

GEOIDE-GSN Summer School 2004
GIS - Spatial Analysis & Statistics
Scott Mitchell

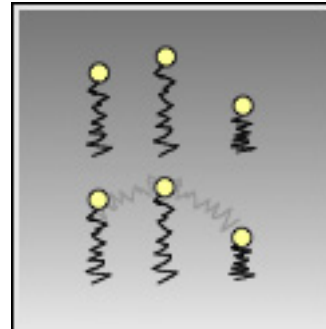
CLASS NOTES

Note: I'll try to tailor the presentation to the backgrounds of participants, but here are some notes corresponding to the slides I will be using (to various degrees).

1. Spatial Analysis: geostatistics, pattern, accuracy, and all that ...

- what's special about space?
- why GIS?
- what are spatial statistics? error?
- applications/examples

- Geography = exploration in space
- analysis of spatially referenced information
- FIRST ORDER effects = local expectation
- SECOND ORDER effects = neighbours



Why GIS?

- understand spatially referenced data
- “massage”, visualize spatial data
- analysis not previously possible
 - community health
 - hydrology / runoff
 - market share
 - ecosystem dynamics / production

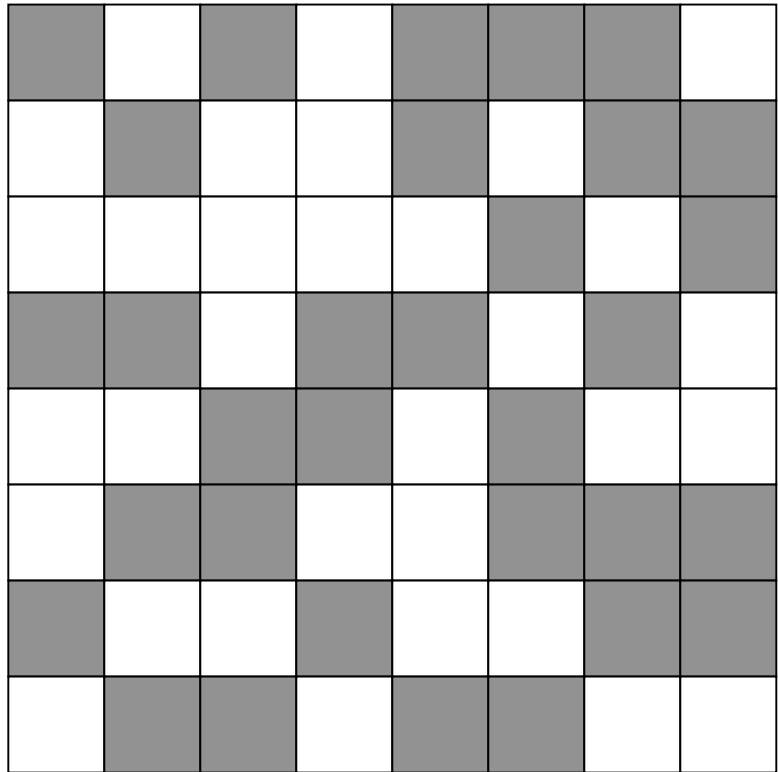
Tobler's Law:

Spatial Statistics

- STATISTICS
 - art and science of summarizing data
 - how surprised should we be?
- SPATIAL STATISTICS
 - exploratory
 - inferential gets complicated (non “iid”)

e.g. the expected number of BB neighbour-pairs:

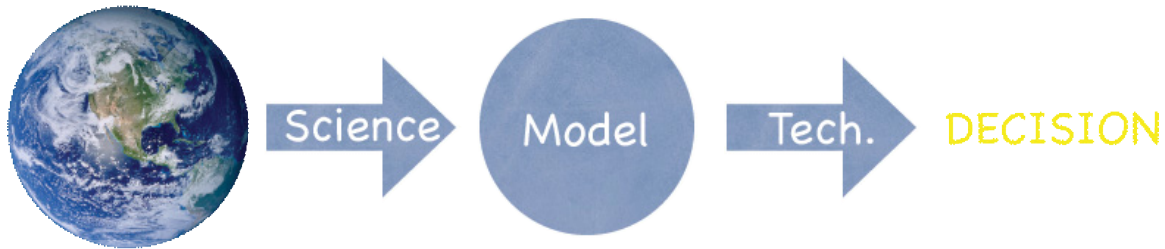
$p(B) * p(B) * N * N_n$
is this random?



