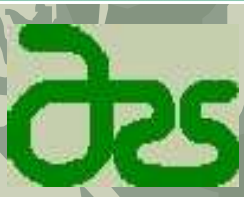
Spatial Statistics Workshop

Agricultural & Environmental Research Applications



March 15 – 16, 2006

Beltsville, MD

Contents

March 15

Morning Session

Afternoon Session

March 16

Morning Session

Afternoon Session

* Supplemental material on the CD in the booklet (for onsite attendees) and, temporarily, at http:\\www.ars.usda.gov\ba\spatialworkshop

An Overview of Spatial Statistics

Bryan Vinyard
Biometrical Consulting Service
USDA, ARS, Beltsville

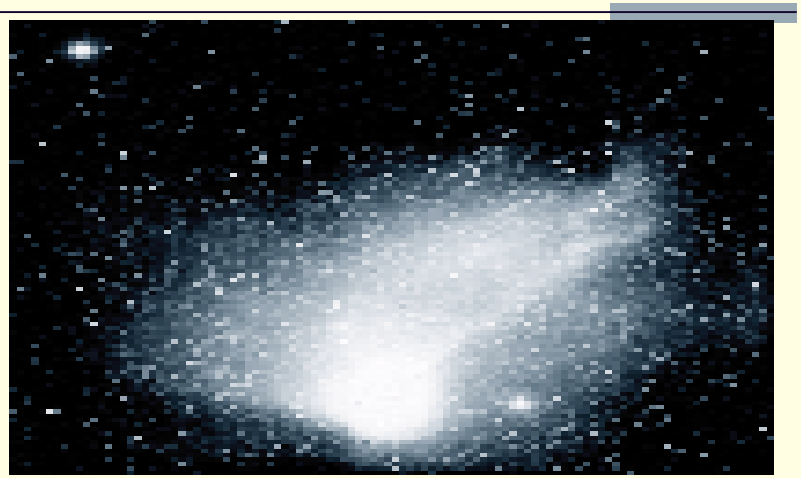
Philosophy of this Workshop

- Focus on Concepts
- Define Terminology
- Illustrate using Graphics
- Avoid excessively Technical Explanations
- Apply the Concepts to Data
- Focus on '*What*?' & '*Why*?' not '*How*?'

"Statistics, the science of uncertainty, attempts to model order in disorder."

Cressie (1991)

Observed Data ('*disorder*' ?)



Characteristics of Interest 'Y'

measured at each observed data point

Examples:

- **Stream Flow**
- **Yield**
- **Muscle Tissue Toughness**
- **CY3, CY5 Image Reflectance**
- **Insect Damage Rating**
- **Bacteria Count**
- **Nitrate Flux**
- **Turbidity**

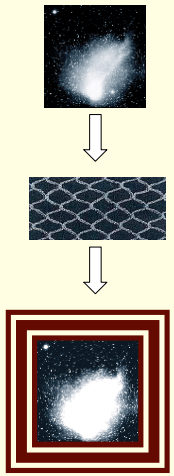
Primary Goal of Applied Statistics

Use observed Y values
together with scientific knowledge

to obtain accurate predictions ($\hat{Y}$)
of unobserved Y values
or to understand a process
(i.e., test the effects of a treatment)

by creating a statistical model:

$\hat{Y}$ =

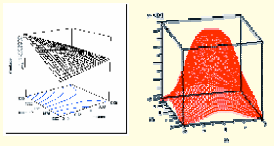


Notes

Fitting a statistical model to data can be viewed as a process of identifying a sequence of "filters" through which the observed data are "sifted".

Initial Attempt to Predict Y
Model the 'Large-Scale' Trend

$\mathbf{Y}$ =



+ ε^*

$\hat{Y}$

where

$\mathbf{Y}$ is predicted by fitting a 'large-scale' trend to the observed data.

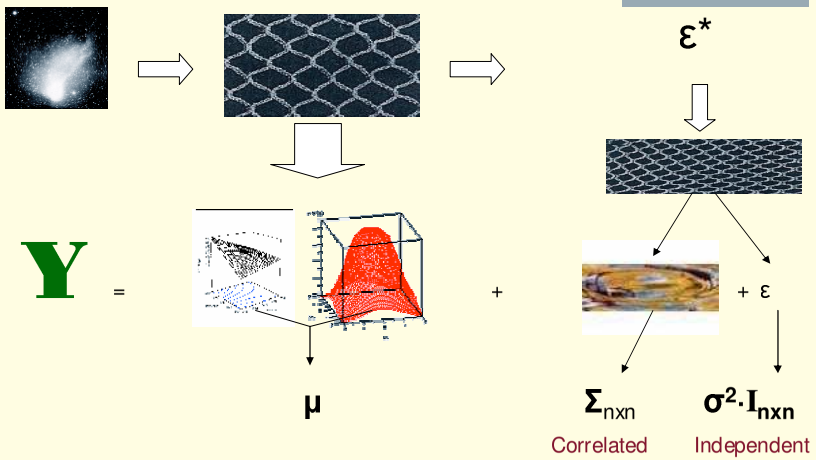
ε^* is data variability remaining after the model is fit.

Notes

When modeling a characteristic of interest, Y, there are typically well established large-scale relationships between Y and one or more fixed-effect “covariates” (i.e., regressors) and/or random “block” effects.

These large-scale effects are modeled first so that the remaining (i.e., residual) variability can be examined in detail on a small-scale to model any spatial dependencies that may be present.

Refine the Model to Predict Y
Model ‘Small-Scale’ Variability

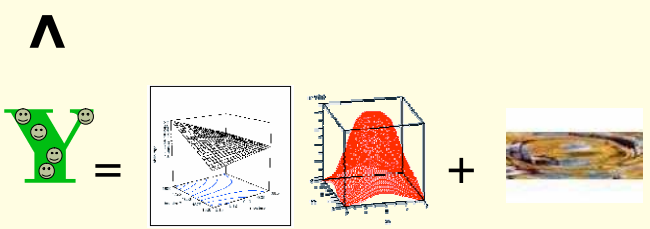


Notes

The initial “filters” capture the large-scale relationships, letting the small-scale relationships remain in the “residual” data to be modeled by a small-scale filter.

Accurate modeling of the small-scale variability, composing the residual data, often requires identification of an appropriate correlation (i.e, covariance) structure.

Refine the Model to Predict Y
Model ‘Small-Scale’ Variability



Notes

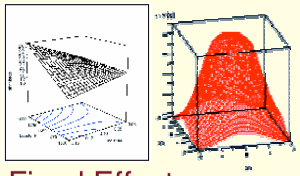
The presence of small-scale dependencies often means that there is correlation among data values located within a certain distance or proximity to one another.

Making use of this "common information" shared by correlated data values improves a model's accuracy.

The primary goal of modeling small-scale dependencies is the identification of an appropriate correlation (i.e., covariance) structure.

Decomposing the Data Variability
ANOVA Terminology

Large-Scale



Fixed Effects

Means
- Deterministic Functions
- Regressors(Covariates)
- Treatments

Small-Scale

All 'residual' variation

(eg., a raindrop on water surface)

Random Effects*

Variance Component
- Variances
- Covariances/Correlations

*Some models may include large-scale random effects (i.e., blocks)

The General Linear Model (GLM)
Perspectives on Model Components

Y	=	**Large-Scale Variation**	+	**Small-Scale Variation**
Y	=	**Fixed Effects**	+	**Random Effects***
Y	=	**Mean** &/or **Covariates**	+	**Variances** & **Covariances***
Y_{nx1}	=	$X_{nxp} \cdot \beta_{px1}$	+	ε_{nx1}
Y_{nx1}	=	μ_{nx1}	+	ε_{nx1}
Y_{nx1}	=	$\hat{y}_{nx1}$	+	ε_{nx1}

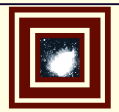
*Some models include random effects (i.e.,blocks) that are considered "large-scale".

Notes

Traditional (General Linear) models typically use only the large-scale effects to model the process; by either predicting Y at various values of regressors or for a collection of experimental "treatments" (i.e., fixed-effects). In traditional GLMs, the small-scale effects do not contribute to improving predictability of Y; rather they are used as precision measures (i.e., root mean-square error) to test hypotheses for the fixed-effects.

Spatial models examine small-scale effects more closely and use the information shared among correlated data values to improve predictability of Y.

The General Linear Model (GLM)
Assumptions – the i.i.d. Mantra


Observed Data - Model Prediction = Model Error

$\mathbf{y}_{nx1}$ - $\hat{\mathbf{y}}_{nx1}$ = $\boldsymbol{\varepsilon}_{nx1}$

Classical GLM assumptions: ε_i are ***i.i.d.***

$\boldsymbol{\varepsilon}_i \sim$ Normal (0, σ^2_ε)
- independent → no correlation among the n residual values.
- identically distributed

Small-Scale Variation
Variances & Covariances that Describe Model Error

For $\boldsymbol{\varepsilon}_i$ ***i.i.d.***
no correlation among the n data values
$\boldsymbol{\Sigma}_{nxn}$ is a diagonal matrix…

$$\boldsymbol{\Sigma}_{nxn}=\sigma^2_\varepsilon \cdot I_{nxn}=\begin{pmatrix} \sigma^2_\varepsilon & 0 & \cdots & 0 \\ 0 & \sigma^2_\varepsilon & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2_\varepsilon \end{pmatrix}_{nxn}$$

When the $\boldsymbol{\varepsilon}_i$ are correlated,
correlations appear as non-zero 'covariances' in the off-diagonals of $\boldsymbol{\Sigma}_{nxn}$ …

$$\boldsymbol{\Sigma}_{nxn}=\begin{pmatrix} \sigma^2_\varepsilon & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma^2_\varepsilon & \cdots & \sigma_{21} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma^2_\varepsilon \end{pmatrix}_{nxn}$$

Notes: The Covariance Matrix

To assist in the visualization of how n observed data points are correlated with one another, statisticians use an n x n "covariance" matrix, denoted as $\boldsymbol{\Sigma}_{nxn}$

The element in row i and column j of $\boldsymbol{\Sigma}_{nxn}$ is the "covariance" between data observation i and data observation j. This covariance is typically denoted as σ_{ij}

Correlation, ρ, is defined to be a standardized covariance, $\rho = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$

By definition, when $\sigma_{ij} = 0$, observation i and data observation j are independent (i.e., not correlated, $\rho = 0$).

By definition, a covariance matrix, $\boldsymbol{\Sigma}_{nxn}$, is symmetric about the main (northwest to southeast) diagonal because $\sigma_{ij} = \sigma_{ji}$

Covariances on the main diagonal are more commonly referred to as variances, $\sigma_{ii} = \sigma_i^2$ (for data observation i).

General Spatial Model

Focus is on modeling Small-Scale Variability
when there is dependence or correlation among observed residual values.

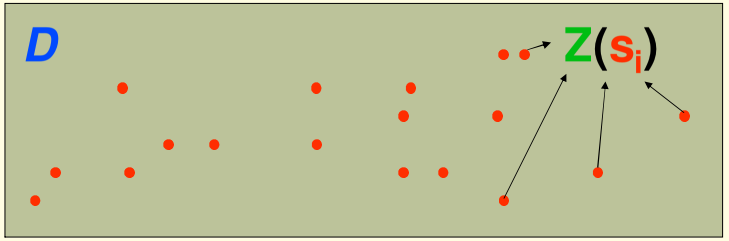
Correlation
implies
$\boldsymbol{\Sigma}_{nxn}$ is not diagonal

$$\boldsymbol{\Sigma}_{nxn}=\begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{21} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}_{nxn}$$

Notes

The key to successfully modeling the small-scale variance for spatially-correlated data is to accurately identify the relationship between the proximity or distance among data points and their correlation to one another.

General Spatial Model
Terminology



D is the spatial domain or area of interest

s_i notates the spatial coordinates

Z is a characteristic of interest measured or observed at the spatial coordinates

Notes

We make a shift in notation here from that used by the traditional general linear model to that used by spatial models.

Characteristic of Interest:

Y	for traditional general linear models
Z(s)	for spatial models; s denotes the location of the measurement.

Spatial Auto-Correlation
A Definition

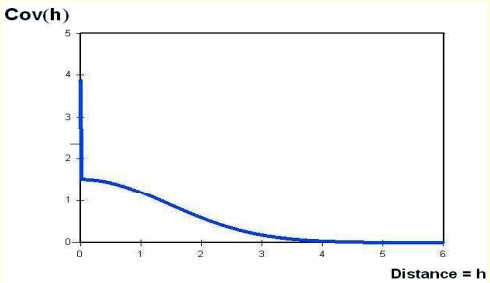
A measure's **correlation with itself** relative to proximity/location.

Data values observed at n locations are **auto-correlated**
when values $Z(s_i)$ and $Z(s_j)$ in close proximity to one another $|s_i - s_j| < h$
are more alike than values located at a further distance $|s_i - s_j| > h$.

As the distance, **h**,

increases between 2 data observations, s_i and s_j,

the correlation between $Z(s_i)$ and $Z(s_j)$ decreases.



Semivariance – A Statistic for Measuring Autocorrelation

Semivariance Formula:

$$\gamma(s_i, s_j) = \tfrac{1}{2}\,\mathbf{Var}[Z(s_i) - Z(s_j)]$$

$$= \tfrac{1}{2}\{\ \mathbf{Var}[Z(s_i)] + \mathbf{Var}[Z(s_j)] - 2\cdot\mathbf{Cov}[Z(s_i), Z(s_j)]\ \}$$

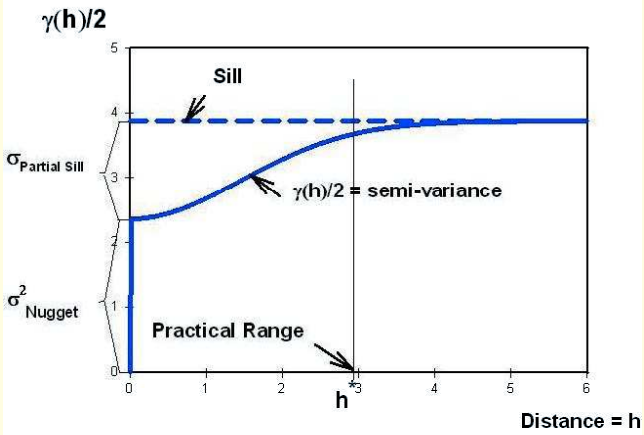
Notes: Semivariance

Semivariance is a statistic and a function that facilitates examining the relationship between the covariance (i.e., correlation) between the characteristic of interest, Z, and the locations where it was measured, s_i and s_j.

The "Auto-Correlation Definition" slide, above, clearly exhibits decreased covariance (and correlation) with increased distance between s_i and s_j.

We will see in a few subsequent slides ("Required Assumptions for Modeling Spatial Data") that under the assumptions of "stationarity", only the distance between observed data points is important to allow accurate estimation of the semivariance (and hence, covariance and correlation).

Semivariogram – A Tool for Measuring Autocorrelation



Notes: Semivariogram

A semivariogram, upon initial consideration, may not be as intuitively interpretable as the covariance plot on the "Spatial Auto-Correlation" slide above.

Most readily, a semivariogram provides the Practical Range, h*, which indicates the distance between any two points in the observed process beyond which those two points are independent of one another (i.e., not correlated).

Also, for any distance ($h=s_i-s_j$), $Cov[Z(s_i), Z(s_j)] = Cov[h] = sill - \gamma(h)/2$.

