultaneous)

Note that the b_0 term in LR is comparable to the prior logit in WofE, and the b's are comparable to the W's. However, instead of 1 coefficient, there are 2 (or more) weights, depending on the number of classes. Therefore, the b's are more comparable to the contrast values

Bonham-Carter, 1999

Solution to Logistic Regression Equation

- The coefficients cannot be solved by ordinary least squares (a direct matrix inversion), because the equation is non-linear
- The method of maximum likelihood is used to maximize the value of a log-likelihood function
  - This requires an iterative solution
- So coefficients are obtained simultaneously without an assumption of conditional independence.

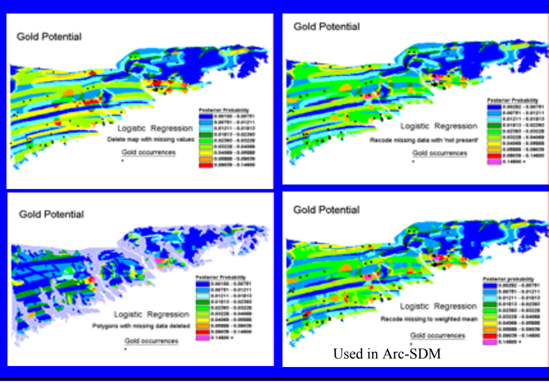
Practicalities

- Can calculate the logistic regression coefficients using the same unique conditions table as for WofE
  - Muti-class themes must be split into binary themes in unique conditions table.
- In ArcSDM deal with missing data and multi-class problem automatically.
- In Arc/Info does not deal with missing data and has another input format.

Problem of Missing Data

- Deleting all unique conditions with missing values in any of the evidential themes.
- Deleting themes that have missing data totally.
- Replacing missing values with zero, or some other constant.
- Replacing missing values with an expected value, e.g. area weighted mean

"Missing Data" Approaches



Can then compare the results from weights of evidence to logistic regression

This is then a check on the effect of conditional dependence on the results of weights of evidence, although if missing data and multi-class categorical evidential themes have been used, then one cannot be absolutely sure what effect the recoding in logistic regression has on the results.

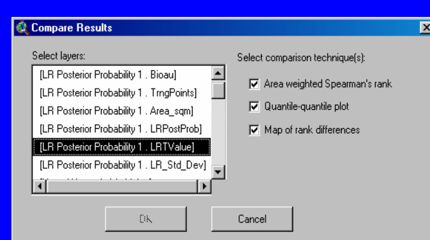
Bonham-Carter, 1999

Compare Results

- ArcSDM includes three techniques for comparing the results of different techniques:
  - Spearman's Area Weighted Rank Correlation
  - Quantile-quantile plot
  - Map of rank differences

Bonham-Carter, 1999

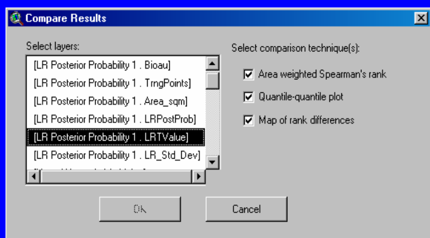
Compare Results



Bonham-Carter, 1999

Compare Results

- Possible inputs:
  - integer grid theme with numeric field(s)
  - floating point grid theme



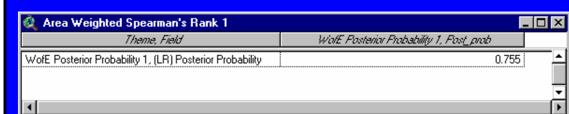
Bonham-Carter, 1999

Compare Results
Spearman's Rank Correlation and Rank Mapping

- Arcview 3 - Classifies both variables into 20 quantiles (ranks). ArcGIS – user specifies number of ranks
- Spearman's Area Weighted Rank Correlation is calculated and written to a dBase file
- Map of rank differences generates a difference map, classifies and symbolizes it to show where the two input evidential themes are similar or dissimilar

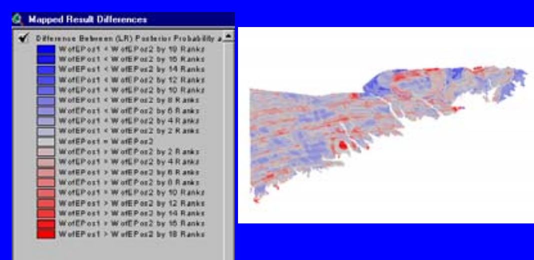
Bonham-Carter, 1999

Compare Results
Spearman's Rank Correlation



Bonham-Carter, 1999

Compare Results
Map of Rank Differences



ArcGIS – User has to symbolize and specify number of classes

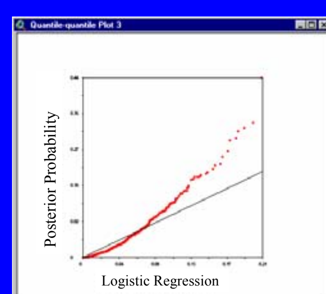
Bonham-Carter, 1999

Compare Results
Quantile-quantile plot

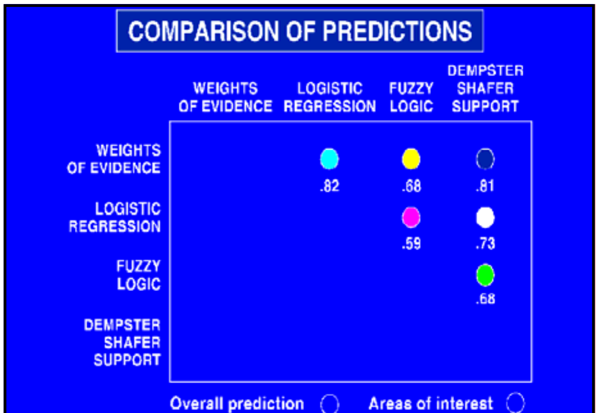
- Sorts the values in each field or theme in ascending order
- if one variable has more observations than the other (for Arcview3), its values are interpolated so that there are equal number of values. ArcGIS: specify number of classes
- values are plotted as x and y coordinates

