Models for Uncertainty in Area-Class Maps

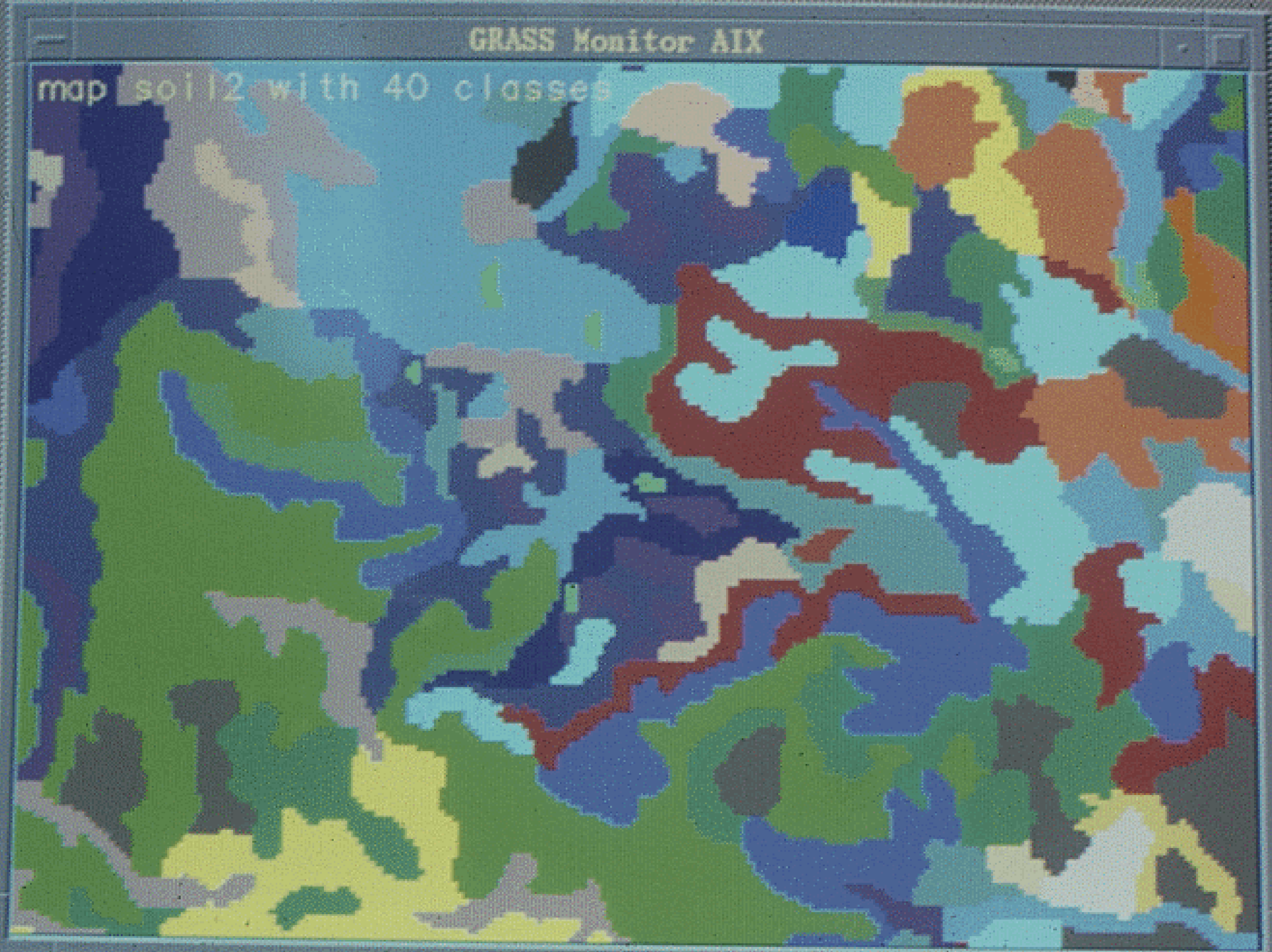
Michael F. Goodchild
University of California
Santa Barbara