The component parts of a semivariogram can be interpreted as:

$sill = \sigma^2_{nugget} + \sigma_{partial\ sill} = Var[Z(s_i)]$

where $\sigma_{partial\ sill}$ is the portion of $Var[Z(s_i)]$ due to variation in Z

σ^2_{nugget} is the portion of $Var[Z(s_i)]$ due to measurement error or small-scale variation in the process

Effective Sample Size in Presence of Autocorrelation

"…*positive autocorrelation results in 'loss of information'.*

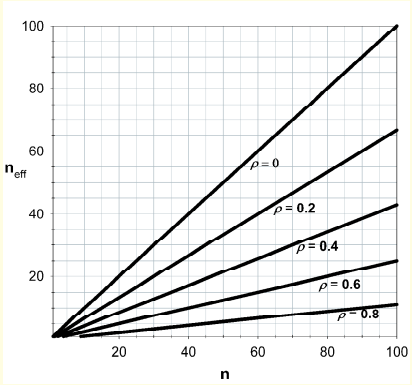
$$n_{effective} = \frac{n_{corr} \cdot (1-\rho)}{(1+\rho)}$$

$n_{effective}$ = uncorrelated (independent) samples

n_{corr} = correlated (dependent) samples

where ρ is autocorrelation

with $0 \le \rho \le 1$.



Notes

The more strongly spatial data are correlated, the less "unique" information is provided by each individual observed data point.

Information shared by data points in closer proximity can improve the ability to accurately model the characteristic of interest, Z.

Simultaneously, strongly correlated data points can reduce the statistical power of inferences (i.e., hypothesis tests).

The effective sample size formula (on the previous slide) results from the assumption (Cressie 1991, p.14-15) that

$$Cov[Z(s_i),Z(s_j)] = \sigma^2 \cdot \rho^{|si-sj|}$$

or equivalently $Cov[h] = \sigma^2 \cdot \rho^h$ where $h=|s_i-s_j|$

Notes

The below table illustrates how correlated data contains less unique information than independent data. For example, if two data points are located at a distance from one another that causes them to have a correlation of ρ=0.2, observing n_{corr} data points provides information equivalent to the amount provided by two-thirds fewer independent (i.e., uncorrelated data points.

ρ	$n_{effective}$
0	n_{corr}
0.2	⅔ n_{corr}
0.5	⅓ n_{corr}
0.8	1/9 n_{corr}
1	0

Autocorrelation Influences Statistical Inference

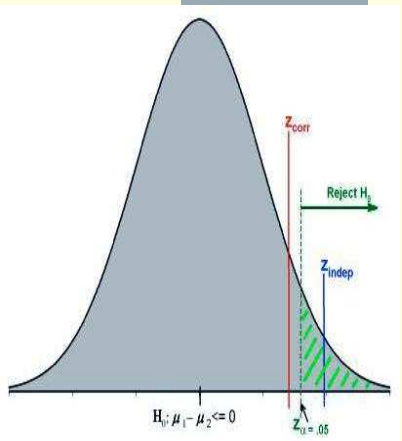
$$Z_{indep} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \cdot \sqrt{\frac{2}{n}}}$$

$$Z_{corr} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \cdot \sqrt{\frac{2 \cdot (1+\rho)}{n \cdot (1-\rho)}}}$$

If positive autocorrelation is present and ignored, a treatment effect can be incorrectly declared significant.

Divisor: n for z_{indep}

$n_{effective}$ for Z_{corr}



Notes

Example:

On the previous slide, a hypothesis test for the equality of 2 treatment means has a divisor of n when the data values for the 2 treatments were independently replicated (i.e., not correlated). In this case, the test statistic $z_{indep} > z_{\alpha=.05}$ and there is sufficient evidence to reject H_0 and declare a significant difference between the treatment means.

However, if the data values observed for the 2 treatments are correlated with one another, the divisor for the test statistic, z_{corr}, is $n_{effective} = \frac{n_{corr} \cdot (1-\rho)}{(1+\rho)}$

which is smaller than the divisor in the independent case.

Hence, $z_{corr} < z_{\alpha=.05}$ so there is an insufficient amount of data (i.e., statistical power) to reject H_0.

Spatial Data

- has **no** independent replications
- consists of a **single** n-dimensional observation:

  $\{ Z(s_1),\ldots, Z(s_n) \}$ at locations $s_1,\ldots,s_n$
- estimates dependency, $\sum$, via semivariance=$\gamma(s_i,s_j)$ using:

  1) the observed $\{ Z(s_1),\ldots, Z(s_n) \}$

  and 2) distances, h, between the $s_1,\ldots,s_n$
- predicts $Z(s_0)$ at an unobserved location, s_0, using the observed $\{ Z(s_1),\ldots, Z(s_n) \}$ and the estimated semivariance=$\gamma(s_i,s_j)$

Required Assumptions for Modeling Spatial Data

Stationary Process

Constant Mean: $Z(s_i)=\mu$ for all s_i in D

Covariance is function of distance ($h=s_i-s_j$), NOT location (s_i):

$Cov(s_i - s_j)$ NOT $Cov(s_i)$

Notes

The validity of all statistical models requires that the data meet some basic assumptions. Typical spatial models require that the data possess characteristics of a "Stationary Process", as defined on the previous slide. If the data do not represent a stationary process, the fitted spatial model will produce incorrect predictions and/or inferences.

Spatial models require that the characteristic of interest, Z, have a constant mean value over the entire domain. This can typically be achieved by modeling the large-scale effects and use the residual variability as the spatial data to which a spatial model is fit.

Required Assumptions for Spatial Data Modeling

The water level of
a calm pond
during
a light rain shower
is an example
of a stationary process:

Photo "Raindrops on the Pond"
by Mark Schretlen 11-May-2003

Notes

The water level on the surface of a pond in a light rain shower is a natural phenomenon that illustrates a stationary process:

1. The average water level is constant over the entire pond surface

2. The water level within a radius from the point where the rain drop strikes the surface depends on the water level at all other locations within that radius. Since the intensity of rain is similar across the entire surface of the pond, the correlation of water levels within the radius is the same regardless of where the rain drop hits the surface of the pond and the strength of correlation within the radius depends only upon the distance from the raindrops point of impact.

Definition: Kriging

Predict unobserved $z(s_0)$ as a weighted average of the observed $z(s_1),\ldots,z(s_n)$ spatially-correlated data

Σ and h (i.e., distance) determine the kriging weights assigned to each of the observed $z(s_1),\ldots,z(s_n)$ in the kriged estimate, $\hat{z}(s_0)$

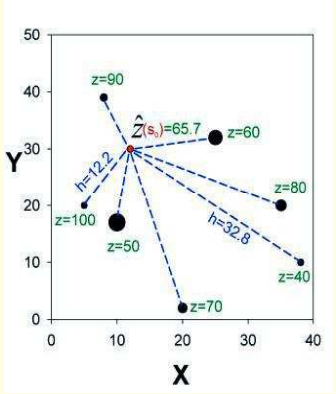
The term Kriging was coined by G. Matheron(1963) in honor of South African mining engineer D.G. Krige, whose work (1951) laid preliminary groundwork for the field of "geostatistics".

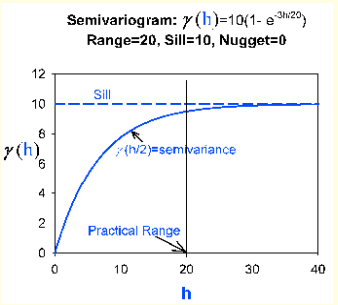
Semivariance determines Kriging Weights

Range=20, Sill=10, Nugget=0

Kriged Estimate, $\hat{z}(s_0)$, at s_0 = (x=12,y=30) is 65.7

$\hat{z}(s_0)$ is a weighted average of the observed $z(s_i)$. The weights sum to 1.
Each point on the graph is sized proportionately to its weight.



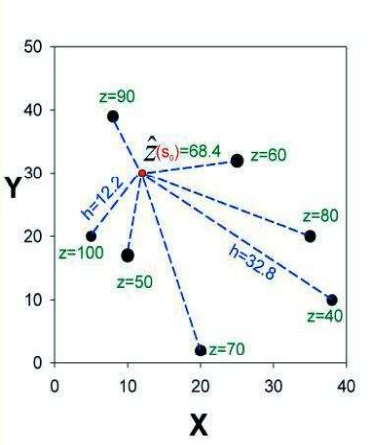


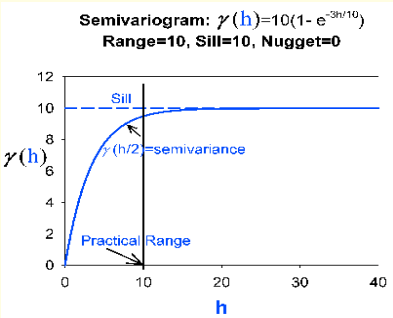
Notes

The previous and next several slides use Isaaks' and Srivastava's (1989, pp. 291, 301-307) small data set of seven observations and one prediction location to examine the effect of semivariogram parameter on ordinary kriging predictions. This example was also given as Example 5.5 in Schabenberger & Gotway (2005).

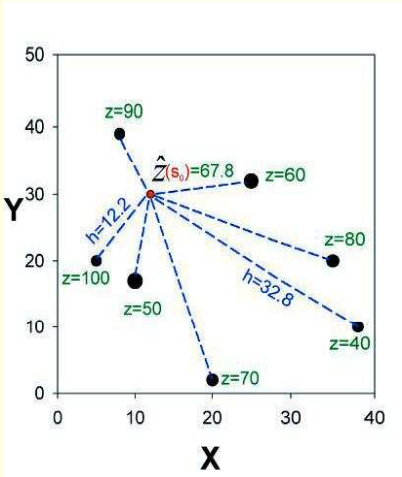
The only difference between the previous and the next semivariogram is the range. The larger practical range in the previous slide causes greater "short-distance" correlations, which results in greater heterogeneity in the weights used to obtain the kriged estimate.

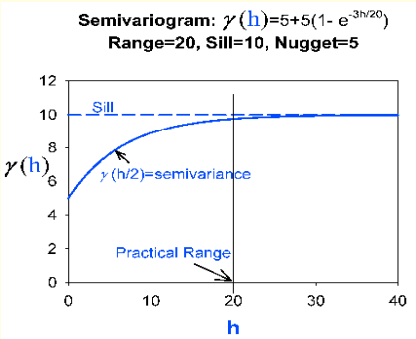
Practical Range Changes from 20 to 10
Kriged Estimate, $\hat{z}(s_0)$, at s_0 = (x=12,y=30)
Changes from 65.7 to 68.4





Nugget changes from 0 to 5
Kriged Estimate, $\hat{z}(s_0)$, at s_0 = (x=12,y=30)
Changes from 65.7 to 67.8

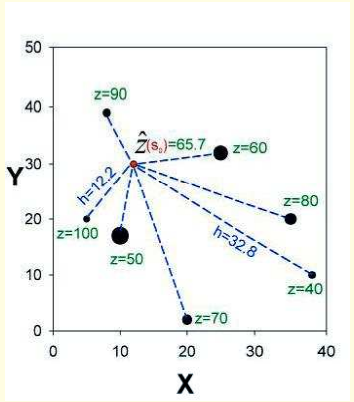


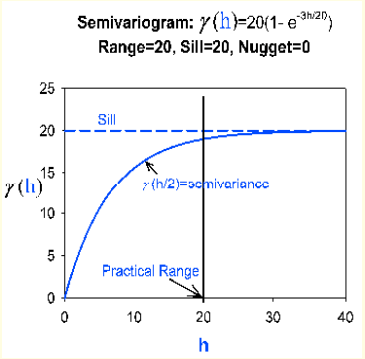


Notes

Introduction of a nugget effect yields more homogeneous kriging weights, similar to the kriging weights resulting from the doubling of the practical range.

Sill Doubles from 10 to 20
Kriged Estimate, $\hat{z}(s_0)$, at s_0 = (x=12,y=30)
Remains Unchanged at 65.7





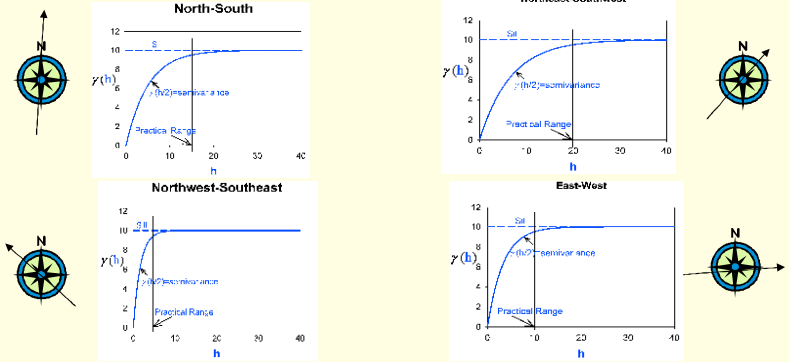
Notes

Compared to the first kriged estimate that used an exponential semivariogram with Range=20, Sill=10, Nugget=0; the previous slide used an exponential semivarigram with Sill=20. Doubling of the sill did not change the kriging weights at all. The larger sill caused only a larger the kriging variance (i.e., variance of Z(s))

Direction in Spatial Modeling

Isotrophy – autocorrelation is equivalent in all directions

Anisotrophy – autocorrelation is direction dependent.



Note

If distance correlations change depending on direction, the appropriate semivariogram for the spatial model also changes with direction. In this case, direction must be considered when fitting the model.

3 Types of Spatial Models

- Geostatistical / Point-referenced
- Lattice / Areal
- Point-Process / Point-Pattern

Notes

The majority of the information presented thusfar most readily lends itself to geostatistical data. However, the general concepts apply (with appropriate adjustments or modifications) to all 3 types of spatial models.

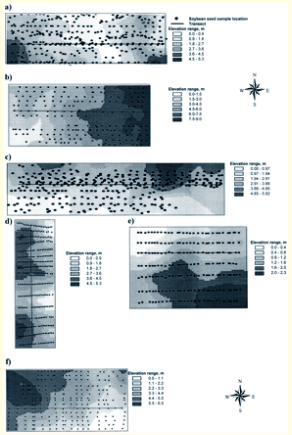
Geostatistical / Point-referenced Models

Specific locations $s_1,\ldots,s_n$ in the domain **D** are selected.

The characteristic of interest, $z(s_1),\ldots,z(s_n)$, is observed.

Example: Six fields, each planted in a different soybean cultivar.
- Locations $s_1,\ldots,s_n$ are n individual soybean plants.
- $z(s_1),\ldots,z(s_n)$ are protein concentration of the plant's yield.

Crop Science 42:804-815 (2002), A. N. Kravchenko and D. G. Bullock



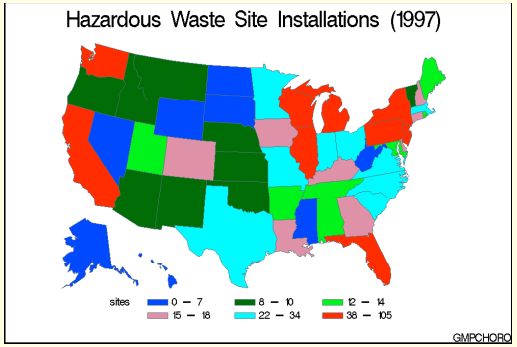
Notes

The figures on the right of the previous slide illustrate how kriging can produce prediction maps.

Lattice / Areal Models

Specific locations $s_1,\ldots,s_n$ represent 'contiguous areas' in the domain **D**. The characteristic of interest, $z(s_1),\ldots,z(s_n)$, is observed for each 'area'.

Example: # of hazardous waste sites in each U.S. state.



SAS/Graph Online Documentation – Proc GMAP

Notes

Lattice or Areal models have the objective of predicting Z(s) where s is an "area" rather than a "point", as in the Geostatistical/point-referenced model case.