Bonham-Carter, 1999

Compare Results
Quantile-quantile Plot



Bonham-Carter, 1999



From Wright, 1996

SUMMARY

- Logistic regression can be compared to weights of evidence to check CI assumption
- The total expected number of deposits is usually slightly underestimated by LR (rounding?)
- In general the results of the two methods are similar in terms of ranks, except the WofE probabilities are usually higher than LR probabilities because of CI

Bonham-Carter, 1999

SUMMARY (2)

- ArcSDM will generate LR automatically (expanding the UC table for categorical themes and substituting area-weighted mean values for missing data) at the same time as running WofE, if desired
- Tools for comparing maps are provided

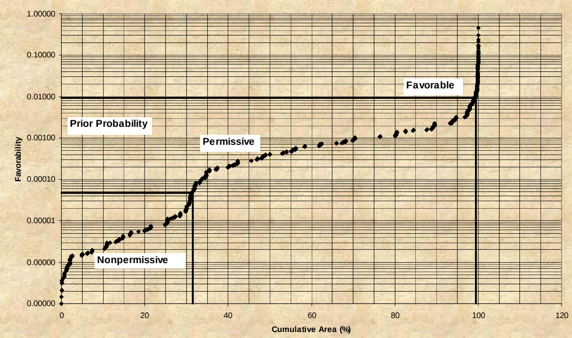
Bonham-Carter, 1999

Carlin Demo

Gary Raines

Sept. 2001

Posterior Probability Classes



Student T Values

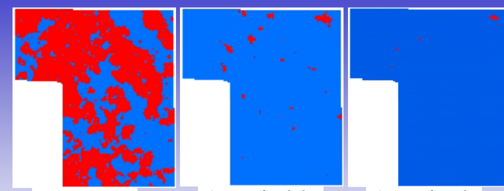
Confidence	T Value
99.5%	2.576
99%	2.326
97.5%	1.96
95%	1.645
90%	1.282
80%	0.842
70%	0.542
60%	0.253

Research Topics and New Tools

Gary Raines and Graeme Bonham-Carter

Winter 2005

Generalization of Evidence



Binary map of quantile classification

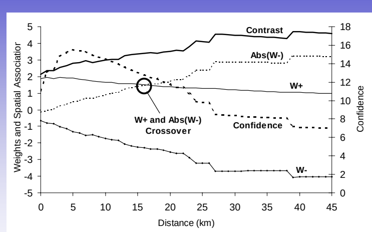
Binary map of standard deviation classification

Binary map of natural breaks classification

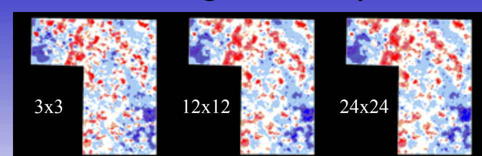
Classification Method	Sb Conc (ppm)	Contrast	W+	W-	Stud C	% Area
Quantile	3.40	3.1693	0.5015	-2.6678	3.1236	58.8600
Std Deviation	33.60	3.2155	2.8950	-0.3205	8.5593	1.6000
Equal Interval	110.60	3.8993	3.7805	-0.1188	7.2529	0.2680
Raw Antimony	121.00	4.0532	3.9341	-0.1191	7.5247	0.2300
Natural Breaks	145.10	4.0347	3.9468	-0.0880	6.5866	0.1700

Gene Lohrmeyer, written communication, March 2004

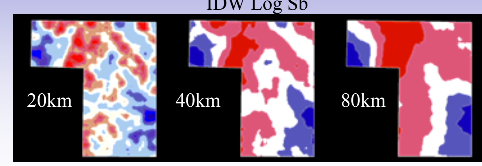
Reclassification of Evidence



Filtering Antimony



IDW Log Sb

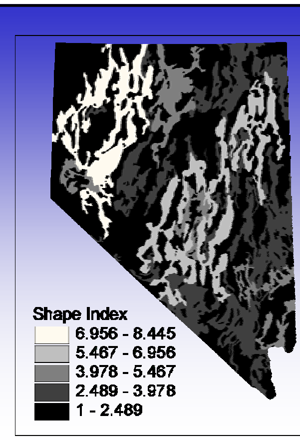


Neighborhood mean from 3x3 surface at various kernal sizes

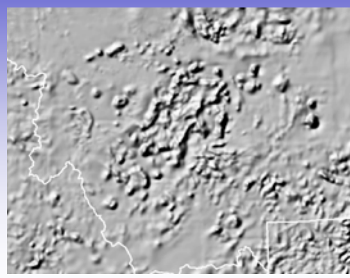
Evidence from Geologic Maps

- Lithology
  - Map Unit
    - Rock Type
    - Age
  - Proximity to map unit
  - Lithodiversity
  - Lithotectonic package
- Structure
  - Orientation
  - Density
  - Proximity
  - Intersection
  - Sinuosity
- Shape
- Topology

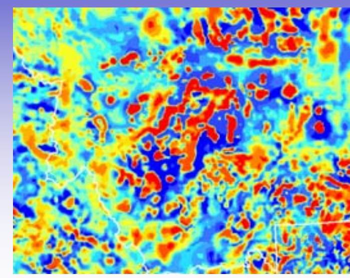
Shape Index of Nevada Geologic Map



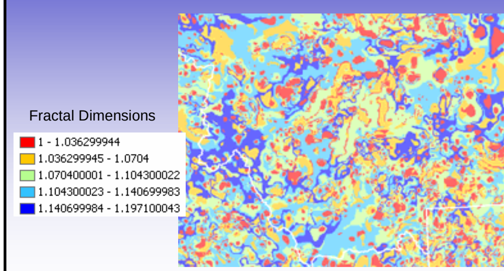
Boulder Batholith Intermediate BandPass Magnetics



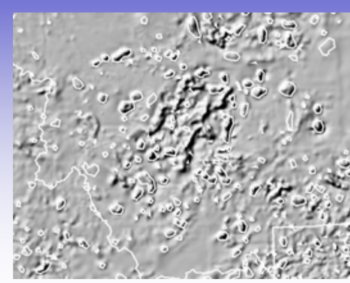
Boulder Batholith – Magnetic Anomaly by SA Method