Or, “what is the difference between two maps?”

http://eratos.erin.utoronto.ca/fcs/PRES/2MAPS/2maps_index.html
(a current GEOIDE project)

2. How do we approach research questions?



- **HOW DO WE REPRESENT SPATIAL DATA?**
 - continuous/ discrete vs raster/ vector
 - measurement frameworks (for variables)
 - neighbours for spatial data...! (topology vs matrix)
- **WHERE DO ERRORS COME FROM?**
 - living with errors... ("absorption")
 - sampling
 - prediction
 - simulation
- **WHAT DO WE USE STATISTICS FOR?**
 - estimation of (selected) model parameters
 - prediction of values using a model
 - simulation of new realizations using a model
- **CAN WE MATCH STATISTICAL MODELS AND GEOGRAPHICAL DATA MODELS?**
 - no one-to-one correspondence ***
- **BLUFFING IS DANGEROUS...**

3. Mapping spatial distributions

- issues in mapping distributions - role of generalization, perception, ...
- information flow versus accuracy
- gradients, abrupt changes, mixtures, ...

Measuring accuracy:

- accuracy "indices":
 - the "usual" - SSQ
 - spatially weighted SSQ
 - boundary index
- but the indices are inter-dependent
- number of classes characterizes complexity
- N pixels with G classes: G^N

Model selection:

e.g. principle in Csillag & Kabos
information theory (Akaike, Bayes)

penalty for overcomplication (parsimony):

e.g. $AIC = f(-\text{ve log-likelihood (explanation), \# parameters})$

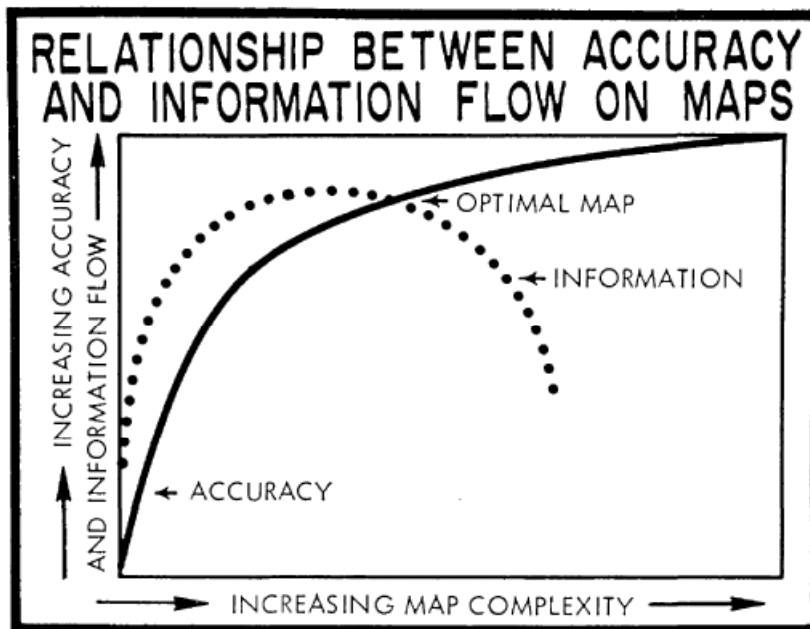


FIG. 22. The accuracy curve has been defined in this paper but the information flow curve (dotted line) is theoretically based. The optimal representation of a distribution will be achieved when both of these map functions are maximized.