Two Methods of Modeling Lattice Data

- Simultaneously Autoregessive
  - Likelihood methodology
- Conditionally Autoregressive
  - Gibbs sampling (Bayesian) methodology

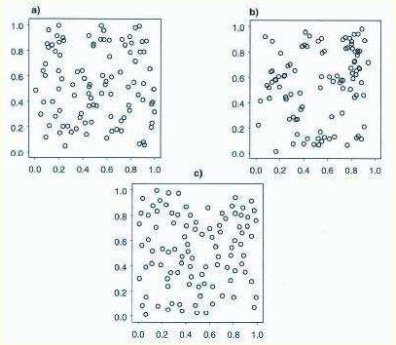
Schabenberger & Gotway (2005) pg.7

Point-Pattern Models

Objective:

Model the 'process' that generated the spatial data.

Fig a) completely random pattern
Fig b) Poisson cluster process
Fig c) process with sequential inhibition regularity

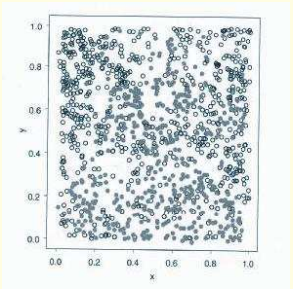


Schabenberger & Gotway (2005), p. 82

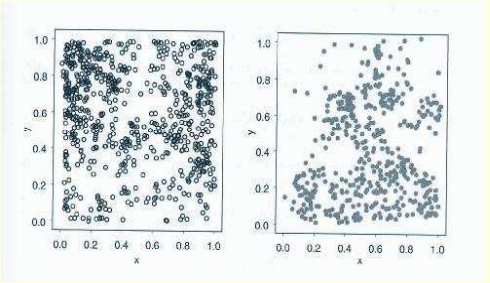
Point-Pattern Data
Example 1: A Marked Process

Distribution of hickory(○) and maple trees(●).

Overlaid



Separate Plots



Schabenberger & Gotway (2005), p.119,121

Notes

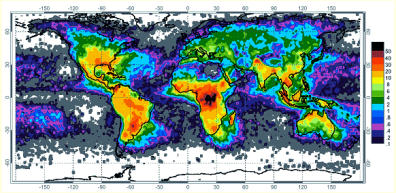
The "mark" in this marked process is whether the species of tree is hickory or maple.

Point-Pattern Data
Example 2: Lightning Strikes

Lightning strikes within 200 miles of the U.S. east coast April 17-20, 2003.

Schabenberger & Gotway (2005) p.13

Kriged predictions can also be obtained for point-pattern data, as shown by the NASA map of global lightning strikes.



Summary of the 3 Model Types

Analogy: A desktop is the domain **D** of locations $\mathbf{s}_i$
Experiment → pour sand on the desktop.

Geostatistical & Lattice Data:

- locations $\mathbf{s}_i$ do not change from one pouring (i.e., experiment) to the next
- $z(\mathbf{s}_i)$= observed sand depth varies at $\mathbf{s}_i$

Point-Pattern:

- specify a sand depth of interest
- observe all locations $\mathbf{s}_i$ in **D** where sand has this depth.

Schabenberger & Gotway (2005)

Data Measured at Multiple Scales

"Even when the disorder is discovered to have a perfectly rational explanation at one scale, there is very often a smaller scale where the data do not fit the theory *exactly*, and the need arises to investigate the new, residual uncertainty."

Cressie (1991)

Hierarchical Models

- Estimate 'parameters' of an experiment using the observed data $z(s_1),\ldots,z(s_n)$
- Assume and impose statistical distributions on the parameters to be estimated
  - distribution choices rely on theory and/or scientific knowledge
  - modeling of distributions uses Bayesian methods
    - GEOBUGS freeware

References

- Schabenberger & Gotway (2005)
  Statistical Methods for Spatial Data Analysis
- Banerjee, Carlin, & Gelfand (2004)
  Hierarchical Modeling and Analysis for Spatial Data
- Cressie (1991)
  Statistics for Spatial Data

Differing World Views in Modeling

Deterministic vs Stochastic

Mary J. Camp

Biometrical Consulting Service
USDA, ARS, Beltsville

Begin at the Beginning

- Observation – a record obtained by an act of recognizing and noting a fact or occurrence often the outcomes of an experiment, investigation, or survey and measuring with instruments.
- Data – a collection of observations. Factual information (as measurements) used as a basis for reasoning, discussion, or calculation.
- What to do with data? Investigate how it came about, what caused it, manipulate conditions to produce it, use it to make predictions.
- To do the above to data usually means – Model It

World Views in Modeling

- How data is modeled and the purpose of modeling will depend on the modeler's world view.
- Deterministic Model (Functional Model)
- Stochastic Model (Statistical Model)
- Classical Statistical Model (General Linear Model)
  – actually a subset of the Stochastic Model

Classical Statistical Modeling

- An observation is thought of as being composed of three parts: a part due to the average of all observations in the population, a part due to manipulation, the level of an applied factor(s), i.e., a treatment(s), and a part due the unique properties of that particular observation in the population

$$Y_{ij} = \mu + \tau_j + \epsilon_{ij}$$

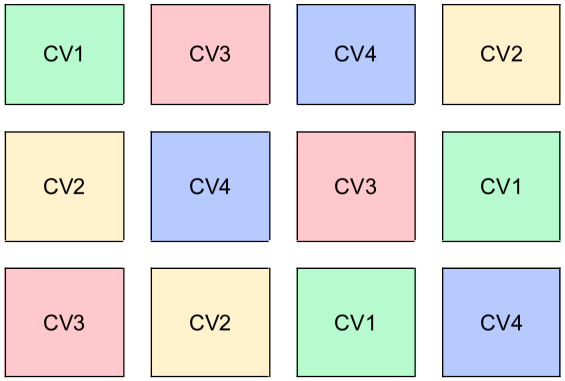
- Rearranging shows that the deviation of an observation from the overall mean is then due to the effect of its factor level, the treatment effect, and its other properties, the error

$$Y_{ij} - \mu = \tau_j + \epsilon_{ij}$$

Assumptions of the Model

- The observations are independent. Measuring or observing one does not affect the measurement or observation of another.
- The error is a sample from a probability distribution. Often in modeling this is a normal distribution, also known as the bell-shaped curve.
- The errors for the observations come from the same probability distribution.

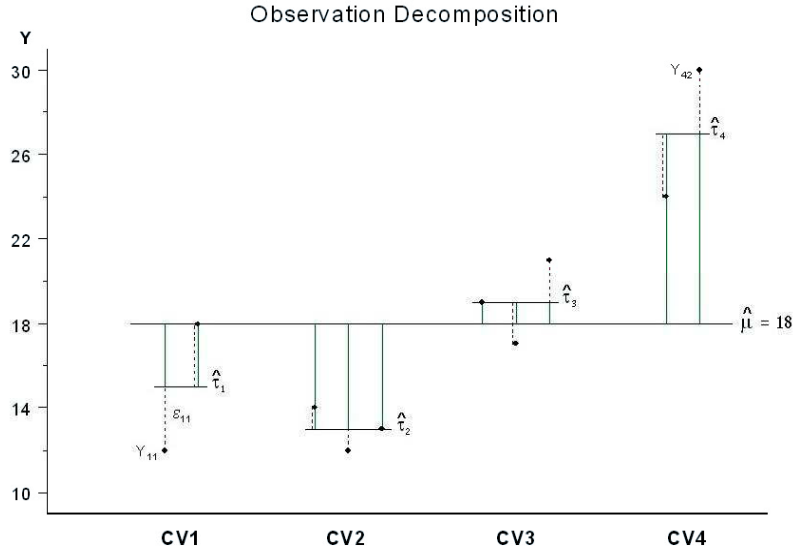
Example 1: High Tunnel Tomato Yield



	Cultivar				
Plot	1	2	3	4	
1	12	14	19	24	
2	18	12	17	30	
3	—	13	21	—	
					Total
Total	30	39	57	54	180
Mean	15	13	19	27	18
Number of Plots	2	3	3	2	10

An observation, the yield for a plot, is viewed as being composed of the average yield of tomatoes in the high tunnel plus an effect due to the cultivar on the plot and an effect due to the individual differences intrinsic to each plot.

$$Y_{ij} = \mu + \tau_j + \epsilon_{ij} \qquad i = 1\ldots3,\ \ j = 1\ldots4$$



Determining if Treatment is Important

- For each observation the square of the distance of the estimated treatment effect from the estimated overall mean is calculated. In the above plot would be squaring and summing the solid green lines. The sum of these distances is known as the *sum of squares treatment*.
- The squared distance of the each observation from the estimated treatment effect is calculated. In the above plot this would be squaring and summing the dotted red lines. The sum of these distances is known as the *sum of squares error*.
- We look at a ratio of the average sum of squares treatment and the average sum of squares error.
- If the ratio is large enough, then we judge that the treatment has an important effect in understanding the differences between the means of the treatment levels.

High Tunnel Tomato Yield Analysis

- The average sum of squares treatment, i.e., the cultivar effect: 258/3 = 86
- The average sum of squares error, i.e., the mean square error: 46/6 = 7.67
- The ratio is: 86/7.67 = 11.21
- Under the assumptions of the model and that the probability distribution for the error is the normal distribution, 11.21 is large enough. The probability of obtaining a ratio this large if the average cultivar yields were not different is only .007. The conclusion is that the cultivar is important in explaining the differences in the average tomato yields.

	Cultivar 1	Cultivar 2	Cultivar 3	Cultivar 4
Average Yield	15	13	19	27

Deterministic (Functional) Model

- Mathematical function(s) is used to model a process, usually chemical or physical.
- Observations or predictions are the results of how the inputs interact in the process.
- The model can be very complex however the more complex the model the more inputs, *parameters*, and terms are needed for prediction.
- The model is only as good as the science used to make it. Assumes the process is understood and the data for it can be collected.
- By changing any of the inputs, any of the values of the parameters, new predictions and “What if?” questions can be asked.

Example 2: Return on an Investment

$$F = P(1 + r/m)^{Ym}$$

Where:

F = Future value
P = Present value,
r = Annual rate,
m = Periods/Year,
Y = number of Years

5-Year Return on $1000 at Federal Funds Rate, June 2004 – January 2006

Rate	Return	Rate	Return	Rate	Return	Rate	Return
1.25	1064.46	2.25	1118.95	3.25	1176.19	4.25	1236.30
1.50	1077.83	2.50	1133.00	3.50	1190.94	4.50	1251.80
1.75	1091.37	2.75	1147.22	3.75	1205.88		
2.00	1105.08	3.00	1161.62	4.00	1221.00		

Example 3: Verhulst-Pearl Logistic Growth

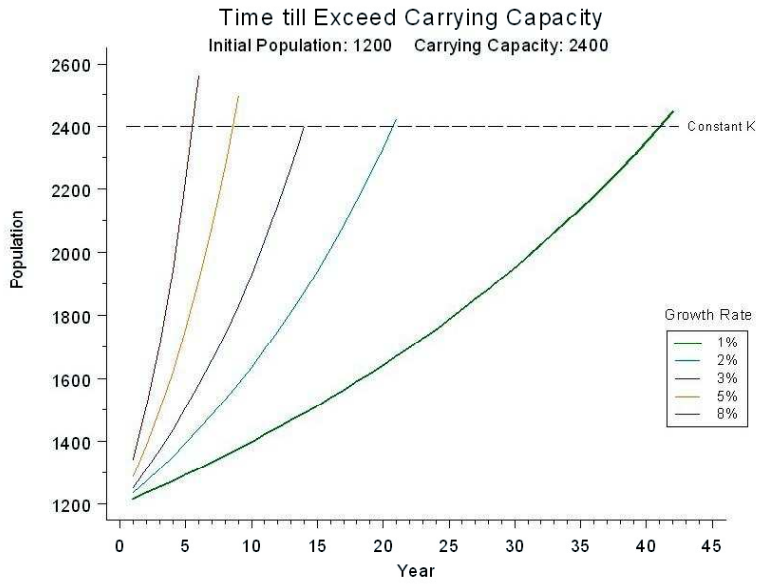
$$N_{t+1} = N_t + rN_t(1 + N_t/K)$$

Where:

N_t is the population size at time t,

r is the growth rate of the population,

K is the carrying capacity of the habitat



Stochastic (Statistical) Model

- Uses mathematical function(s) to model a process.
- At least one model parameter is a random variable described by a probability distribution.
- Ability to reproduce and predict observations are based on patterns of previous data, not necessarily the underlying physical or chemical processes.
- Some view these models as 'black-box' models.

Example 4: Verhulst-Pearl Logistic Growth

$$N_{t+1} = N_t + rN_t(1 + N_t/K_{f(t)})$$

Where:

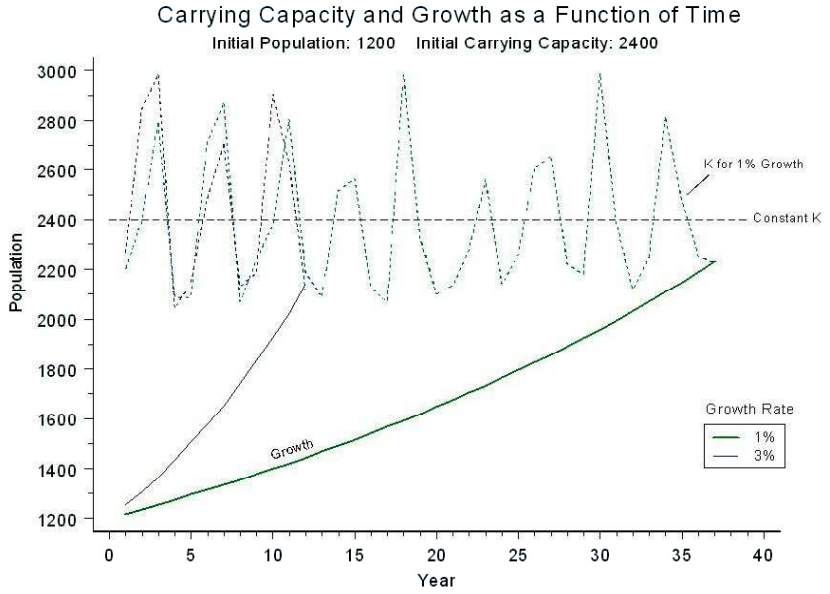
The carrying capacity at time t, $K_{f(t)}$, is the initial carrying capacity multiplied by a random variable, *p*. $K_{f(t)} = K_0 p$

In years 1,4,5,8,9,12... p has an equal chance of taking any value between 0.85 and 0.95.

In years 2,3,6,7,10,11... p has an equal chance of taking any value between 0.95 and 1.25

Mathematically this is

$$\begin{cases} 0.85 \le p \le 0.95 \text{ uniformly distributed when } t = 1,4,5,8,9,12... \\ 0.95 \le p \le 1.25 \text{ uniformly distributed when } t = 2,3,6,7,10,11... \end{cases}$$



Example 5: Number of Marriage Licenses

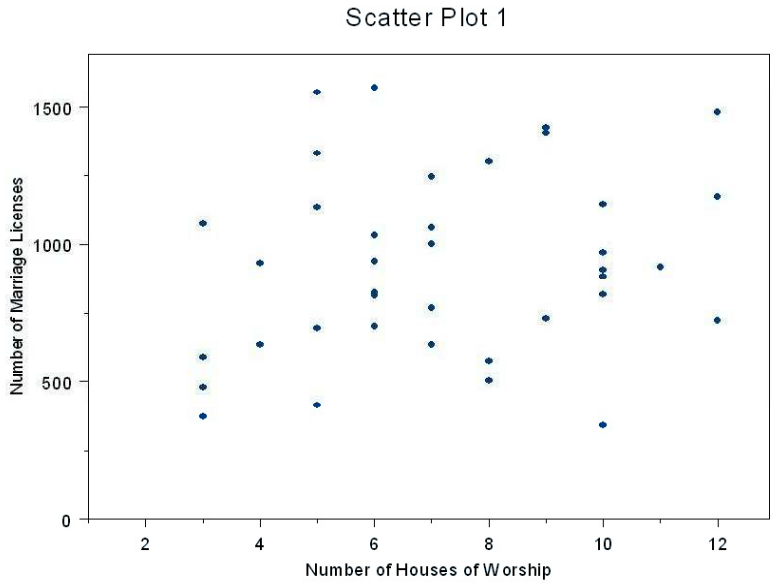
In an attempt to model the number of marriage licenses issued in March from 38 randomly selected county courthouses, a linear regression model was used.