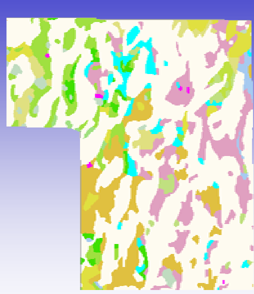
Boulder Batholith – Fractal Dimensions of Magnetic Anomalies



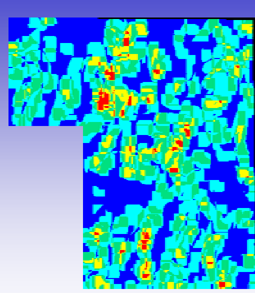
Boulder Batholith Plutons



Variety – Neighborhood Statistic



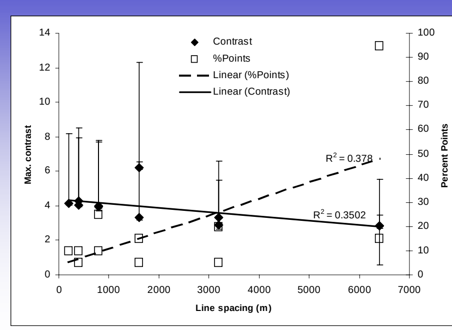
Geology – Carlin area



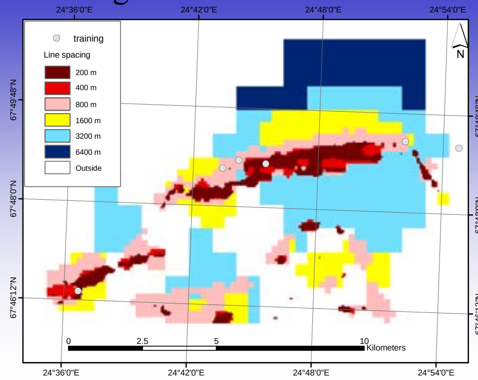
Variety of map units

Lithodiversity

Contrast vs. Line Spacing Aeromagnetic Data - Finland

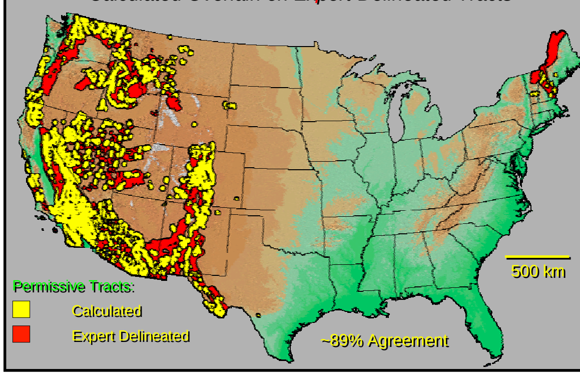


Aeromagnetic Anomalies - Finland



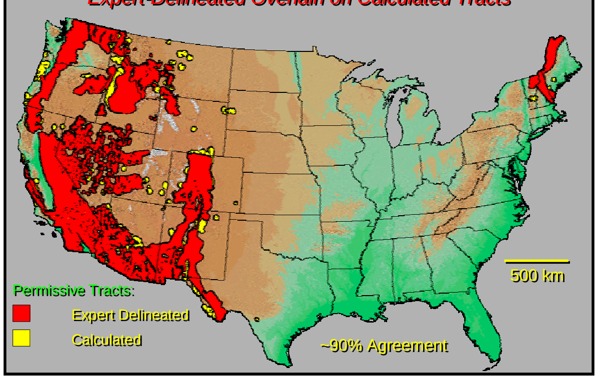
Porphyry Cu & Related Deposits Tract Map

Calculated Overlain on Expert-Delineated Tracts

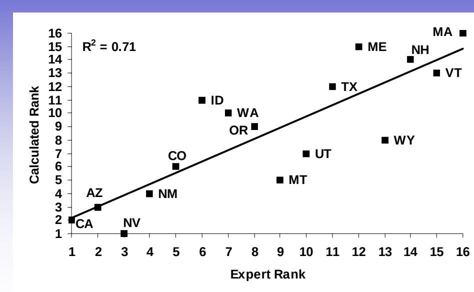


Porphyry Cu & Related Deposits Tract Map

Expert-Delineated Overlain on Calculated Tracts



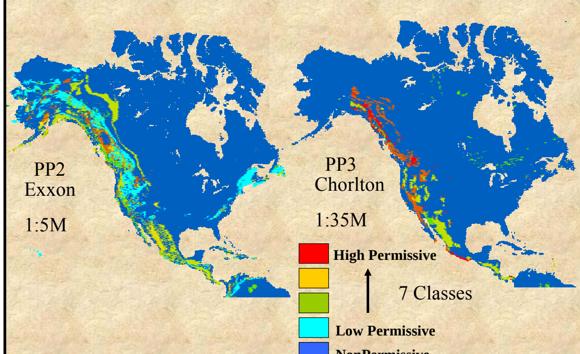
Correlation Expert vs. Calculated US Porphyries (1:2.5M)



Confusion Matrix

Confusion Matrix			
	Calculated		
Expert	*Nonpermissive*	*Permissive*	Row Sum
Nonpermissive	82.1	4.6	86.7
Permissive	8.4	4.9	13.3
Column Sum	90.5	9.5	87 / 40 / 28

North America Comparison



Cross Tabulations

Conterminous U.S.

US		PP2 NP	P	
Expert	NP	76.5	9.6	86.1
	P	6.8	7.0	13.9
		83.3	16.7	83.5

Kappa = 36.5

US		PP3 NP	P	
Expert	NP	80.1	6.0	86.1
	P	6.2	7.7	13.9
		86.3	13.7	87.8

Kappa = 48.7

World		PP3 NP	P	
PP2	NP	75.3	7.9	83.1
	P	13.2	3.6	16.0
		88.5	11.5	78.9

Kappa = 13.7

Gray – marginal sums
Red – Percent Agreement
NP – Nonpermissive area
P – Permissive area
PP2 – Exxon map
PP3 – Chorlton's map