Discrete objects and continuous fields

- Discrete objects
  - accuracy of position
  - accuracy of attributes
- Continuous fields
  - $z = f(\mathbf{x})$
  - correct attribute at wrong location, or wrong attribute at correct location?
  - unless singularities can be found and independently located in the real world
  - hilltops, ridges, cliffs

The area class map

- Assigns every location **x** to a class
  - Mark and Csillag term
  - $c = f(\mathbf{x})$
  - a nominal field (or perhaps ordinal)
  - classified scene
  - soil map, vegetation cover map, land use map
- We have no adequate models of uncertainty for this type of map


GRASS Monitor AIX
map soil2 with 40 classes

Uncertainty modeling

- Area-class maps are made by a long and complex process involving many stages, some partially subjective
- Maps of the same theme for the same area will not be the same
  - level of detail, generalization
  - vague definitions of classes
  - variation among observers
  - measuring instrument error
  - different classifiers, training sites
  - different sensors

Error and uncertainty

- Error: true map plus distortion
  - systematic measurements disturbed by stochastic effects
  - accuracy (deviation from true value)
  - precision (deviation from mean value)
  - variation ascribed to error
- Uncertainty: differences reflect uncertainty about the real world
  - no true map
  - possible consensus map
  - combining maps can improve estimates

Models of uncertainty

- Determine effects of uncertainty/variation/error on results of analysis
  - if there is known variation, the results of a single analysis cannot be claimed to be correct
  - uncertainty analysis an essential part of GIS
  - error model the preferred term

Traditional error analysis

- Measurements subject to distortion
  - $z' = z + \delta z$
- Propagate through transformations
  - $r = f(z)$
  - $r + \delta r = f(z + \delta z)$
- But *f* is rarely known
  - complex compilation and interpretation
  - complex spatial dependencies between elements of resulting data set

Spatial dependence

- In true values *z*
- In errors *e*
- cov(e_i,e_j) a decreasing positive function of distance
  - geostatistical framework
- Scale effects, generalization as convolutions of *z*

Realization

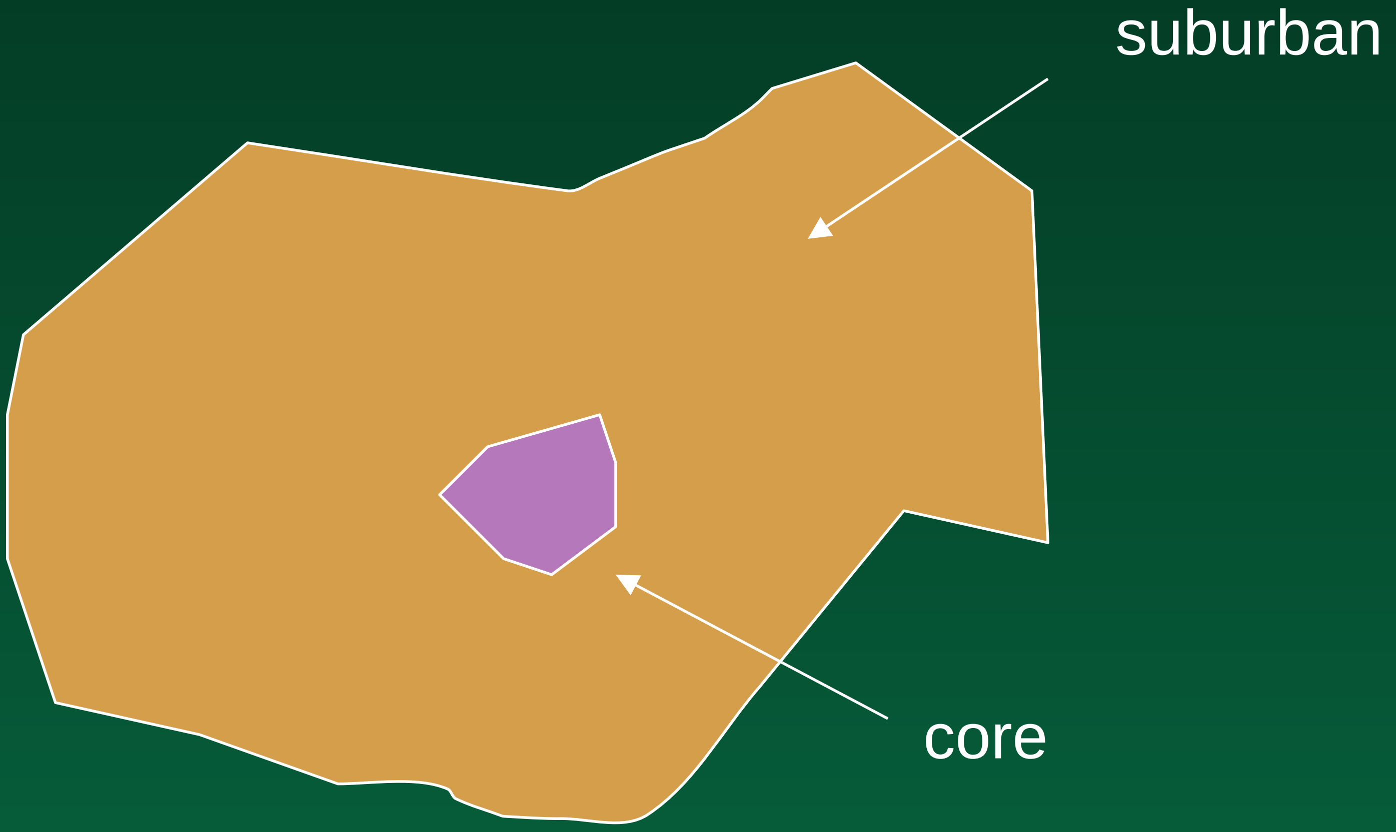
- A single instance from an error model
  - an error model must be stochastic
  - Monte Carlo simulation
- The Gaussian distribution metaphor
  - scalar realizations
  - a Gaussian distribution for maps
    - an entire map as a realization