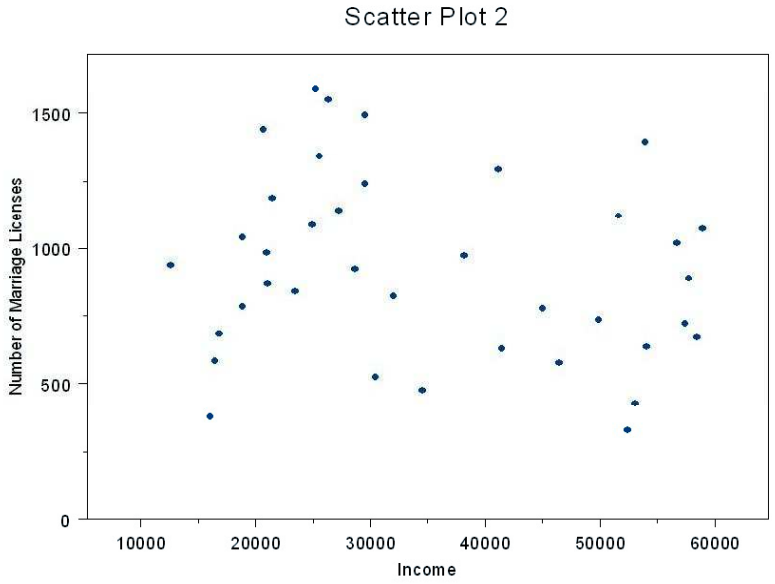
The input variables were:

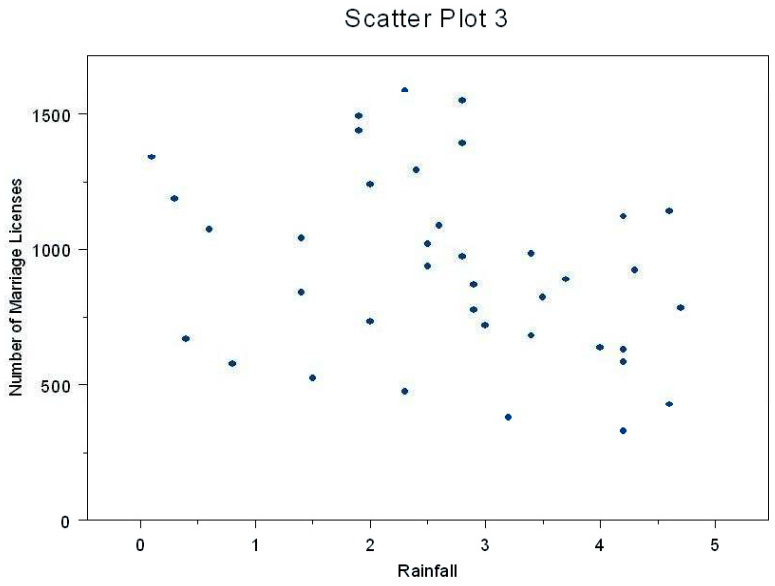
Income	=	average household income in the county.
Liquor	=	number of liquor stores within a 10 block radius of the courthouse.
Rain	=	rainfall in inches for the county in March.
Robins	=	number of robins reported in the county's March bird survey.
TV	=	average number of television sets per household in the county.
Worship	=	number of houses of worship within a 3 mile radius of the courthouse.

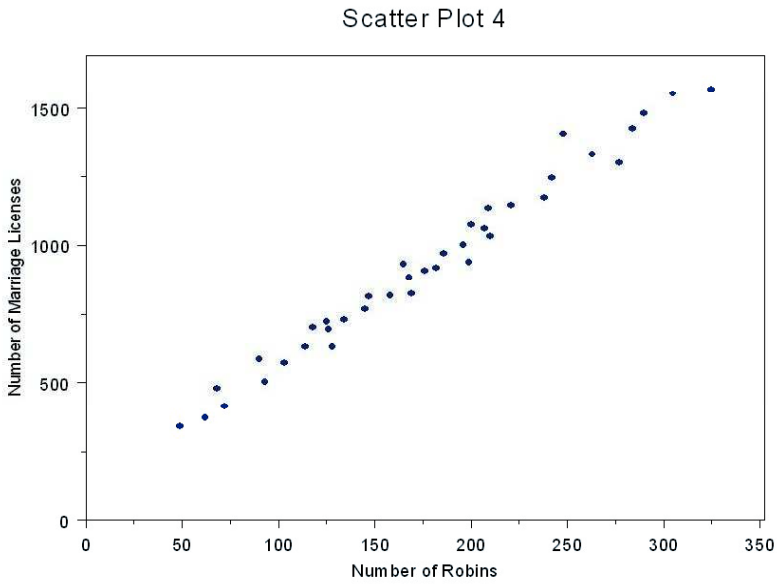
Result

Marriage Licenses = 108.90 + 4.63(Robins)









Comparing Models

- Deterministic and Stochastic models look the same when the deterministic model has model error, a random part.
- Deterministic model with error

$$Q = f_d(P|\Omega) + \epsilon$$

Q is the observation based on a function, $f_d(\cdot)$, of the process P, for the model parameters, Ω, and ϵ is the model error.

- $f_d(P|\Omega)$ is the deterministic component and ϵ is the stochastic part.
- The model error can occur either through mis-specifying the model, i.e., leaving out factors that explain the process, or measurement error.
- Goal is usually to minimize ϵ by some means and focus on the deterministic element.

- Stochastic model with error

$$Q = f_s(P|\Omega) + \nu$$

Q is the observation based on a function, $f_s(\cdot)$, P is the inputs, for the model parameters, Ω, and ν is the model error.

- $f_s(P|\Omega)$ is the deterministic component and ν is the stochastic part.
- The model error occurs because the model is based on mathematical function(s), measurement error and unexplained variability.
- In the stochastic model the deterministic element, $f_s(P|\Omega)$, is derived to insure reproduction of characteristics of Q without regard to underlying physical processes.

- The Stochastic model's weakness is that it is does not necessarily represent observed internal physical laws or processes. Consequentially, the model may not be useful in understanding how the observations occur.

- The Deterministic model's weakness is that it can not reproduce the variance of observed model outputs. As long as the model residuals (observed value – predicted value) are independent of the model inputs

$$Var[Q] = Var[f_d(P|\Omega)] + Var[\epsilon]$$

it will always hold that $Var[f_d(P|\Omega)] < Var[Q]$, unless $Var[\epsilon] = 0$ which means there is no model error.

Relationship to Spatial Modeling

Spatial models comprise two sources of variation:
Large Scale Variation (modeling the mean structure) and
Small Scale Variation (modeling the covariance structure).

Large Scale Variation = Trend

- Involves the entire region of the study or experiment area.
- All points are used equally to predict an observation.
- In a deterministic model this would be the functional part that describes a process, e.g., modeling how fast water flows down a slope.
- In a stochastic (statistical) model this would be the treatments in an analysis of variance, independent variables in a regression, blocks.

Small Scale Variation

- Once large scale variation has been removed, only neighboring points are used to estimate a nearby observation.
- Observations are viewed as being correlated. Observations close together are more correlated than observations further apart. As observations become further apart a distance is reached where the correlation is negligible.
- A perfect deterministic model would have no small scale variation. That these models do have small scale variation is largely a matter of measurement error.
- A statistical model will have small scale variation, since the model is based on mathematical functions and proxy variables that do not fully explain the process, plus it will have measurement error.

Some physiologists will have it that the stomach is a mill; --others, that it is a fermenting vat;--others again that it is a stew-pan;--but in my view of the matter, it is neither a mill, a fermenting vat, nor a stew-pan--but a *stomach,* gentlemen, a *stomach.*

William Hunter 1718–1783

References

J.L. Cisne. How Science Survived: Medieval Mansuscripts' "Demography" and Classic Texts' Extinction. *Science*, **307**:1305–1370 (2005).

J. W. Hayse and I. Hlohowskyj. Comparison of Deterministic and Monte Carlo Analyses for Evaluating Risks to Ecological Receptors With Contaminant Uptake Models. (1998) http://web.ead.anl.gov/jfield/PPT_presentations/ArmyPresentation/index.htm

A. J. Lembo, Jr. Lecture 3: Model Use and Development Spatial Modeling and Analysis. www.css.cornell.edu/courses/620/lecture3.ppt

J. Neter and W. Wasserman. *Applied Linear Statistical Models*. Richard D. Irwin, Inc, Homewood, 1974

O. Schabenberger and C. A. Gotway. *Statistical Methods for Spatial Data Analysis*. Chapman and Hall/CRC Press, Boca Raton, 2005

R. M. Vogel. Stochastic and Deterministic World Views. *Journel of Water Resources Planning and Management*, **125(6)**: 311–313 (1999)

An Introduction to Statistical Models for Spatial Data in Ecology

By

Jay M. Ver Hoef

National Marine Mammal Lab

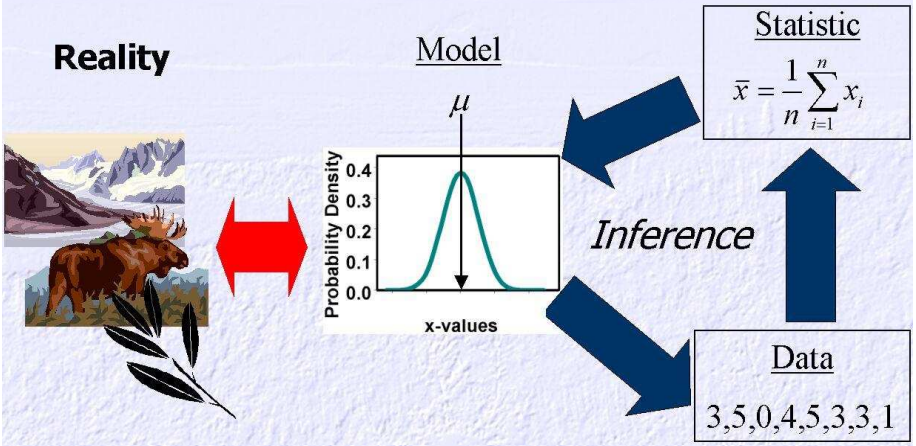
7600 Sand Point Way, NE

Seattle, WA 98115

jay.verhoef@noaa.gov

1

What do Statisticians Do?



2

What is a Model?

What does it look like?

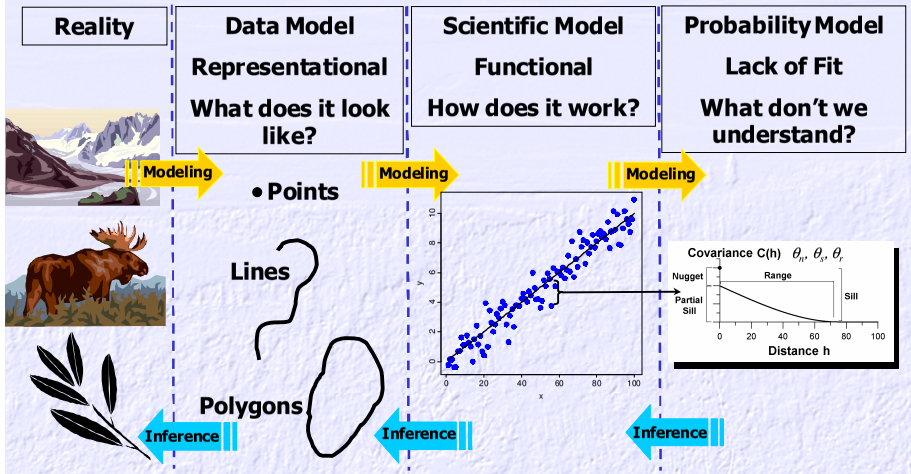
Representational

How does it work?

Functional

3

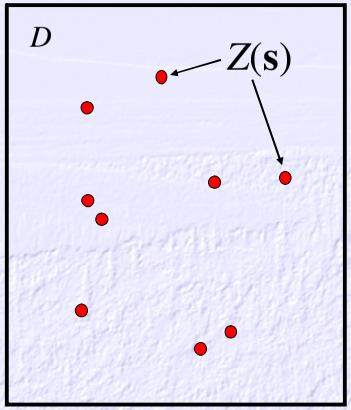
What are Spatial Statistics



"All models are wrong. We make tentative assumptions about the real world which we know are false but which we believe may be useful."- George Box 1976

4

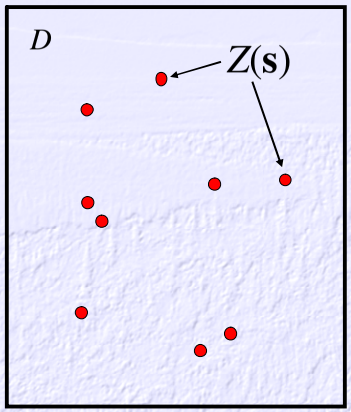
Notation



- D is the spatial domain or area of interest
- **s** contains the spatial coordinates
- Z is a value located at the spatial coordinates

5

Types of Spatial Data



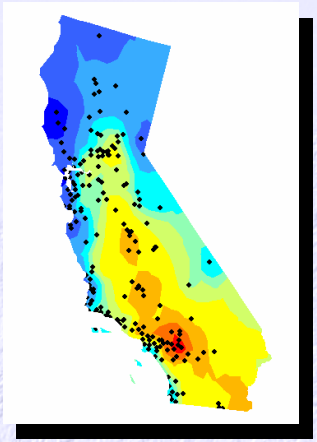
$\{Z(\mathbf{s}): \mathbf{s} \in D\}$

- **Geostatistical Data**: Z random; D fixed, infinite, continuous
- **Lattice Data**: Z random; D fixed, finite, (ir)regular grid
- **Point Pattern Data**: $Z \equiv 1$; D random, finite