Spatial-Temporal Modeling

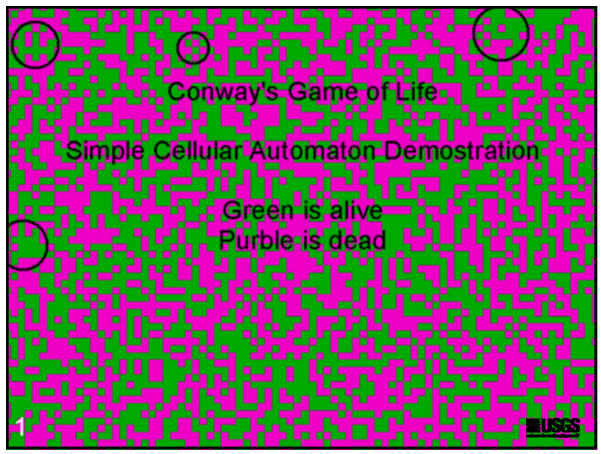
Cellular Automata

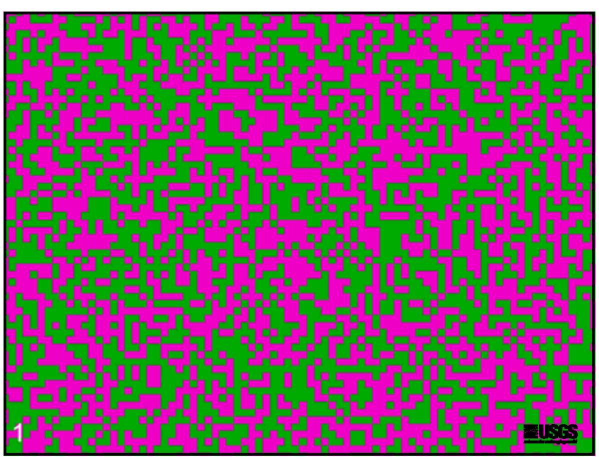
Further Reading

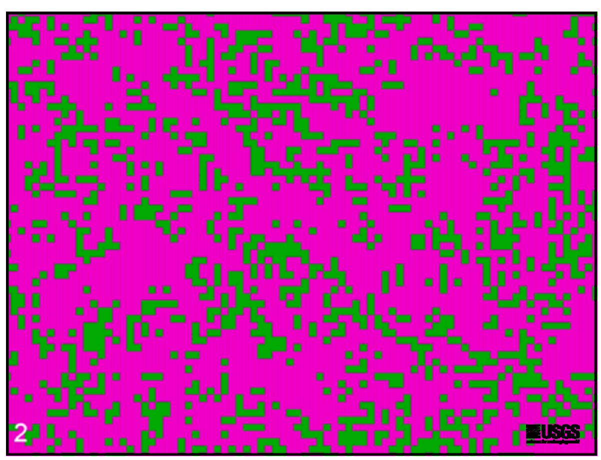
Toffoli, Tommaso, and Margolus, Norman, 1987, Cellular automata machines – a new environment for modeling: Mass., MIT Press, 259p.

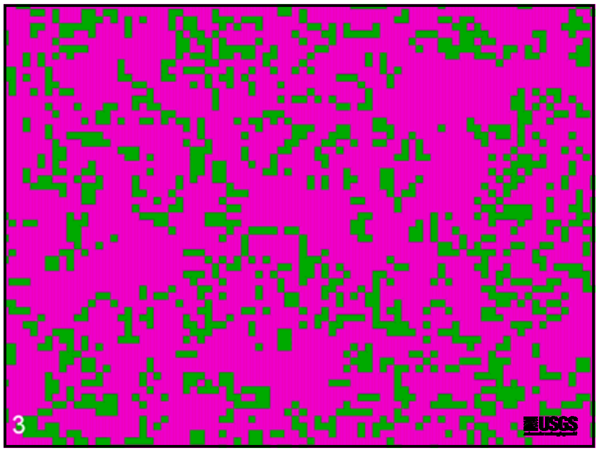
What is a cellular automata?

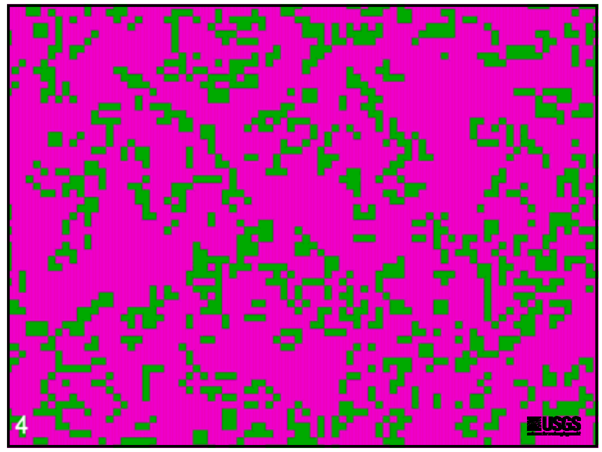
- Cellular automata (CA) are defined by an array of *cells*.
- The state of each cell evolves by a simple transition rule, the *automaton*.
- Implementation of a CA in a GIS involves a summation filter with an if-then or logic rule.