Spatial processes (using statistics lingo):

Y = random* variable

y = realization / obs

$f_Y(y)$ = probability distribution

Expected value:

$E(Y) = \sum y f_Y(y)$ if Y is discrete, or

$E(Y) = \int y F_Y(y) dy$ if Y is continuous

- expected values of functions of the random variable are also of interest, especially the expected squared deviation of the random variable from its mean (VARIANCE):

$$\text{VAR}(Y) = E((Y-E(Y))^2)$$

When one has two random vbls (X, Y), joint probability distribution,

$$f_{XY}(x,y)$$

COVARIANCE = expected tendency for values of X to be 'similar' to values of Y

$$\text{COV}(X,Y) = E((X-E(X))(Y-E(Y)))$$

(divide by product of the SD's to get correlation)

Returning to space:

$Y(s)$ = a spatial stochastic ($s \in \mathbf{R}^2$; or $s_i=1, \dots, N$)

first order: variations in the mean:

$E(Y(s))$: global trend

second order: local variations from the mean:

if there is spatial dependence, these have a local covariance structure

(COV_{ij})

both can vary in space!

if they don't, called stationary

- $Y(s)$ is stationary (homogeneous) if its statistical properties are independent of location in real space R^2
- $E(Y(s))$ and $VAR(Y(s))$ are constant in space, and $COV(Y(s_i), Y(s_j))$ depends only on their relative locations, not their individual absolute locations in R^2
- also isotropic if COV only depends on distance between points, not direction

Measuring second order effects:

- look at neighbours – values and distances
- $r = \sum W_{ij} D_{ij}$ where W describes neighbours and D describes values
 - Geary: $D_{ij} = (y_i - y_j)^2$
 - Moran: $D_{ij} = (y_i - \bar{y})^2 (y_j - \bar{y})^2$

Modelling second order effects:

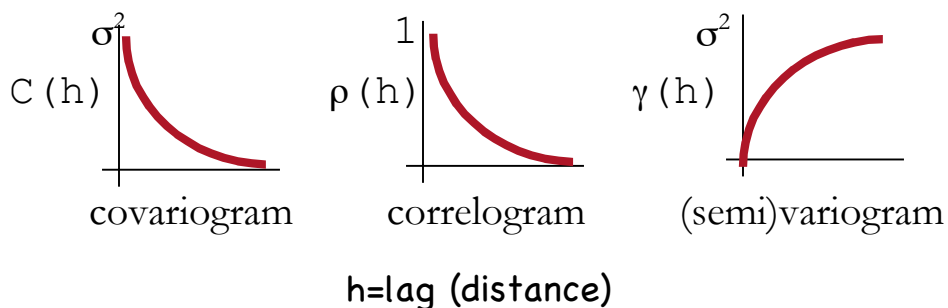
specifying the joint distribution with GEOSTATISTICS:

$$E(Y(s+h) - Y(s)) = 0$$

$$VAR(Y(s+h) - Y(s)) = 2\gamma(h) \text{ (variogram)}$$

ignoring anisotropy, and assuming $Y(s)$ is stationary:

$$COV_{ij} / (\sigma_i \sigma_j) = \rho(h) = COV(h) / \sigma^2 \text{ and } \gamma(h) = \sigma^2 - COV(h)$$

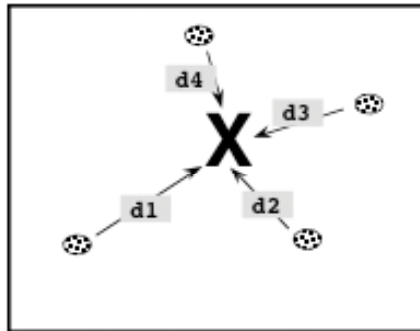


Let's see that in action – DEMO in R

R is an open-source statistics package using the S statistics language
(which is also used in S-Plus)

(<http://www.r-project.org>)

3b) Spatial Interpolation



...what is $z(x_0)$?

- managing errors - confidence in prediction?
- accuracy - what is the truth? (think is truth)
- precision - how detailed is our knowledge?

- **Classify – simplest approach**
 - assign single category to intermediate areas
 - can be global or local
 - variance partitioning
 - Voronoi/Thiessen/Dirichlet polygons: only the nearest $z(s)$ counts
- **Trend Surface Analysis**
 - fit a function which describes a surface to / through the samples
 - essentially spatial multiple regression
 - higher order -> better fit, but number of parameters becomes “silly”
- **Distance Weighted – IDW**
 - inverse distance weighting
 - no stochastic component (!), but closer observations have more weight
 - $\hat{z}(X_0) = \sum \lambda_i Z(X_i)$ with $\sum \lambda_i = 1$
- **Other local functions:**
 - various possibilities for fitting local functions to the points, piecewise
 - e.g. thin plate splines (ANUDEM, s.surf.tps) - fits local areas quite well (retain local features), can be exact or inexact, but no direct measurement of errors, and can be unrealistically smooth
- **Kriging**
 - based on Krige (1966) (but... Matheron, ...)
 - also weighted average / partition between trend and local covariance
 - minimize variance:
 - $\sigma^2(X_0) = E\{(Z(X_0) - \hat{z}(X_0))^2\}$
 - $\hat{\sigma}_{\min}^2(X_0) = \sum \lambda_i \gamma(X_i, X_0) + \psi$, achieved when