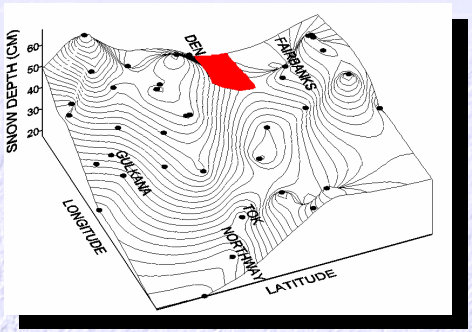
6

Examples of Geostatistical Data

Ozone Predictions



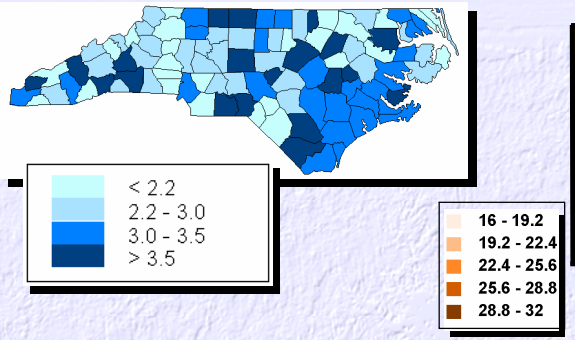
Average Snow Depth



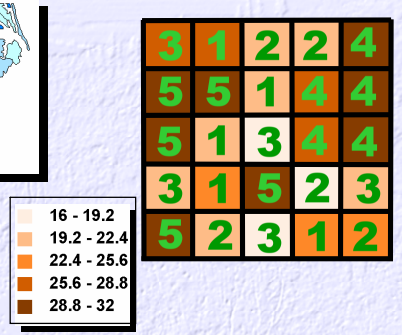
7

Examples of Lattice Data

Transformed SIDS rates



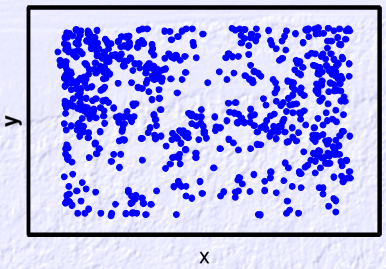
Plots in a Designed Experiment



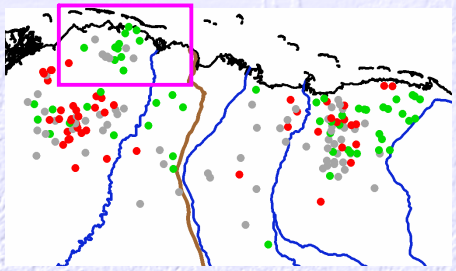
8

Examples of Point Patterns

Lansing Woods Hickory Locations



Arctic Caribou Calving Locations



9

Statistical Models

Linear Model

$$\begin{pmatrix} \mathbf{z}_{observed} \\ \mathbf{z}_{unobserved} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathrm{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

Nonlinear Model

$$\begin{pmatrix} \mathbf{z}_{observed} \\ \mathbf{z}_{unobserved} \end{pmatrix} = g(\mathbf{X}, \boldsymbol{\varepsilon}, \boldsymbol{\theta})$$

Prediction **Estimation**

10

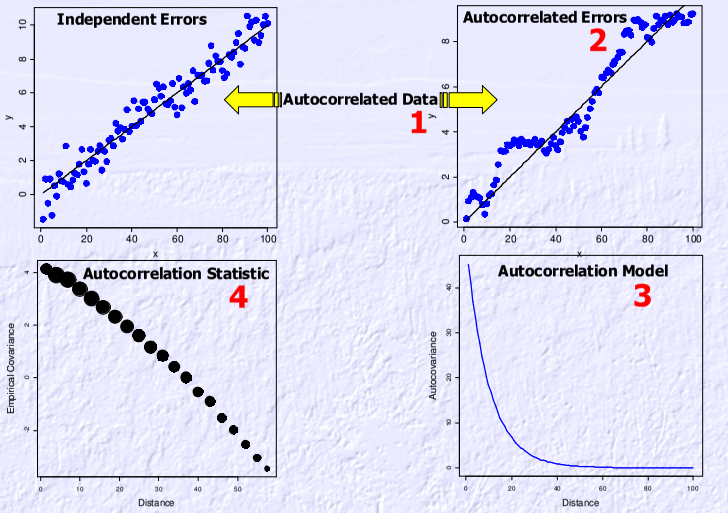
Five Meanings of Autocorrelation

- Description of data
- Property of a stochastic process
- Model for a stochastic process
- Statistic
- Function in Fourier analysis

11

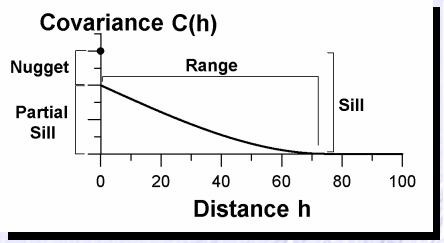
Four Meanings of Autocorrelation

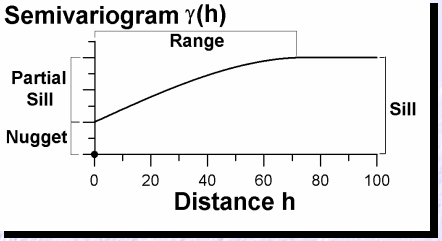
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$ $\qquad$ $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ \mathrm{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$

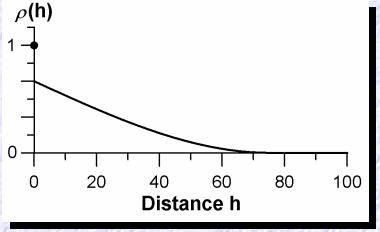


12

Autocorrelation Models

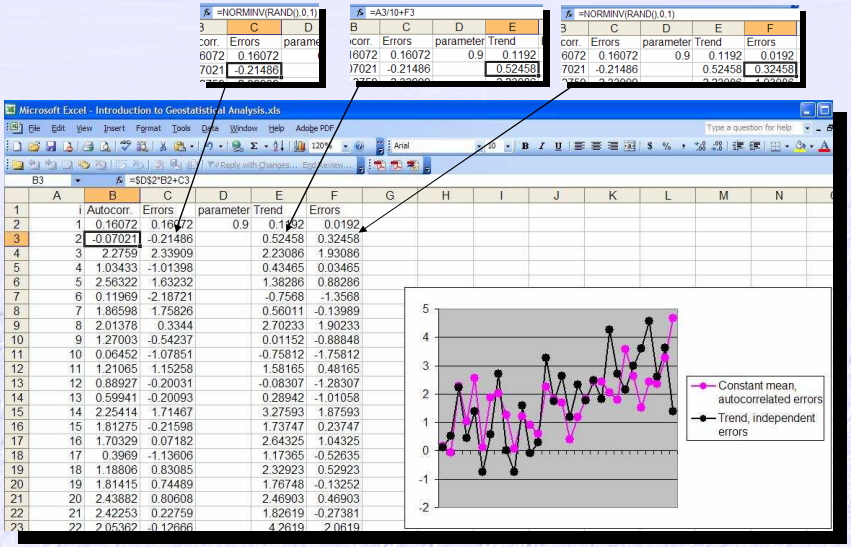






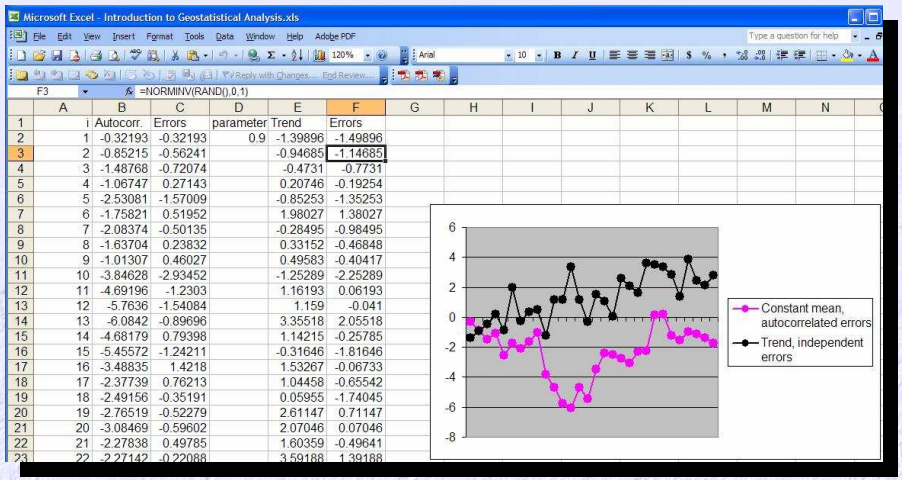
13

Autocorrelation



14

Try it! F9

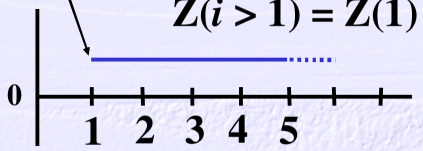


15

Why Spatial Statistics?

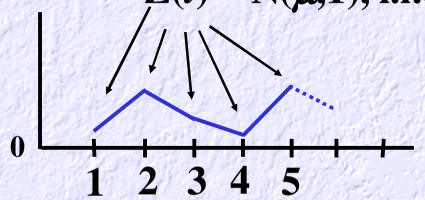
$$Z(1) \sim N(\mu,1)$$

$$Z(i>1) = Z(1)$$



$$\Sigma = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

$$Z(i) \sim N(\mu,1), \text{ i.i.d.}$$

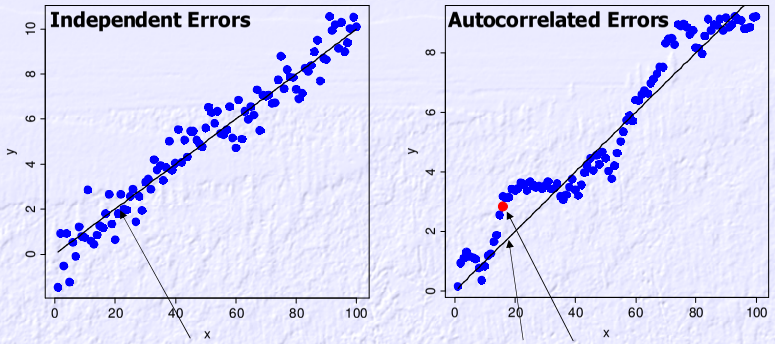


$$\Sigma = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

16

Fits vs. Prediction

$$\begin{pmatrix} \mathbf{z}_{observed} \\ \mathbf{z}_{unobserved} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \ \text{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$



Variances different in both cases

Estimation and Prediction

$$\begin{pmatrix} \mathbf{z}_{observed} \\ \mathbf{z}_{unobserved} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \ \text{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

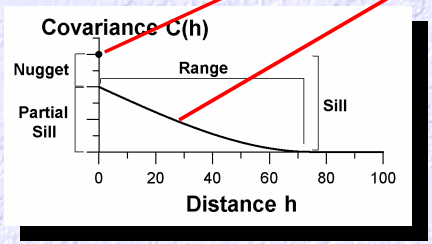
Prediction
- Mapping
- Sampling

Estimation
- Regression
- Designed Experiments

Why Do We Need Autocorrelation Models?

$$\begin{pmatrix} \mathbf{z}_{observed} \\ \mathbf{z}_{unobserved} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \ \text{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{21} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$



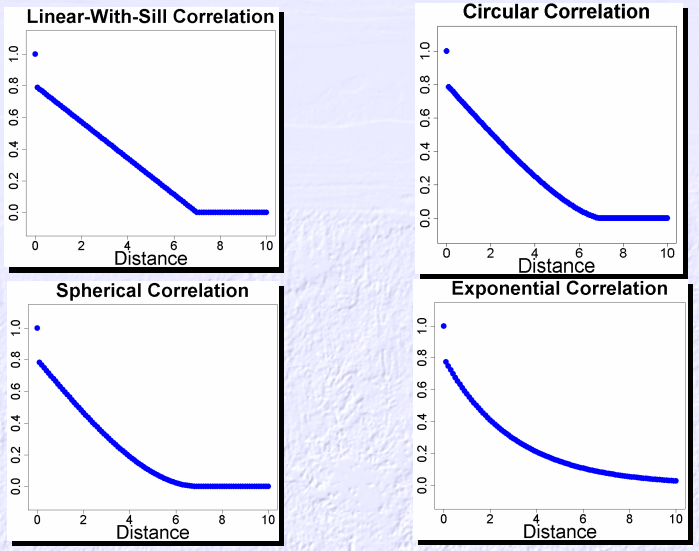
Estimation?!

Leave it to the Statisticians!

- Weighted Least Squares
- Generalized Least Squares
- Maximum Likelihood
- Restricted Maximum Likelihood
- Bayes (Markov Chain Monte Carlo MCMC)

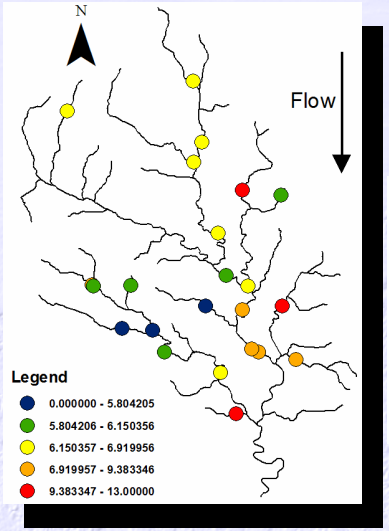
$$\begin{pmatrix} \mathbf{z}_{observed} \\ \mathbf{z}_{unobserved} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \ \text{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

Pitfalls: Valid Autocorrelation Models

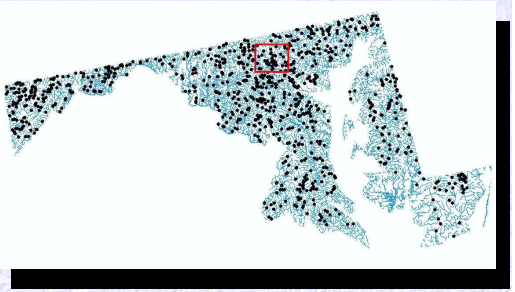


21

Stream Network Models

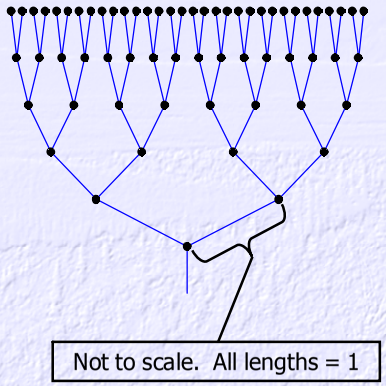


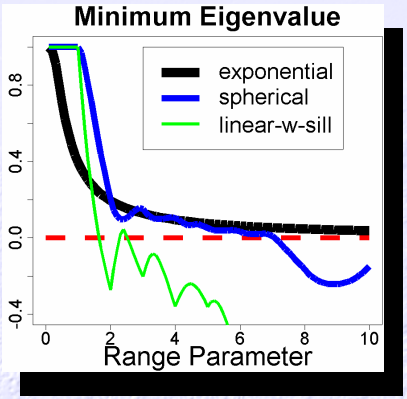
SO_4 Concentration



22

Pitfalls: Valid Models for Stream Networks



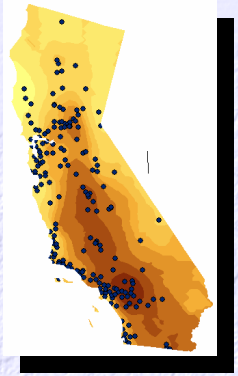


23

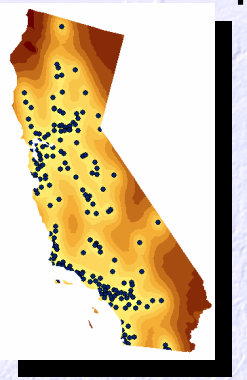
Prediction - Mapping

$$\begin{pmatrix} \mathbf{z}_{observed} \\ \mathbf{z}_{unobserved} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathrm{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