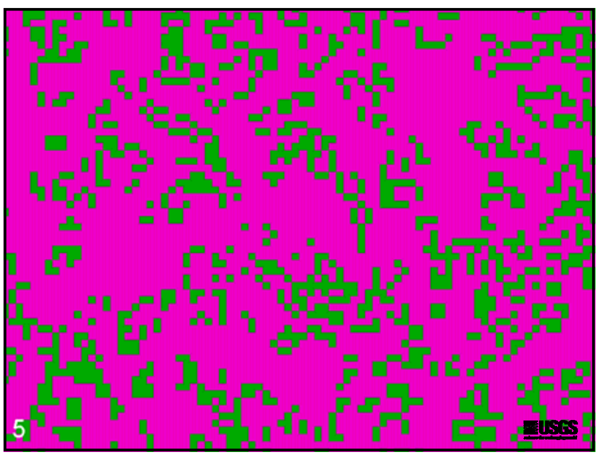




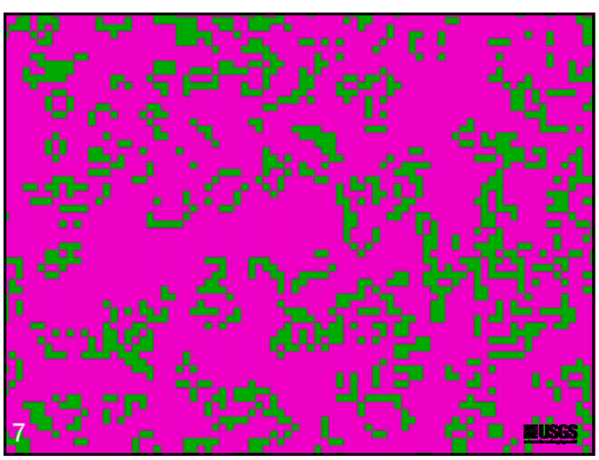


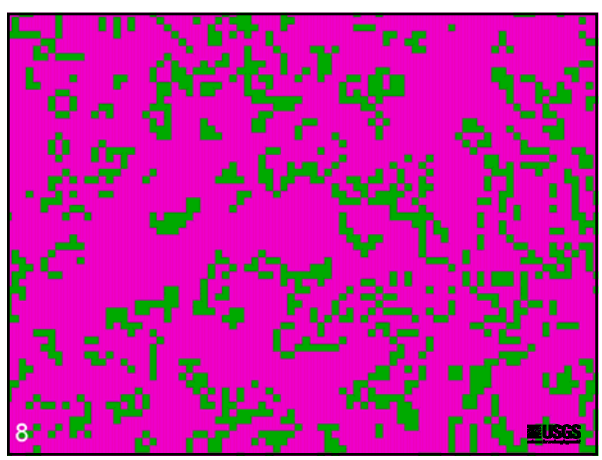

3

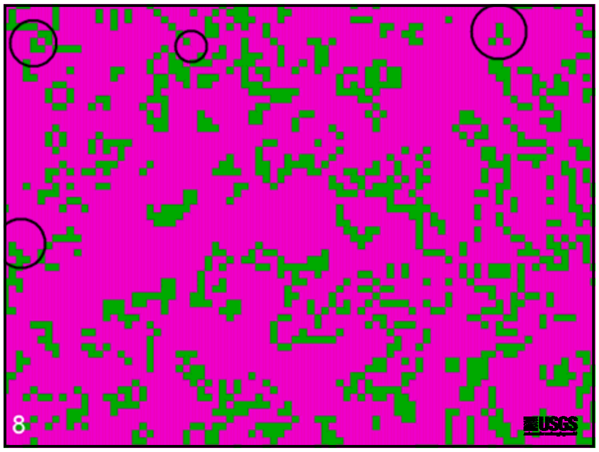

4

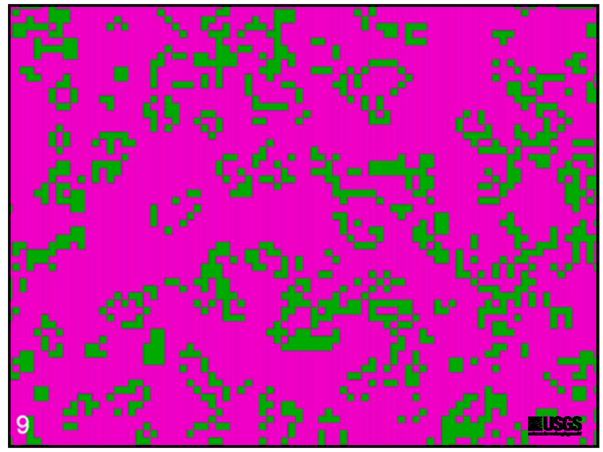

5

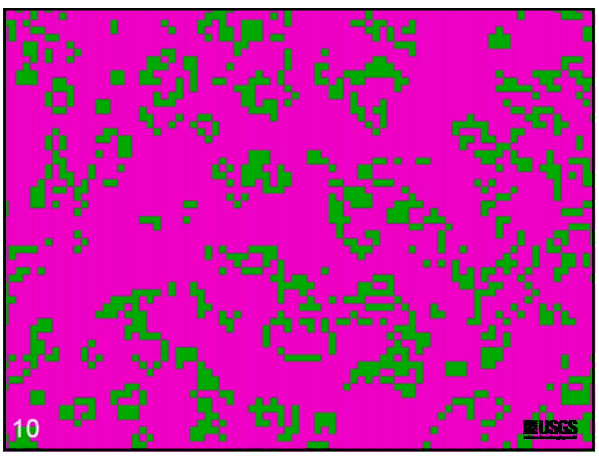

6


7

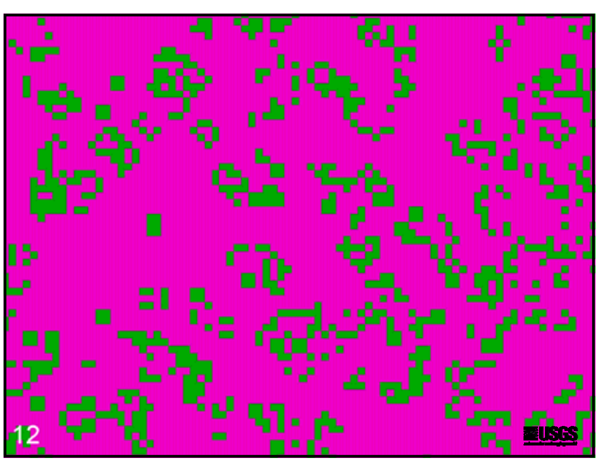

8

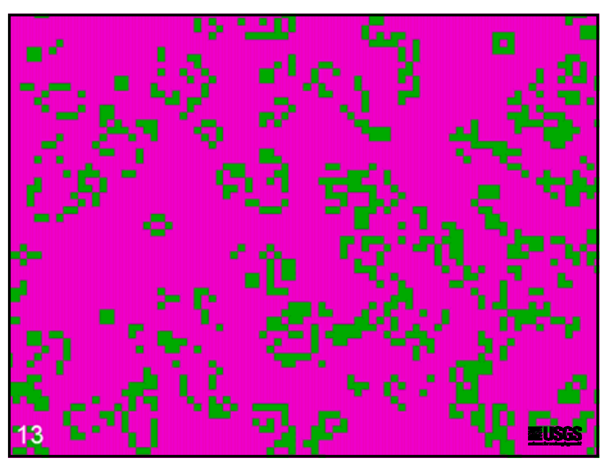

8

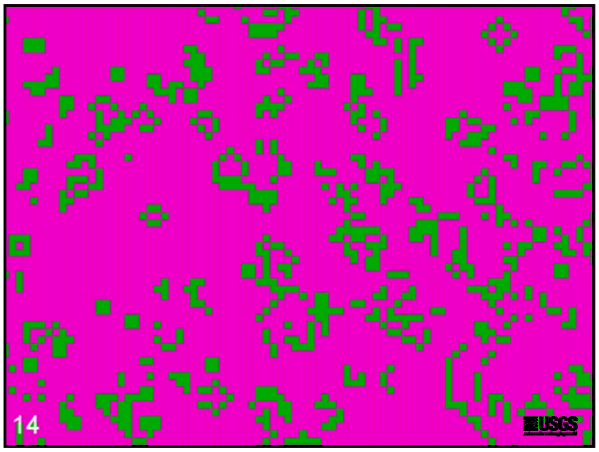

9

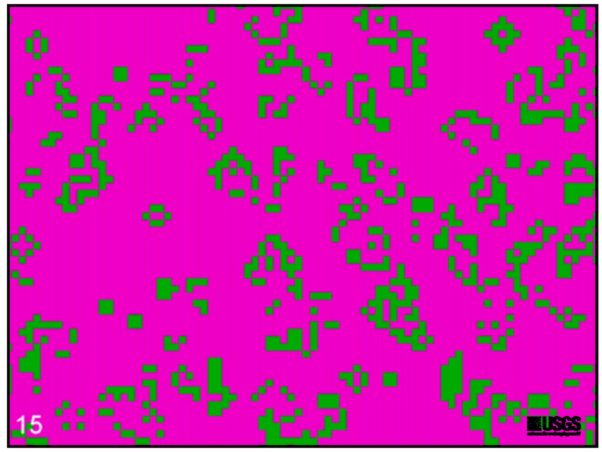

10

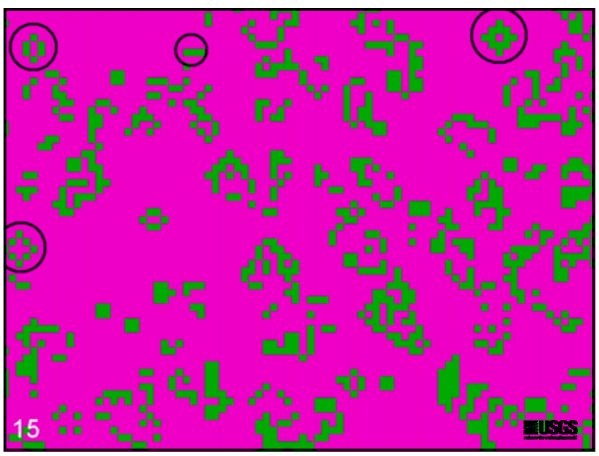

11

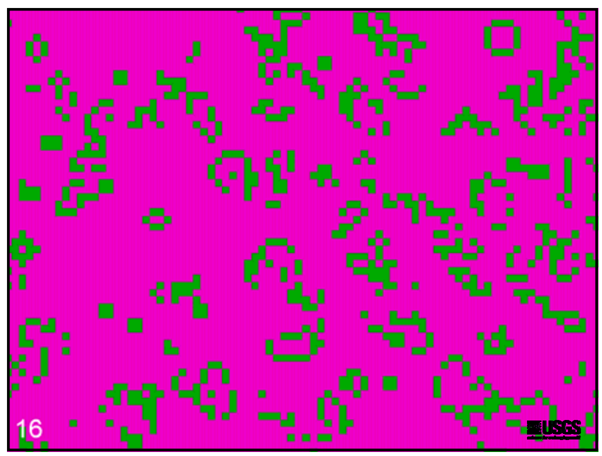

12

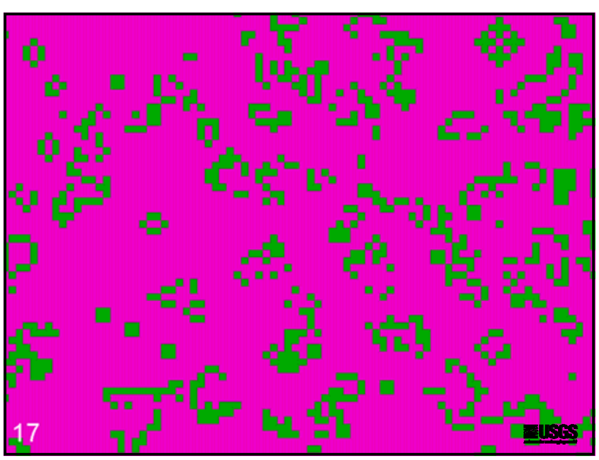

13