$$\sum \lambda_i \gamma(X_i, X_j) + \psi = \gamma(X_i, X_0) \text{ for all } j$$

- Types of kriging:
 - simple - GLS regression, assume 2nd order stationarity with a known mean
 - “ordinary” (no a priori mean) (also corrections for anisotropy, use of nested variograms)
 - stratified - classify area into meaningful sub-areas
 - universal kriging - with trend
 - block kriging - changing resolution, dealing with large local variability
 - co-kriging - “cheap” covariates (>1 attribute)
 - indicator kriging, probabilistic, ...
- **Comparing interpolations:**
 - characterize the prediction error:
 - bias / accuracy
 - precision
 - spatial arrangement (map)
 - non-spatial summary (RMSE)
 - analytical - fully specified model (parameters)
 - empirical - cross-validation (resampling, bootstrapping)
- **General comparison:**

TYPE	Spatial Dependence	Exact?	Comments
trend surface	global	no	very smooth
proximity polygons	local (abrupt)	yes	nearest neighbour
distance weighted	local (neighbours)	no/yes	smoothing independent of samples
geostatistics: kriging	local (stochastic)	yes	stationary smooth

PRACTICAL SESSION (R)

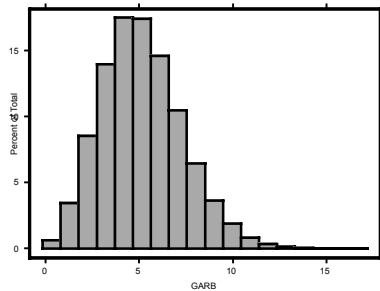
Alternatives:

- Simulation - if don't know enough for deterministic modelling, perhaps should use the knowledge we do have to control stochastic simulation in order to get equally-likely surfaces instead of overly-smoothed interpolations
- e.g. can combine stochastic techniques such as Monte-Carlo simulation with traditional interpolations such as kriging

PRACTICAL: Monte Carlo, Autoregressive models

4. Point Pattern Analysis

- events in space with coordinates
- deviation from randomness - regular OR clumped
- 1st order stat: intensity $\lambda(s) = \lim_{ds \rightarrow 0} \{E[Z(s)] / ds\}$
- 2nd order: neighbourhood effect
 $\gamma(s_i, s_j) = \lim \{E[Z(s_i)Z(s_j)] / ds_i ds_j \}$
- Random location?
 - (homogeneous) Poisson model
 - $A = \sum z(a_i)$ (locations/ areas), and $z(a_i), z(a_j)$ are independent



- First order effects:
 - quadrat counts (classical surveys)
 - 2-D histograms (or at random locations)
 - counts ($z(a_i)$) proportional to area, but get into trouble if area too small
 - dispersion test
 - sampling test
- Second order effects:
 - nearest-neighbour tests
 - K-function (Ripley): $E[\# \text{events within dist of an arbitrary location}]$
 - gain confidence intervals
 - ...

5. Local statistics

- classical: kernel estimation (moving windows)
- shape parameter; can be edge corrected, optimized
- excellent tools for EXPLORATORY work on spatial data (test stationarity, look for boundaries, hot spots, ...)
- e.g. local versions of Moran, Geary, Getis statistics

6. Partitioning

- “divvying up” / classification
- various purposes – determine common features, trends, identify homogeneous units, separate differences
- tree-based methods:
 - = flexible (growable, prunable) tools for data description
 - CART: classification and regression trees
 - Quadtrees and wavelets
- DEMONSTRATIONS:
 - Partitioning for environmental modeling across space
 - GRASS/R interface