The area class map

- Field of nominal values $c(\mathbf{x})$, $n>1$
  - spatially autocorrelated
  - in raster, count of i,i joins greater than expected
- In vector, collection of discrete objects
  - nodes, edges, areas in coverage model
  - polygons in shapefile model

A collection of discrete objects

- Three conflicts with the observed nature of area-class maps
- In repeated mappings, positions, attributes, and *numbers of objects* will vary (*topological* variation)
- Positional uncertainties will vary widely depending on boundary clarity
- Confusion of attributes will vary within polygons
  - may be greatest in the center
  - contrary to the *egg-yolk* model



land use type urban

Effects of refining classification

- Boundaries in coarsely classified maps preserved in finer classifications
- Boundaries in coarsely classified maps become polygons at finer classifications

Six requirements of an error model for area-class maps

1. Address confusion *at every point* between observed class *c*' and consensus class *c*
2. Variation between realizations should emulate variation between repeated mappings
3. Autocorrelations in outcomes at nearby points
4. Emulate effects when maps are generalized, both thematically and cartographically

Continued requirements:

5. Realizations should be invariant under changes in underlying representation, *e.g.*, raster cell size
6. Nominal case: results invariant under reordering of classes

- Review known models against these requirements

1. The confusion matrix

- Useful descriptive device
  - quality control
- Comparing classifiers, observers, scales, accuracies
- $p(c' \mid c)$
- Applied per-pixel or per-polygon
- Per-polygon case:
  - no within-polygon variation (violates 1)
  - no variation in topology (violates 2)
- Per-pixel case: no spatial dependence in outcomes (violates 3, 5)

2. The epsilon band

- Addresses only positional accuracy in a fixed topology
- Assumes uniform degree of positional accuracy
- Violates 1, 2, 4

Models based on vectors of probabilities

- At every location:
  - $P(\mathbf{x}) = \{p_1, p_2, ..., p_n\}$
  - assume raster representation, *P* constant over cell
- Simple random assignment
  - no spatial dependence
  - violates 3
  - violates 5 since cell size would be evident in outcomes
- How to induce spatial dependence?