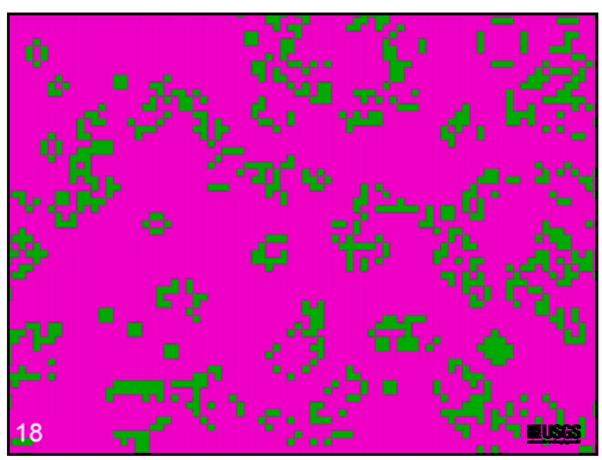

14

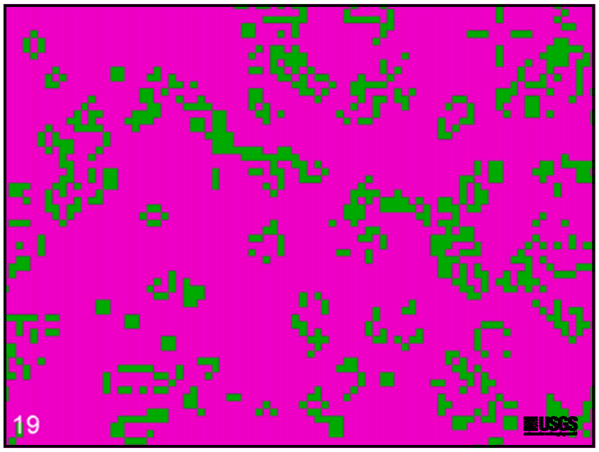

15

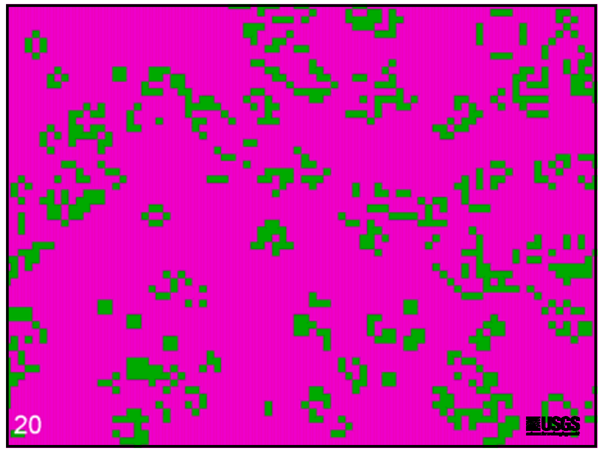

15


16


17


18

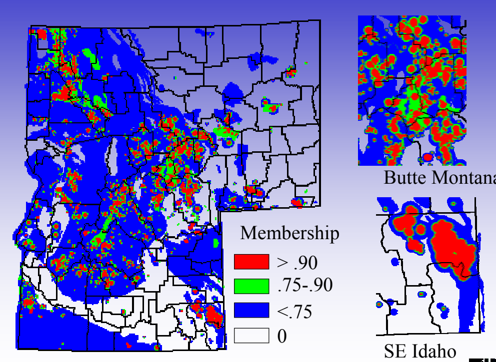




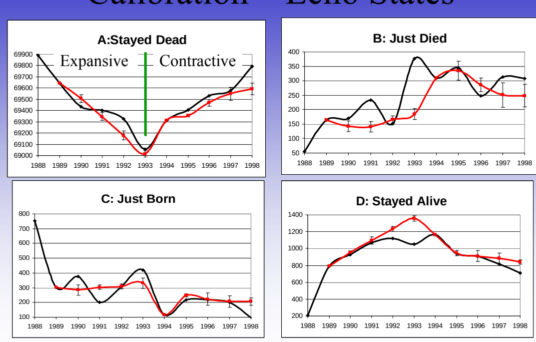
How have CAs been used?

- Modeling evolution of cities
  - Project Gigalopolis
    http://www.ncgia.ucsb.edu/projects/gis/project_gig.htm
- Flow of lava
- Evolution of forest fires
- Physics – diffusion, Brownian motion, defraction
- Biology – life processes

Resource Thresholds

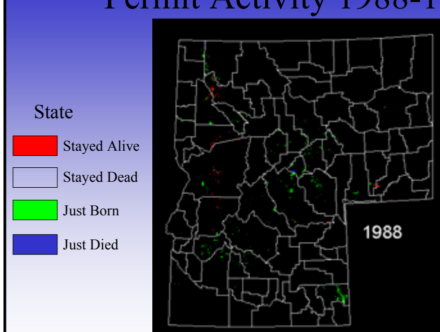


Calibration – Echo States



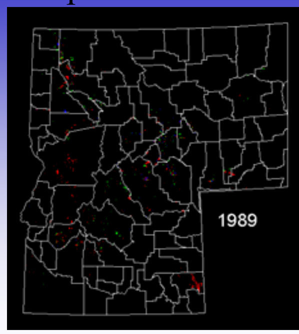