Prediction Map

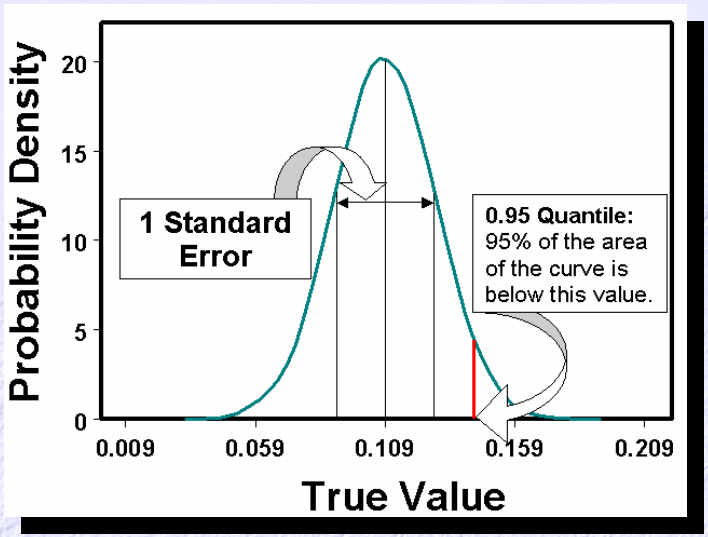


Standard Error Map



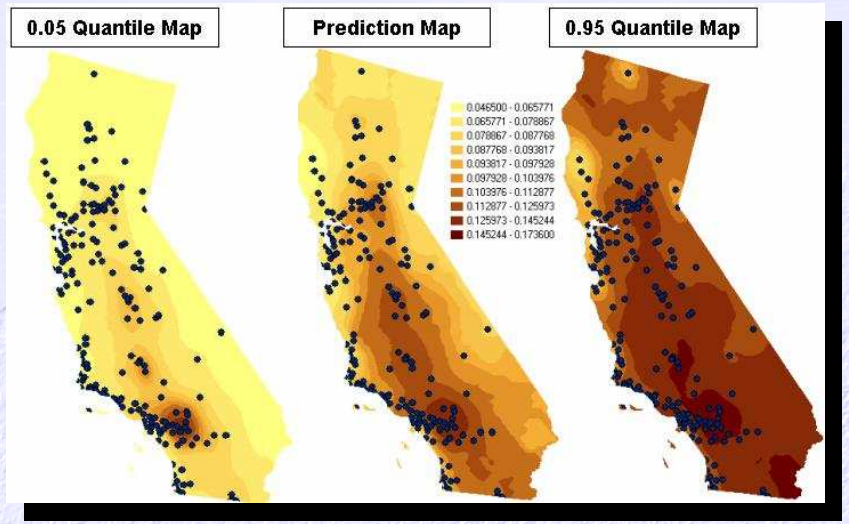
24

Mapping - Quantiles



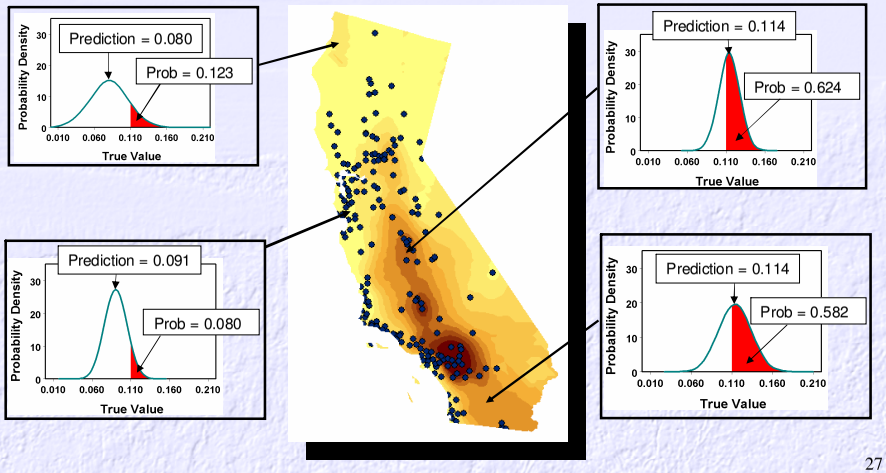
25

Quantile Maps



26

Probability Maps



27

Spatial Regression

- Whiptail Lizard
- 148 locations in Southern California
- Measured the average number caught in traps over 80 – 90 trapping events in one year
- Data log-transformed, one outlier removed

28

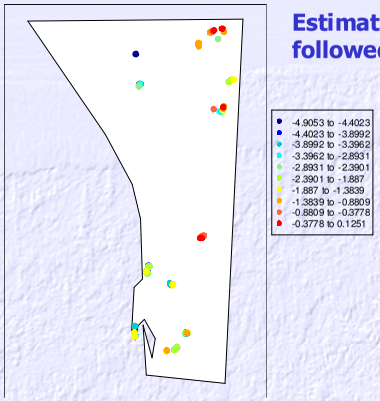
Whiptail Lizard Example

- There were 37 explanatory variables in 5 broad categories: vegetation layers, vegetation types, topographic position, soil types, and ant abundance

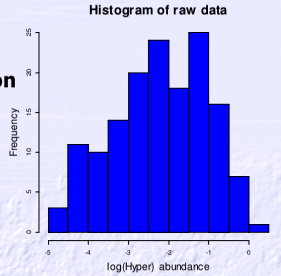
California Lizard Data

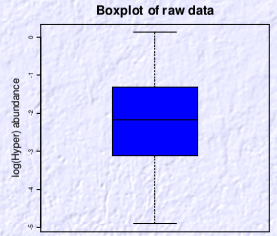
$$\begin{pmatrix} \mathbf{z}_{observed} \\ \mathbf{z}_{unobserved} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \ \text{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

- Ant abundance
- Percent sandy soil
- Spherical autocorrelation
- Isotropic



Estimation: REML followed by GLS

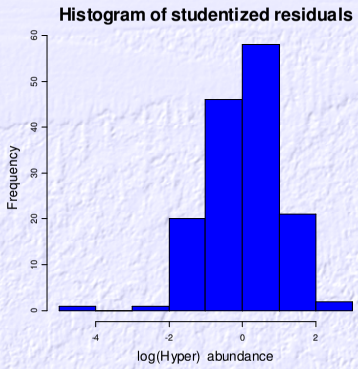


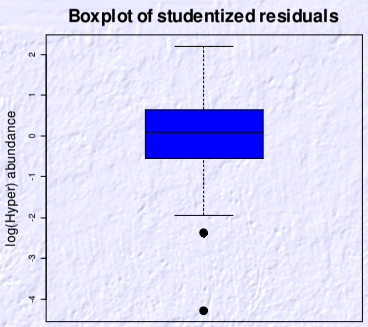


Exploratory Data Analysis on Residuals

Studentized Residual:

$$\frac{z_i - \hat{\mu}_i}{\sqrt{MSE(1-h_{ii})}} \qquad \mathbf{H} = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-}\mathbf{X}'\boldsymbol{\Sigma}^{-1/2}$$

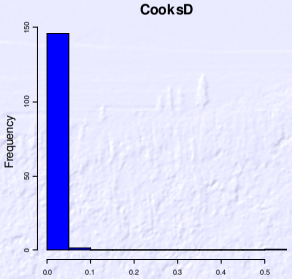




Model Diagnostics

Based on deleting observations

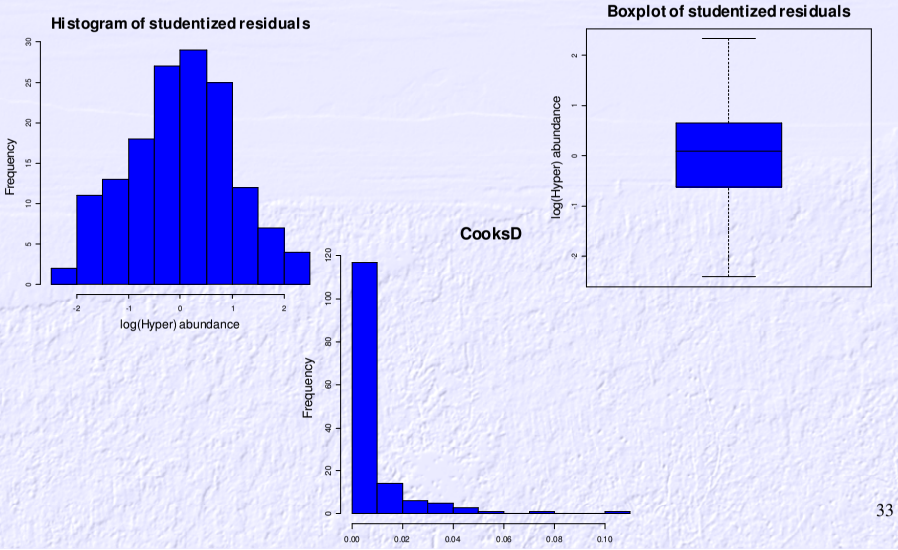
- Likelihood Distance
- Cook's D and Leverage
- Covariance Trace



Mostly for outlier detection

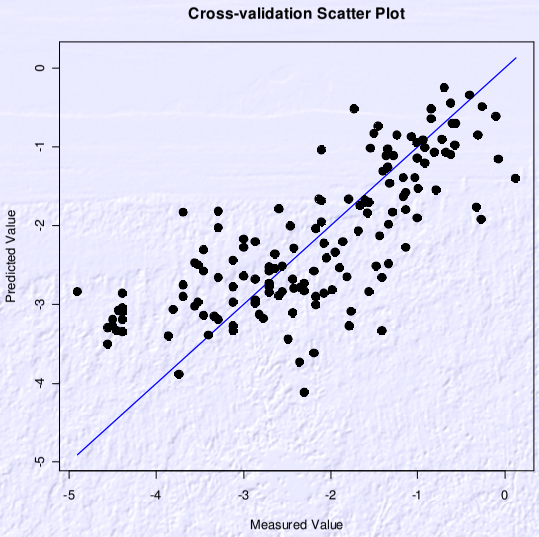
$$\left(\mathbf{z}_{observed-1}\right) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \ \text{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

Remove Outlier and Re-fit Lizard Data



33

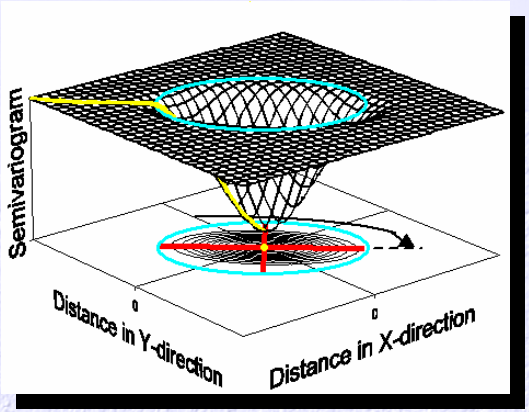
Cross-validation



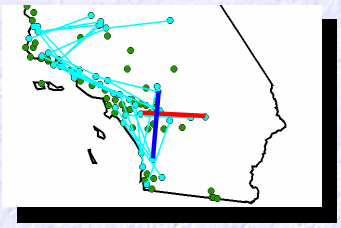
34

Directional Autocorrelation

5 Parameters



Isotropy vs. Anisotropy



35

Cross-validation

$$\begin{pmatrix} \mathbf{z}_{observed-i} \\ Z_i \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \ \ \mathrm{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

Standardized Bias :

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{Z}_i - z_i)}{\sqrt{\hat{\mathrm{var}}(\hat{Z}_i)}}$$

Root Mean Squared Prediction Error :

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Z}_i - z_i)^2}$$

Prediction Interval Coverage :

$$\sum_{i=1}^{n} I\left(\frac{|(\hat{Z}_i - z_i)|}{\sqrt{\hat{\mathrm{var}}(\hat{Z}_i)}} < 1.96\right)$$

- Spherical autocorrelation
- Isotropic

Standardized Bias	-0.0037
RMSPE	0.7989
80% Coverage	76.4%
90% Coverage	86.5%
95% Coverage	93.9%

- Spherical autocorrelation
- Anisotropic

Standardized Bias	-0.0027
RMSPE	0.7785
80% Coverage	77.7%
90% Coverage	86.5%
95% Coverage	95.9%

36

Model Selection

- AIC
- AICc
- BIC
- etc.!

} -2*loglikelihood + (Penalty for number of parameters)

Choose the model with the Minimum of these:
Be careful! Some software uses 2*loglikelihood – (Penalty for number of parameters), in which case you choose the maximum.

Can also use RMSPE and other criteria. Why not?

37

Model Selection

Variogram	Anis.	AIC	RMSPE	95%PI
Spherical	No	398.86	0.874	95.3%
Exponential	No	398.62	0.873	95.3%
Spherical	Yes	394.68	0.841	96.0%
Exponential	Yes	394.67	0.834	95.3%

38

Final Fitted Model

$$\begin{pmatrix} \mathbf{z}_{observed} \\ \mathbf{z}_{unobserved} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathrm{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

Partial Sill	0.723
Major Range	12.78
Nugget	0.524
Minor Range	0.012
Rotation	163.6

effect	estimate	std.err	df	t.value	prob.t
Intercept	-3.994	0.5003	145	-7.984	<0.00001
Ant_Abund	0.306	0.1007	145	3.037	0.00283
Sandy_Soil	1.080	0.2345	145	4.606	0.00001

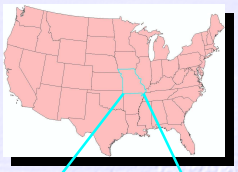
39

Spatial Regression References

- Ver Hoef, J.M. 1993. Universal kriging for ecological data. Pages 447 – 453 in Goodchild, M.F., Parks, B., and Steyaert, L.T. (eds.) *Environmental Modeling with GIS*, Oxford University Press, 488 p.
- Ver Hoef, J.M., Cressie, N., Fisher, R.N., and Case, T.J. 2001. Uncertainty and spatial linear models for ecological data. Pages 214 – 237 in Hunsaker, C.T., Goodchild, M.F.. Friedl, M.A., and Case, T.J. (eds.), *Spatial Uncertainty for Ecology: Implications for Remote Sensing and GIS Applications* Springer-Verlag.
- Maier, J.A.K., Ver Hoef, J.M., McGuire, A.D., Bowyer, R.T., Saperstein, L. and Maier, H.A. 2006. Distribution and density of moose in relation to landscape characteristics: Effects of scale. In press, *Canadian Journal of Forest Research*.

40

Glades in Ozarks



41

Glades in Ozarks

42

Designed Experiment

32	26	24	24	24
26	25	22	22	23
23	21	21	20	24
26	23	26	22	25
25	23	24	24	27

Add Trts

Estimate

Treatment	Effect
1	0
2	–3
3	–5
4	+6
5	+6

3	1	2	2	4
5	5	1	4	4
5	1	3	4	4
3	1	5	2	3
5	2	3	1	2

Contrast	True Value
$c_1 = (\tau_2 + \tau_3)/2 - \tau_1$	-4.00
$c_2 = (\tau_4 + \tau_5)/2 - \tau_1$	6.00
$c_3 = (\tau_4 + \tau_5)/2 - (\tau_2 + \tau_3)/2$	10.00
$c_4 = (\tau_2 - \tau_3)$	2.00
$c_5 = (\tau_4 - \tau_5)$	0.00

43

Estimation and Prediction

$$\begin{pmatrix} \mathbf{z}_{observed} \\ \mathbf{z}_{unobserved} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathrm{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

Prediction
- Mapping
- Sampling

Estimation
- Regression
- Designed Experiments

44

Linear Models

$$Z(\mathbf{s}_i) = \tau_j + \varepsilon(\mathbf{s}_i) \text{ or } \mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad E(\boldsymbol{\varepsilon}) = \mathbf{0}$$

Covariance Models

$\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ **Independence Models**

$\text{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$ **Geostatistical Models**

Exponential, Spherical, etc.

$\text{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$ **Lattice Models**

CAR, SAR, etc.