Spatial dependence in outcomes

- Independent outcomes
  - zero spatial dependence between pixels
  - perfect positive spatial dependence within pixels
  - implies pixel size is meaningful
- Induce spatial dependence
  - range >> pixel size
  - spatial dependence falls smoothly
  - independent of pixel size

3. Simple convolution

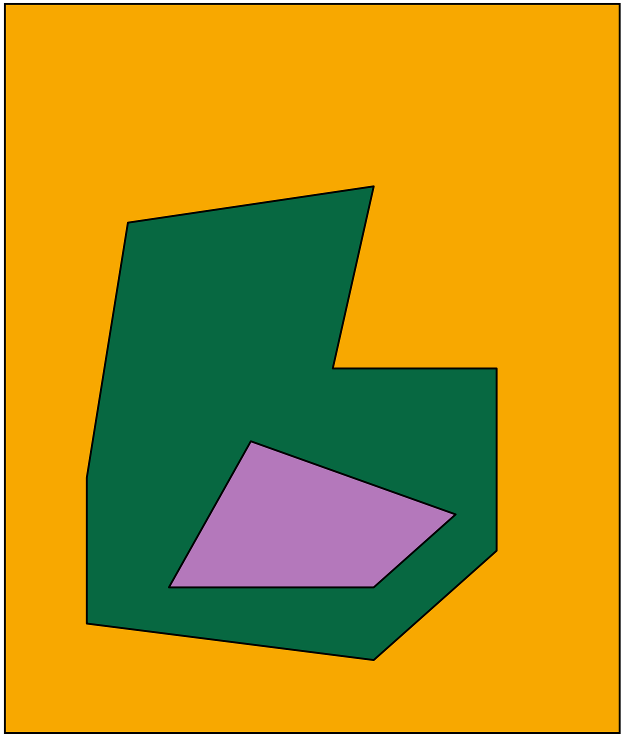
- Generate independent outcomes in each pixel
- Convolve using a modal filter
  - induces spatial dependence
  - size of filter determines range of dependence
  - satisfies 3, 5
- Posterior proportions not equal to prior probabilities
  - convolution favors more probable classes

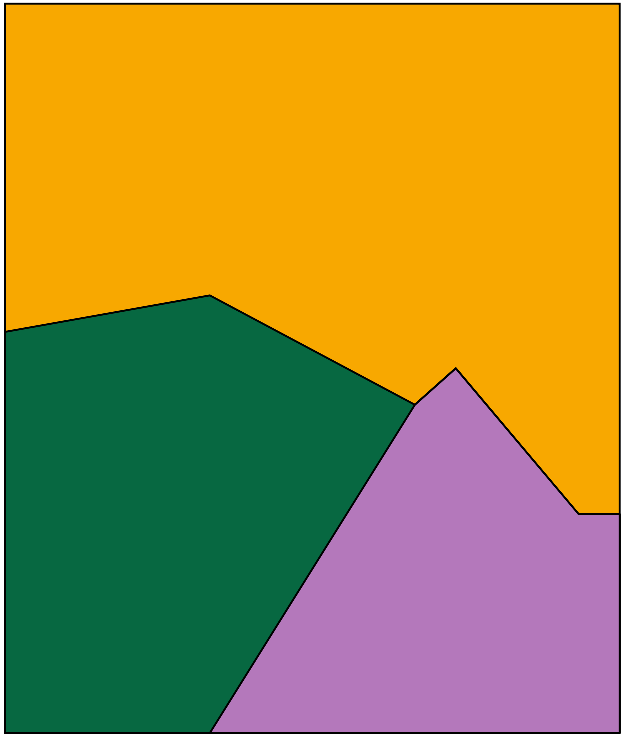
4. Sequential assignment

- Goodchild, Sun, and Yang *IJGIS* (1992)
- Random field *z* with controlled spatial dependence
  - U(0,1)
  - assign class
  - e.g. {0.2,0.3,0.5}
    - 1 (0.0,0.2); 2 (0.2,0.5); 3 (0.5,1.0)
    - z = 0.1, *c* = 1
    - *z* = 0.3, *c* = 2
    - *z* = 0.8, *c* = 3