Actual CA Calculated

Permit Activity 1988-1998



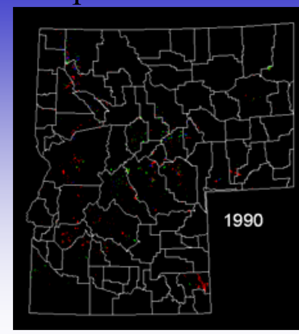
Expansive Period

Expansive Period



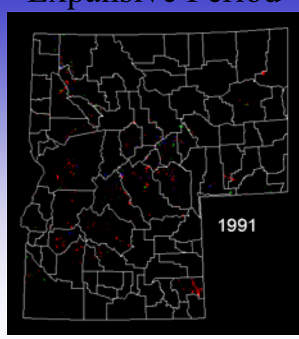
State
- Stayed Alive
- Stayed Dead
- Just Born
- Just Died

Expansive Period



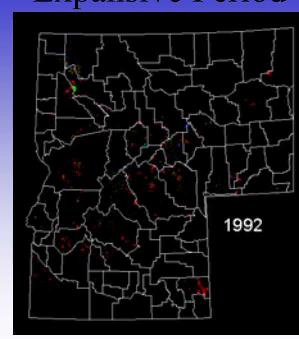
State
- Stayed Alive
- Stayed Dead
- Just Born
- Just Died

Expansive Period



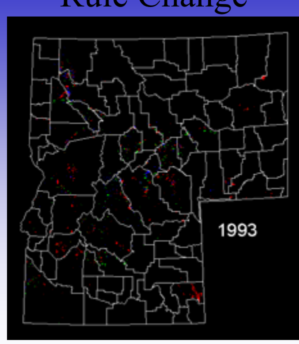
State
- Stayed Alive
- Stayed Dead
- Just Born
- Just Died

Expansive Period



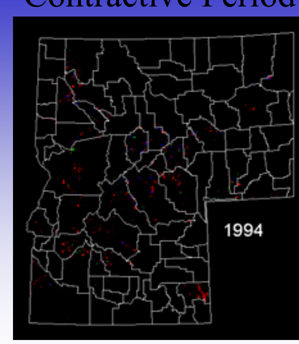
State
- Stayed Alive
- Stayed Dead
- Just Born
- Just Died