45

Estimating Treatment Effects

Contrast	True Value	OLS Est	OLS se	Freq Geo Est	Freq Geo se	Freq Lat Est	Freq Lat se	Bayes Geo Est	Bayes Geo se	Bayes Lat Est	Bayes Lat se
$c_1 = (\tau_2 + \tau_3)/2 + \tau_1$	-4.00	-2.40	1.29	-2.95	0.87	-2.94	0.90	-2.65	1.12	-2.76	1.07
$c_2 = (\tau_4 + \tau_5)/2 + \tau_1$	6.00	6.60	1.29	6.81	1.05	6.81	1.02	6.72	1.18	6.81	1.12
$c_3 = (\tau_2 + \tau_3)/2 + (\tau_4 + \tau_5)/2$	10.00	9.00	1.05	9.77	0.84	9.75	0.86	9.37	0.89	9.57	0.98
$c_4 = (\tau_2 - \tau_3)$	2.00	0.40	1.49	0.53	1.07	0.71	1.16	0.42	1.47	0.71	1.38
$c_5 = (\tau_4 - \tau_5)$	0.00	-2.40	1.49	-1.94	1.68	-2.29	1.60	-1.96	1.86	-2.44	1.63
			nugget	0.00	0.07			0.70	1.44	0.87	1.88
			parsil	13.54	21.22	10.96		10.78	21.74	11.70	5.94
			range	3.73	6.54	0.90		2.66	6.71	0.64	0.25

32	26	24	24	24
26	25	22	22	23
23	21	21	20	24
26	23	26	22	25
25	23	24	24	27

3	1	2	2	4
5	5	1	4	4
5	1	3	4	4
3	1	5	2	3
5	2	3	1	2

46

Designed Experiment Experiment

32	26	24	24	24
26	25	22	22	23
23	21	21	20	24
26	23	26	22	25
25	23	24	24	27

1600 times

Estimate

Treatment	Effect
1	0
2	–3
3	–5
4	+6
5	+6

3	1	2	2	4
5	5	1	4	4
5	1	3	4	4
3	1	5	2	3
5	2	3	1	2

Contrast	*True Value*
$c_1 = (\tau_2 + \tau_3)/2 - \tau_1$	-4.00
$c_2 = (\tau_4 + \tau_5)/2 - \tau_1$	6.00
$c_3 = (\tau_4 + \tau_5)/2 - (\tau_2 + \tau_3)/2$	10.00
$c_4 = (\tau_2 - \tau_3)$	2.00
$c_5 = (\tau_4 - \tau_5)$	0.00

47

Designed Experiment Results

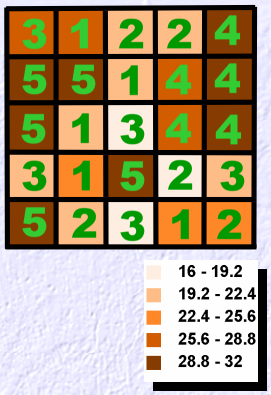
		ANOVA	GLS-Variogram	Spatial ML Estimation	Spatial REML Estimation
MSE	c_1	1.810	1.166	1.103	1.037
	c_2	1.822	1.183	1.105	1.040
	c_3	1.139	0.766	0.724	0.687
	c_4	2.364	1.546	1.457	1.380
	c_5	2.433	1.608	1.551	1.461
Coverage	c_1	0.9450	0.9440	0.9300*	0.9505
	c_2	0.9495	0.9415	0.9240*	0.9500
	c_3	0.9575	0.9495	0.9365*	0.9560
	c_4	0.9500	0.9390*	0.9250*	0.9490
	c_5	0.9435	0.9385*	0.9215*	0.9465
Power	c_1	0.8255	0.9750	0.9845	0.9825
	c_2	0.9960	1.0000	1.0000	1.0000
	c_3	1.0000	1.0000	1.0000	1.0000
	c_4	0.2155	0.3370	0.4095	0.3560
	c_5	0.0565	0.0615	0.0785	0.0535

48

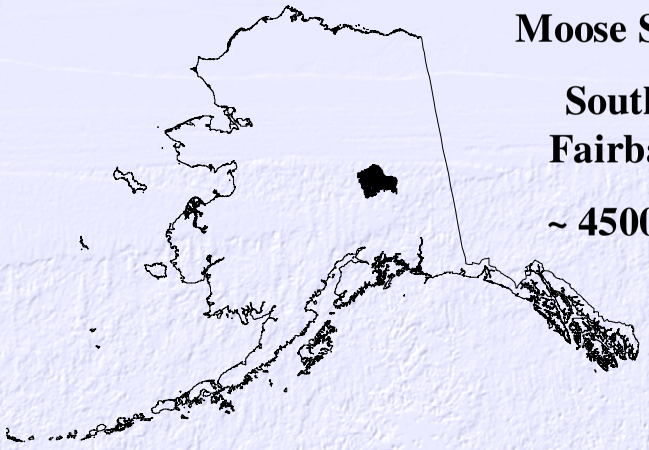
Designed Experiments References

- Ver Hoef, J.M. and Cressie, N. 2001. Spatial statistics: Analysis of field experiments. In Scheiner, S.M. and Gurevitch, J. (eds.), *Design and Analysis of Ecological Experiments, Second Edition*, Oxford University Press, p. 289-307.
- Lenart, E.A., Bowyer, R.T., Ver Hoef, J., and Ruess, R.W. 2002. Climate change and caribou: effects of summer weather on forage. *Canadian Journal of Zoology* **80**: 664 – 678.

Plots in a Designed Experiment



Spatial Sampling

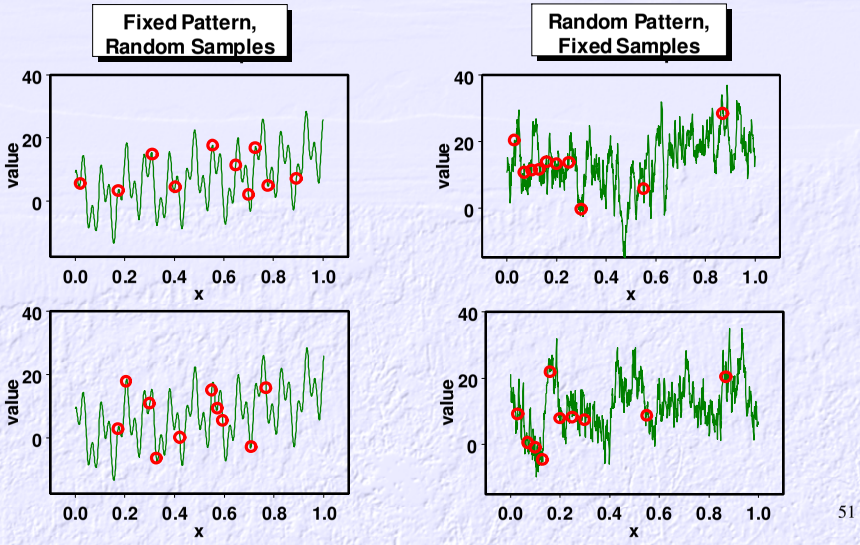


Moose Survey

South of Fairbanks

~ 4500 mi²

Sources of Randomness



Source of Randomness

Fixed Pattern, Random Samples

$$z(x) = \alpha_{s1}\sin(\beta_{s1}x) + \alpha_{s2}\sin(\beta_{s2}x) + \alpha_{c1}\cos(\beta_{c1}x) + \alpha_{c2}\cos(\beta_{c2}x) + \alpha_e(\exp(x) - 1)$$

Random Pattern, Fixed Samples

$$z(x_i) = \rho z(x_{i-1}) + \varepsilon(x_i);\ \ \varepsilon(x_i) \sim N(0, \sigma^2)$$

Sampling and Geostatistics

		Method	
		Classical Sampling	Geostatistics
Population	Infinite (Spatially Continuous)	Classical Sampling Methods	Block Kriging
	Finite (Spatially Discrete)	Classical Sampling Methods	Finite Population Block Kriging

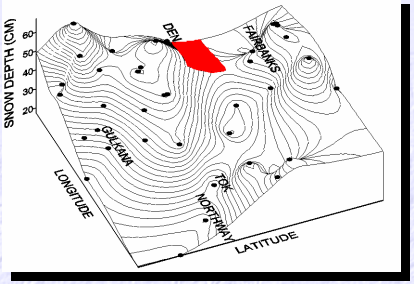
Infinite Population Parameters

Total

$$\tau = \int_A z(\mathbf{s})d\mathbf{s}$$

Mean

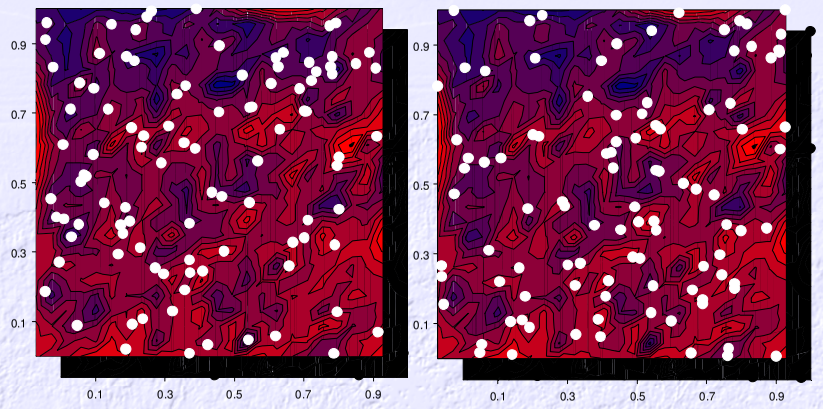
$$\alpha = \int_A z(\mathbf{s})d\mathbf{s} / |A|$$



$$\begin{pmatrix} \mathbf{z}_{observed} \\ \mathbf{z}_{unobserved} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathrm{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

Simulation Study

Fixed Pattern, Random Samples



Simulation Results

Table 1. Comparison of random sampling and block kriging. 1000 random samples were generated from a fixed continuous spatial pattern. Sample sizes were 100. For block kriging, an isotropic exponential covariance model was estimated from the sample data using REML.

Validation Statistics	*SRS*[1]	*BK*[2]
Bias	0.002	-0.020
RMSPE[3]	1.28	1.02
RAEV[4]	1.29	1.00
80%CI[5]	0.813	0.806

[1] Simple Random Sampling
[2] Block Kriging
[3] root mean squared prediction errors
[4] root average estimated variance
[5] 80% confidence interval coverage

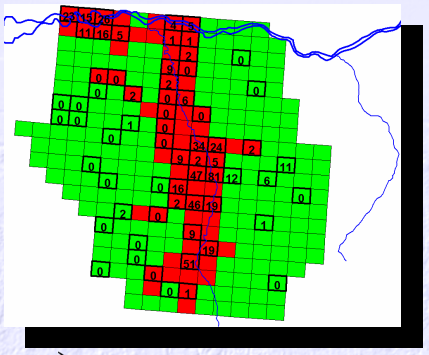
Sampling for Finite Populations

		Method	
		Classical Sampling	Geostatistics
Population	Infinite (Spatially Continuous)	Classical Sampling Methods	Block Kriging
	Finite (Spatially Discrete)	Classical Sampling Methods	Finite Population Block Kriging

57

Finite Population Parameters

$$\tau = \sum_{i=1}^{N} z(\mathbf{s}_i) \qquad \alpha = (1/N)\sum_{i=1}^{N} z(\mathbf{s}_i)$$

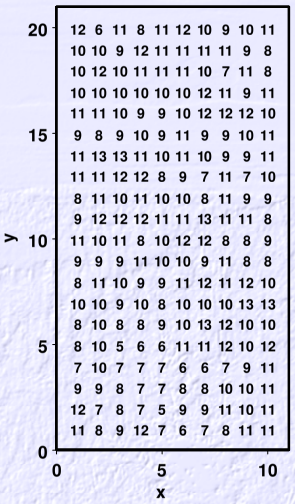


$$\begin{pmatrix} \mathbf{z}_{observed} \\ \mathbf{z}_{unobserved} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathrm{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

58

Simulation Study

Number of plant species in 70 x 70 cm plots



Fixed Population, $N = 200$

Random Sample, $n = 100$

59

Simulation Results

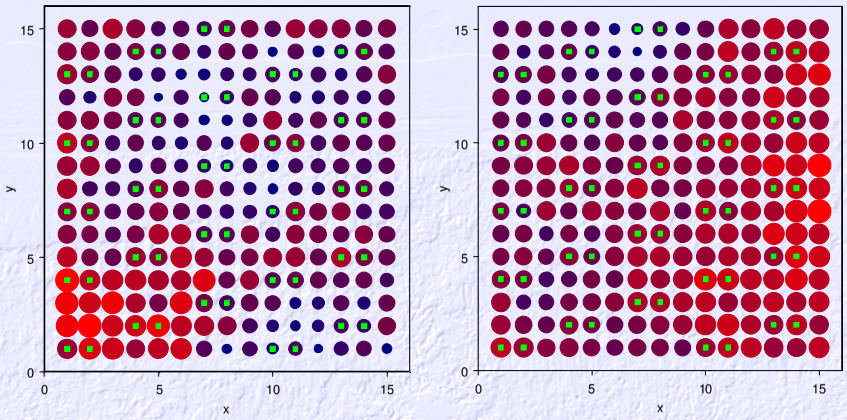
Table 2. Comparison of Random Sampling and Finite Population Block Kriging. 1000 random samples were generated for the fixed spatial pattern given by the species diversity data. Sample sizes were 100. For FPBK, an isotropic exponential covariance model was estimated from the sample data using REML.

Validation Statistics	*SRS*[1]	*FPBK*[2]
Bias	-0.002	-0.001
RMSPE[3]	0.121	0.106
RAEV[4]	0.122	0.105
80%CI[5]	0.802	0.806

[1] Simple Random Sampling
[2] Finite Population Block Kriging
[3] root mean squared prediction errors
[4] root average estimated variance
[5] 80% confidence interval coverage

60

Simulation Study



61

Simulation Results

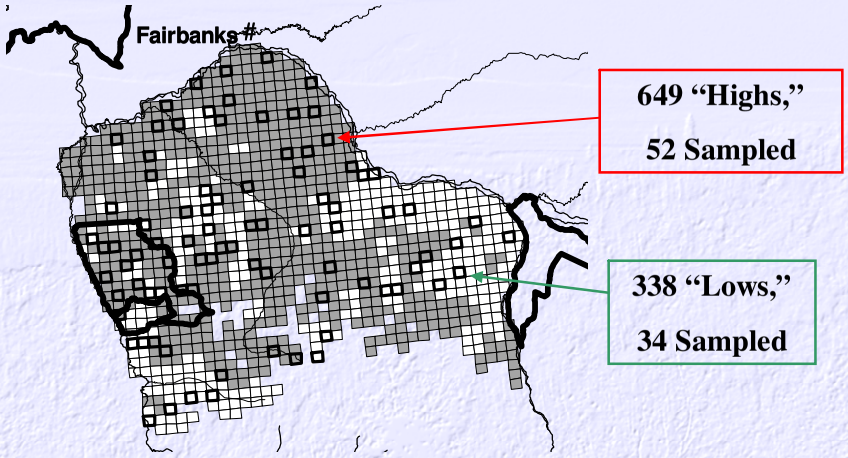
Table 3. Comparison of Random Sampling and Finite Population Block Kriging. 1000 patterns were generated using a spatially autocorrelated stochastic process, and fixed and random samples were taken. Sample sizes were 50. For FPBK, an isotropic exponential covariance model was estimated from the sample data using REML.