Comparing to criteria

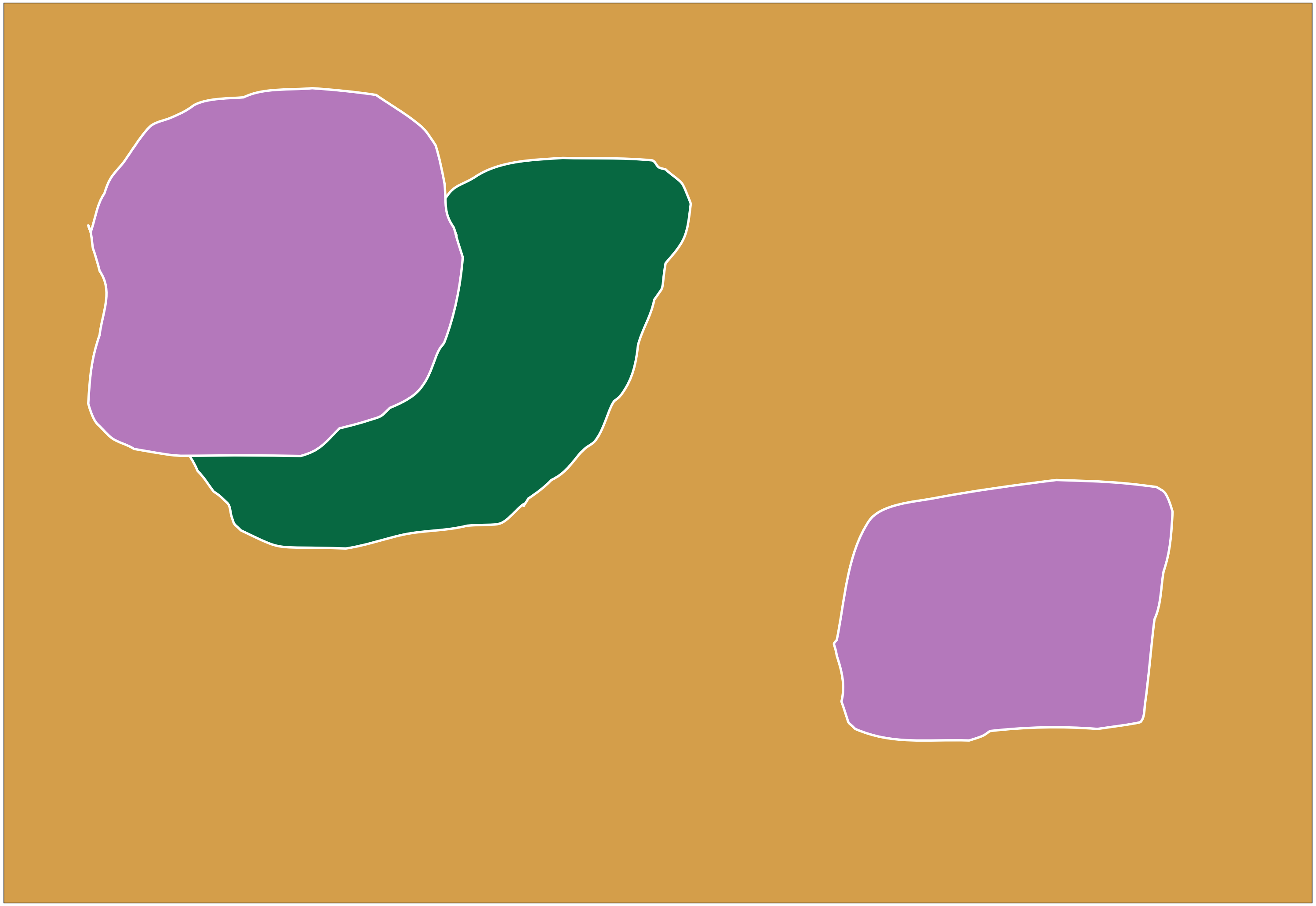
1. Within-polygon variation: yes
2. Topological variation: yes
3. Spatial dependence: yes
4. Generalization: increase cell size, smooth *z*, smooth *P*
5. Independent of cell size: yes
6. Invariant under reordering: no





Indicator Kriging

- Assign Class 1, notClass 1
- Among notClass 1, assign Class 2, notClass 2
- Continue to Class *n*-1
  - notClass *n*-1 = Class *n*