Rule Change

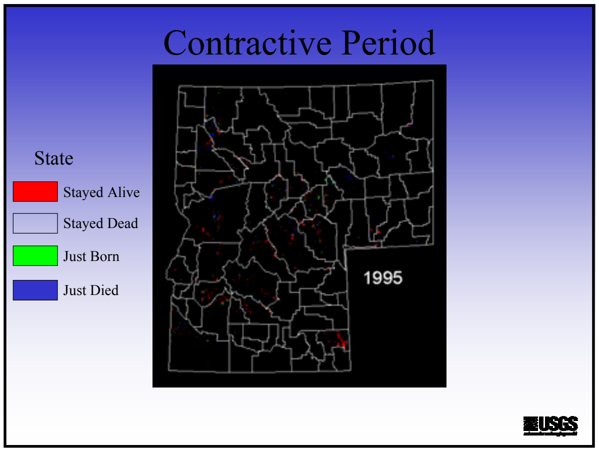


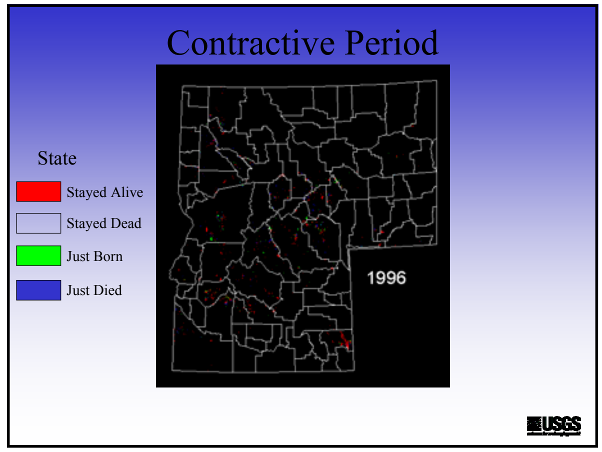
State
- Stayed Alive
- Stayed Dead
- Just Born
- Just Died

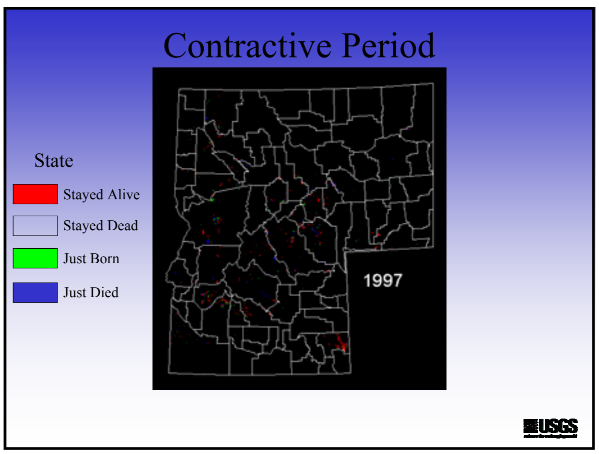
Contractive Period

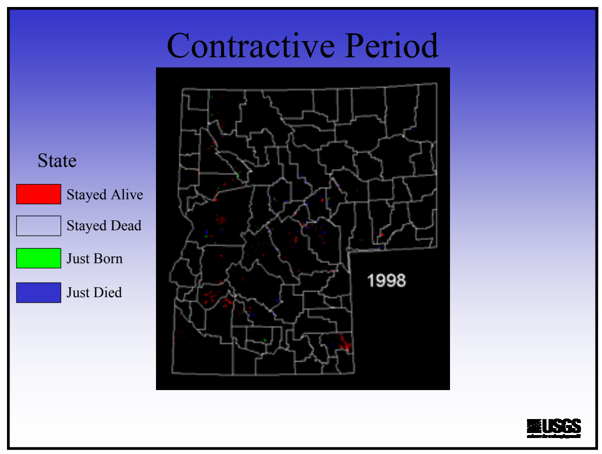


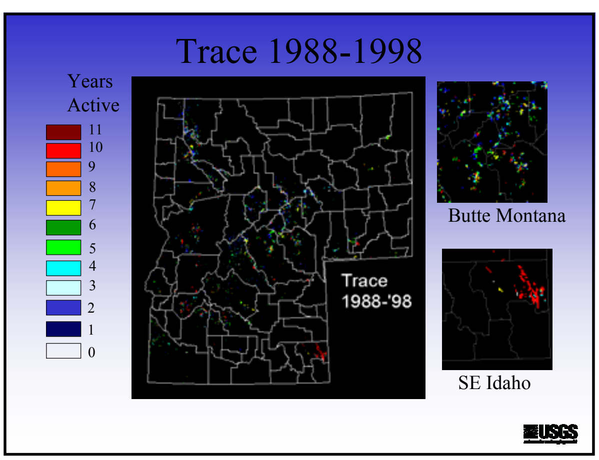
State
- Stayed Alive
- Stayed Dead
- Just Born
- Just Died

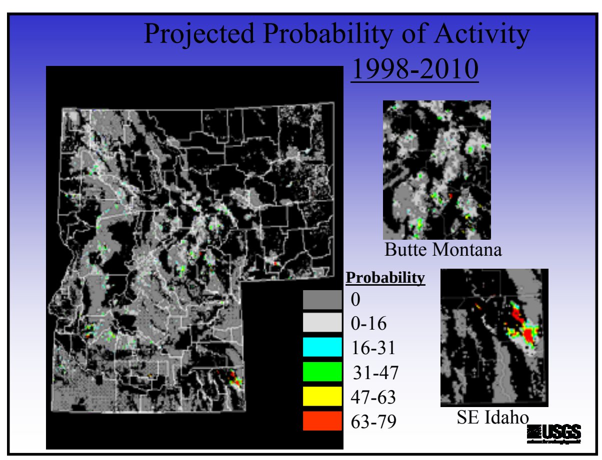

Contractive Period
State
Stayed Alive
Stayed Dead
Just Born
Just Died
1995
USGS


Contractive Period
State
Stayed Alive
Stayed Dead
Just Born
Just Died
1996
USGS


Contractive Period
State
Stayed Alive
Stayed Dead
Just Born
Just Died
1997
USGS


Contractive Period
State
Stayed Alive
Stayed Dead
Just Born
Just Died
1998
USGS


Trace 1988-1998
Years Active
11
10
9
8
7
6
5
4
3
2
1
0
Trace 1988-'98
Butte Montana
SE Idaho
USGS


Projected Probability of Activity
1998-2010
Butte Montana
Probability
0
0-16
16-31
31-47
47-63
63-79
SE Idaho
USGS