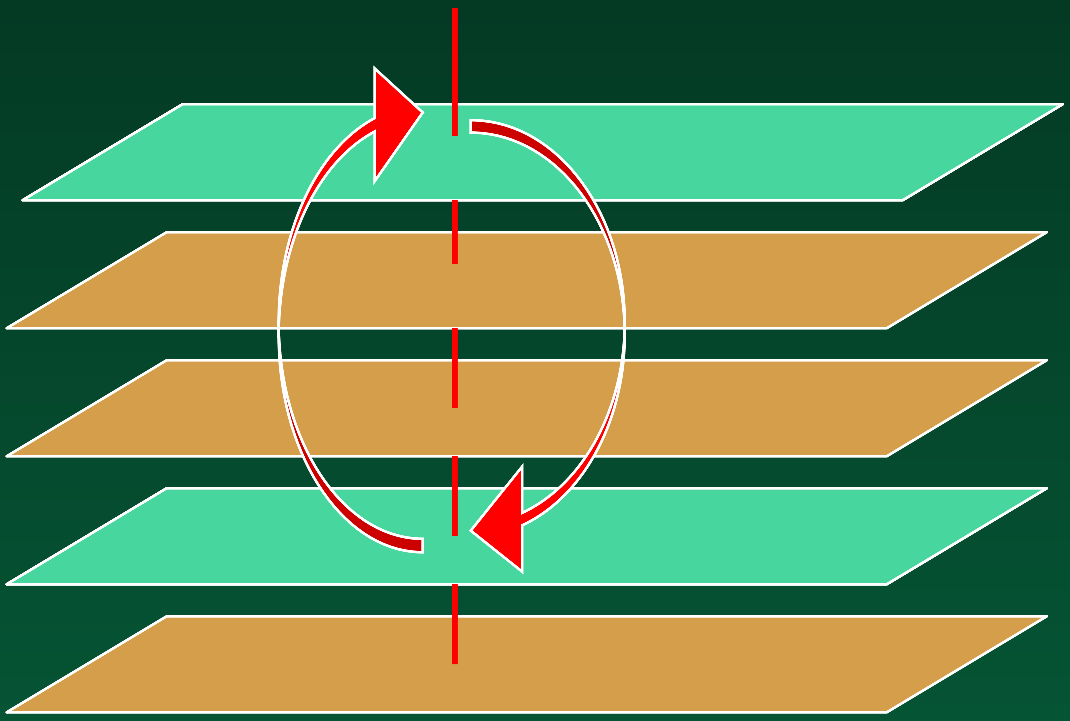


Process-based interpretation

- Class *i* antecedent to class *i*-1
  - e.g. agriculture invaded by urban
  - e.g. grassland invaded by forest
  - shape of boundary between class *i* and class *i*-1 determined by class *i*
  - some applications have inherently ordered classes
    - but in this model all classes are ordered

5. Shuffling across realizations

- Shuffling within realizations unacceptable because of heterogeneity
- Generate *N* realizations with random assignment
- Establish target spatial dependencies
- Pick random pixel
  - pick random pair of realizations
  - swap if closer to target in both realizations

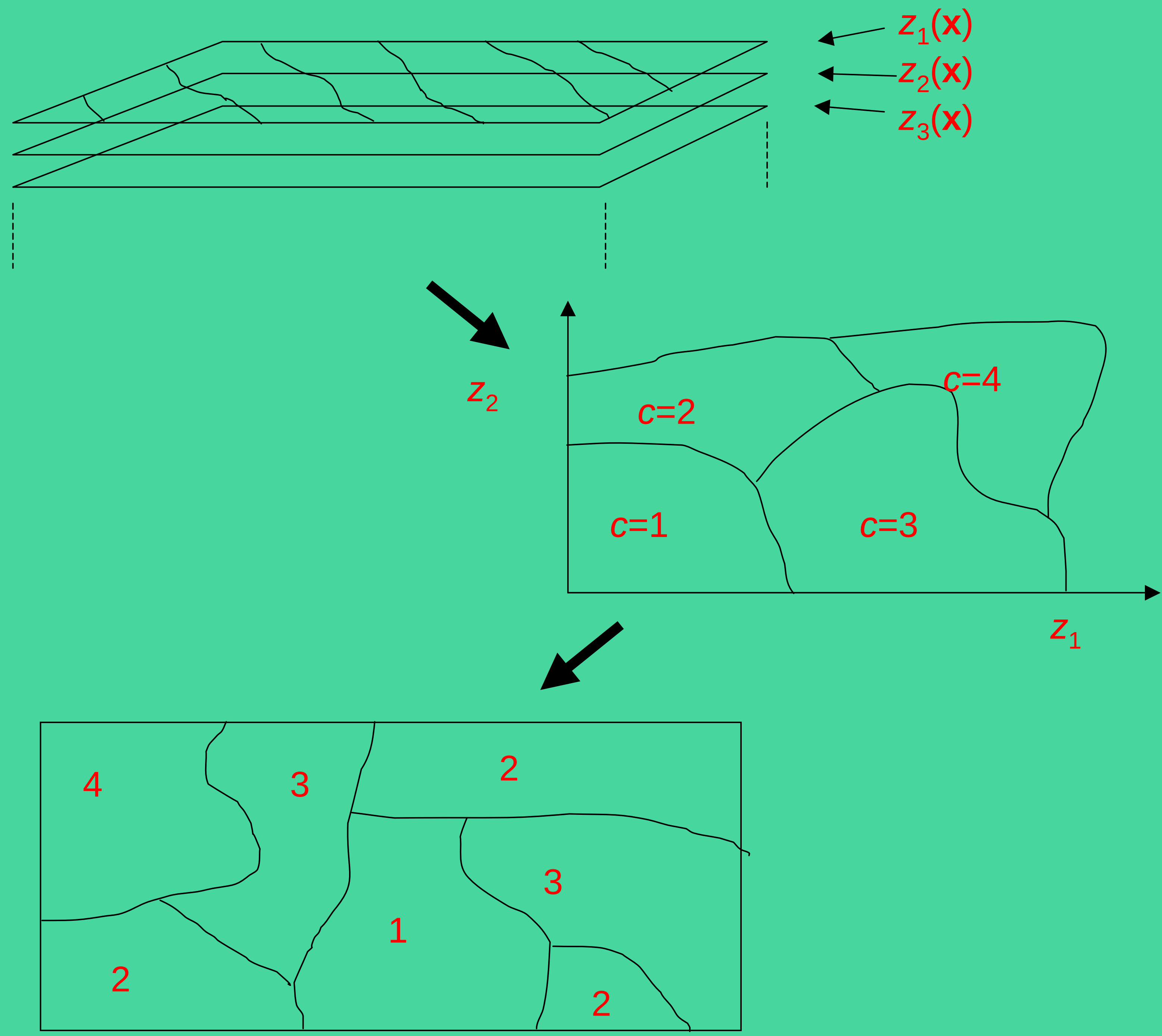


Properties

- No justifying interpretation
  - it works
- Spatial dependence characterized at pixel level
  - no generalization possible, violating 4
- It satisfies all other criteria

6. Phase-space model

- *m* dimensional "phase" space defined by field variables
  - partition into *n* regions
- Generate *m* random fields to locate **x** in phase space
- Assign **x** to one of *n* classes
  - compare classifiers
- Goodchild and Dubuc (1987)


$z_1(\mathbf{x})$
$z_2(\mathbf{x})$
$z_3(\mathbf{x})$
z_2
c=2
c=4
c=1
c=3
z_1
4
3
2
3
1
2
2

Properties of the model

- No specification of $P(\mathbf{x})$
  - two versions of the model
- Independent generation of z in each realization
  - no memory between realizations
  - constant proportions
- Fix z and generate distortions in each realization
  - memory determined by z not by P
  - varying proportions
  - size of distortion determines amount of variation between realizations

Properties of the model

- Only classes adjacent in phase space can be adjacent geographically
- Classifiers provide obvious basis
  - but how to calibrate variances, covariances of random fields?
  - how to calibrate in other cases?
  - model is over-specified
    - but strongly motivated by process
- All criteria satisfied
  - generalization by smoothing *z*, coarsening phase-space classification

Conclusions

- Understanding of uncertainty should be process-based
  - phase space
  - ordinal field
- Spatial dependence and topological variation are critical
  - for applications
  - missing in the simpler models
- Some useful methods
  - shuffling most practical
  - phase space most satisfying