Validation Statistics	*SRS*[1]	*FPBKr*[2]	*FPBKf*[3]
Bias	0.522	-0.181	0.127
RMSPE[4]	28.0	20.7	17.3
RAEV[5]	28.0	20.3	17.5
80%CI[6]	0.801	0.791	0.796

[1] Simple Random Sampling
[2] Finite Population Block Kriging from random sample
[3] Finite Population Block Kriging from fixed sample
[4] root mean squared prediction errors
[5] root average estimated variance
[6] 80% confidence interval coverage

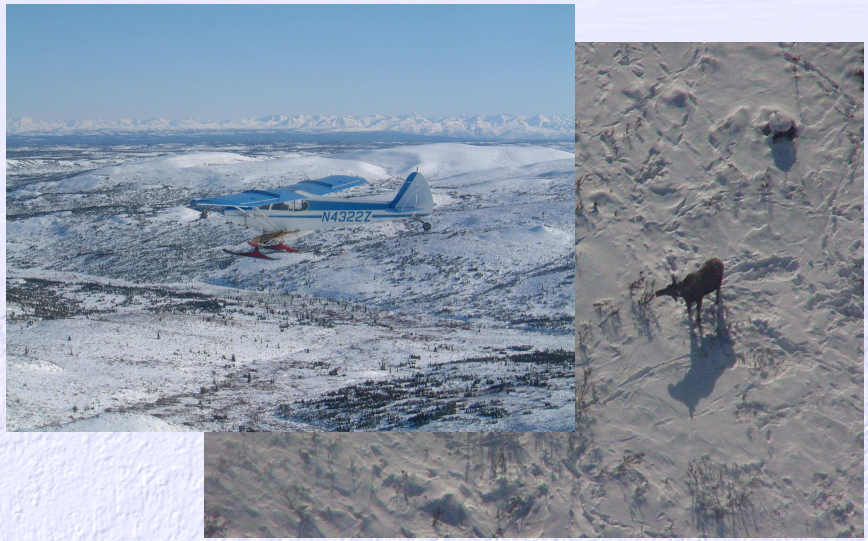
62

Real Example – Moose Survey



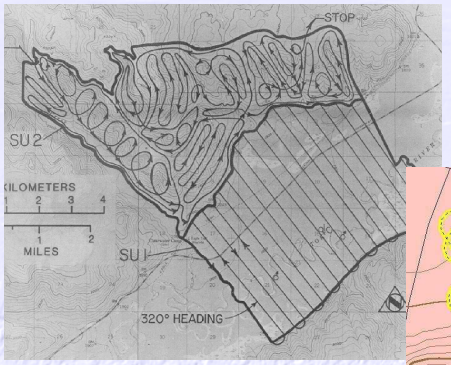
63

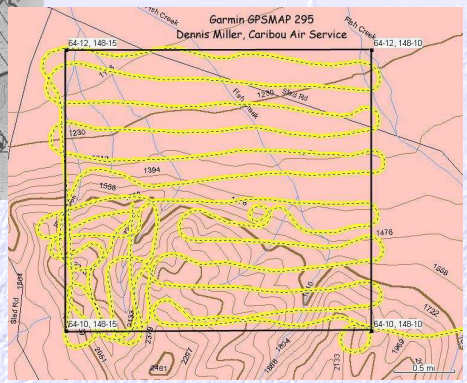
Conducting the Survey



64

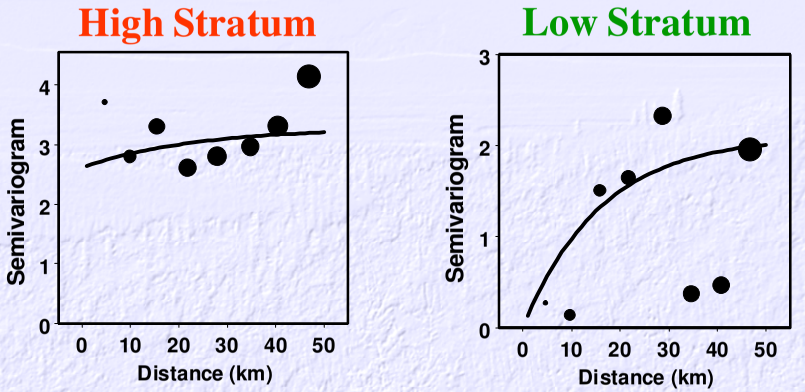
Conducting the Survey





65

Modeling Covariance



66

Results

Small Area

FPBK

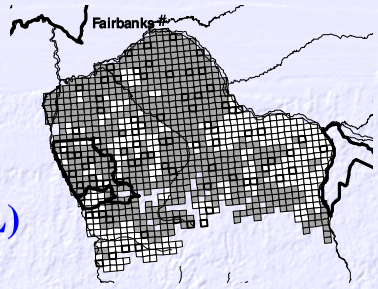
$\hat{t} = 1437$

$se(\hat{t}) = 153$

SRS (13H, 4L)

$\hat{t} = 1535$

$se(\hat{t}) = 227$



Total Area

FPBK

$\hat{t} = 11327$

$se(\hat{t}) = 978$

SRS

$\hat{t} = 11535$

$se(\hat{t}) = 985$

67

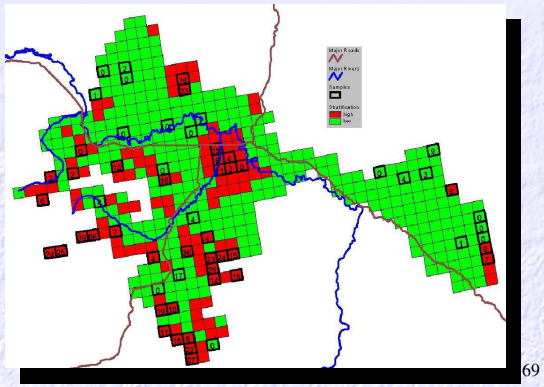
Summary

- Geostatistical Methods for Sampling are often more precise
- Geostatistical Methods for Sampling allow small area estimation
- Geostatistical Methods for Sampling do not require randomized designs
- Geostatistical Methods require modeling

68

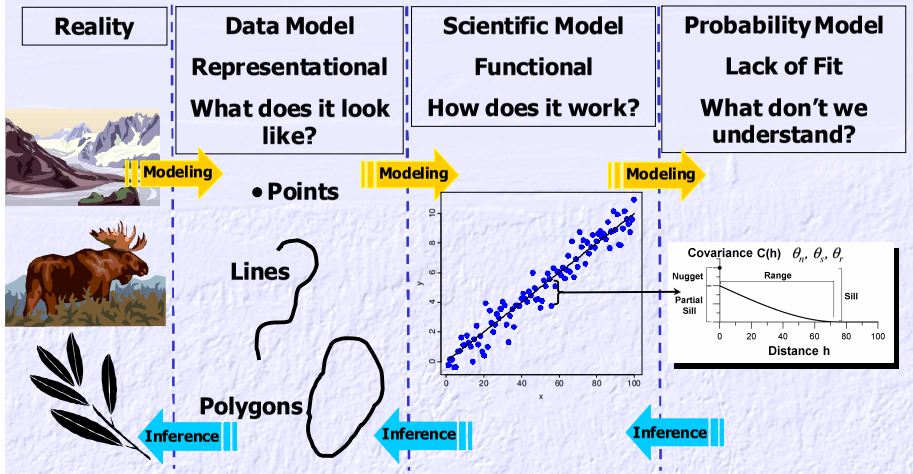
Geostatistics and Sampling References

- Ver Hoef, J.M. 2001. Predicting finite populations from spatially correlated data. *2000 Proceedings of the Section on Statistics and the Environment of the American Statistical Association*, pgs. 93-98.
- Ver Hoef, J.M. 2002. Sampling and geostatistics for spatial data. *Ecoscience* **9**: 152 – 161.
- Ver Hoef, J.M. 2006. Spatial Methods for Plot-based Sampling of Wildlife Populations. In press, *Environmental and Ecological Statistics*



69

Good Science is a Team Effort



"All models are wrong. We make tentative assumptions about the real world which we know are false but which we believe may be useful."- George Box 1976

70

Why Include Spatial Dependencies?

Mark Otto

U.S. Fish and Wildlife Service

15 March 2006

Statistical Models
Biased Estimates of Standard Errors
Geostatistical Spatial Models
Lattice Models
Experimental Design

I want to start with the concepts of some basic statistical models. Concepts you are familiar with. Then, I want to build and apply them to models used on correlated data. I'll start with repetition and IID, extend it to regression models and transformation. We can then see how transformations are also used to handle correlations data taken over time and space. Finally, I will talk about the benefits and difficulties of correlated data in mapping and prediction, regression, and experimental design.

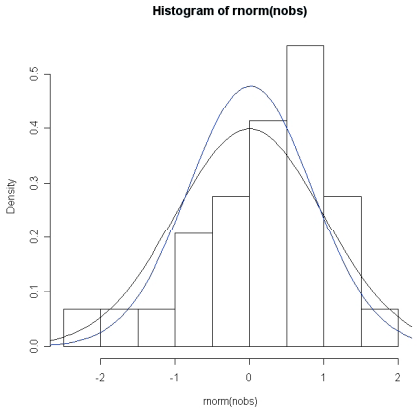
Statistical Models

- Variable data
- Consistent patterns:
  - Parameter estimates
  - Fitted values
  - Predictions
- Repetition
- Model assumptions

Statistical Models
- Variable data
- Consistent patterns:
  - Parameter estimates
  - Fitted values
  - Predictions
- Repetition
- Model assumptions

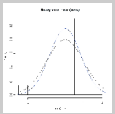
In statistics, we collect variable data that we hope to pull out consistent patterns from. Probability distributions were made to characterize common forms of variability: binomial, and multinomial for categorical data, Poisson data for counts, and with all data no matter what the distribution has estimates the approximate a normal distribution. To extract those patterns, they must repeat in the data many times. In practice we don't know what processes generated the data, so we have to pay attention to the assumptions of the models that we are using and check that the data do not deviate greatly from them

Data from a Normal



6

Data from a Normal



Here are 29 values from a standard distribution, the mean is 0 and the variance is 1. The histogram is of the data, the black line is from standard normal the sample was generated from. The blue line is the normal curve with mean and variance estimated from the sample data. The estimated curve is a reasonable approximation.

Independent, Identically Distributed

- Identically Distributed: each observation is sample from the same distribution
- Independent: One observation does not give information about preceding or succeeding draws

8

Independent, Identically Distributed
- Identically Distributed: each observation is sample from the same distribution
- Indepenent: One observation does not give information about preceding or succeeding draws

IID is one of the most basic assumptions we make for statistical modeling. In term of the covariance independence means no correlation among observations.

Regression

- Mean allowed to vary
- Variance constant

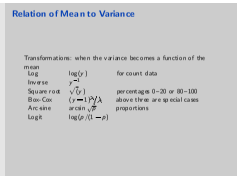
Regression
- Mean allowed to vary
- Variance constant

The first order effects vary, but the relation between the independent variables is constant, $\mathbf{y} = \mathbf{X}\beta$. The variance or second order effects are constant, $\mathbf{I}\sigma^2$. Their is repetition in the data: repetition around a varying mean according to a constant variance. Each observation gives information about β, that relates y to X. The residuals, the data after removing the mean function, are $N(0, \sigma^2)$.

Relation of Mean to Variance

Transformations: when the variance becomes a function of the mean

Log	$\log(y)$	for count data
Inverse	y^{-1}	
Square root	$\sqrt(y)$	percentages 0–20 or 80–100
Box-Cox	$(y-1)^{\lambda}/\lambda$	above three are special cases
Arc-sine	$\arcsin\sqrt{p}$	proportions
Logit	$\log(p/(1-p)$	



For data that does not fit the usual assumptions, we can build more flexible and complex models or we can transform the data back to something that fits the usual assumptions. Here we transform to remove the relation between the mean and variance. Back to IID and a constant variance.

Time Series

Simplified spatial data

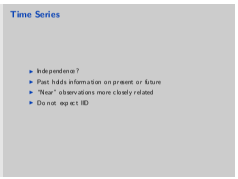
- Both correlated data
- One dimension: time
- Usually sampled regularly: daily, monthly, annually
- Order to observations: past, present, future

14

Time Series

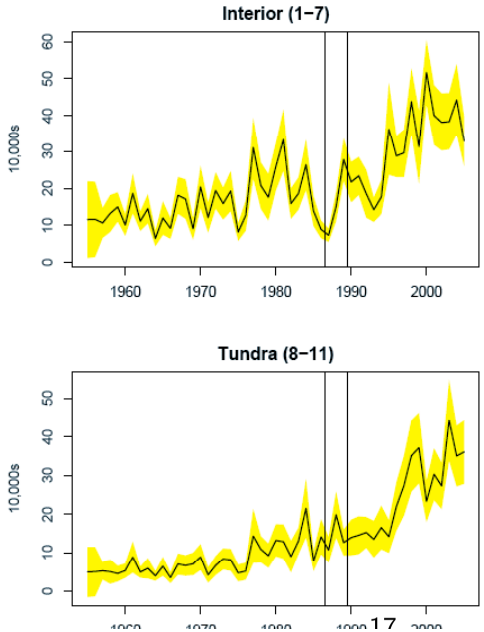
- Independence?
- Past holds information on present or future
- "Near" observations more closely related
- Do not expect IID

15

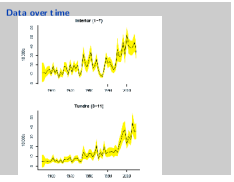


Very reason take measurements periodically is to look for the patterns over time. Look for trends and periodicities. If we did not would not plot the data, just take the information react to it and throw it away.

Data over time



2006-03-07
Why Include Spatial Dependencies?
Statistical Models
Data over time

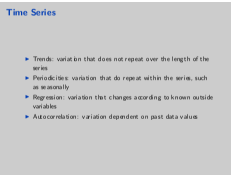


Here we have mallard counts in Alaska. The observations are not independent. Overall the series rises over time. On small scales the deviations are similar. We no longer have IID and do not expect it.

Time Series

- Trends: variation that does not repeat over the length of the series
- Periodicities: variation that do repeat within the series, such as seasonally
- Regression: variation that changes according to known outside variables
- Autocorrelation: variation dependent on past data values

2006-03-07
Why Include Spatial Dependencies?
Statistical Models
Time Series



Time series analysis is the characterization of the correlations as a function of time lag

Where is the Repetition?

- Each observation depends on the observations that came before
- Observation at each time point could be a regression on its past
- Repetitions are of the variation between observations a given number of time lags apart

21

Stationarity

Everything is relative.

- Strong: distribution is the same regardless of where
- Weak: Mean and covariance are the same regardless of position
- Dependence of the data on its past does not change with the mean or with time

22

Stationarity

Everything is relative.
- Strong: distribution is the same regardless of where
- Weak: Mean and covariance are the same regardless of position
- Dependence of the data on its past does not change with the mean or with time

This is a similar situation to the variance being dependent on the mean. Here the autocovariances do not depend on the mean or time. Whether you care about the temporal or spatial effects or not, when working with correlated data, you need to pay attention to the assumptions of the model. The consequences are that your analysis may not make sense (or appear reasonable but be wrong).

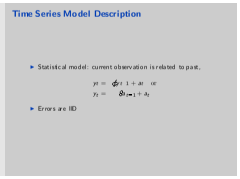
Time Series Model Description

- Statistical model: current observation is related to past,

$$y_t = \phi y_{t-1} + a_t \quad \text{or}$$
$$y_t = \theta a_{t-1} + a_t$$

- Errors are IID

24



Looks like regression but on the series itself, (independent variables are fixed and without error)??? Data are multivariate normal, with a variance as a function of the time series parameters, say ϕ and or θ

$$\mathbf{y} \sim \mathrm{N}(\text{fixed effects} + \text{trend} + \text{periodicities}, \mathbf{V}(\phi, \theta))$$

If we knew the time series parameters this would be generalized least squares (GLM, not GLIM)

Startin' Up

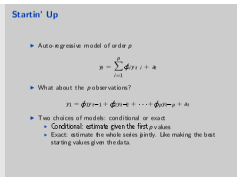
- Auto-regressive model of order p

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + a_t$$

- What about the p observations?

$$y_1 = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + a_t$$

- Two choices of models: conditional or exact
  - Conditional: estimate given the first p values
  - Exact: estimate the whole series jointly. Like making the best starting values given the data.



There is no data for $y_{t-p}, \ldots, y_{t-1}$. Before PCs these were "complex" computations, this was an issue. In fact the first solution was to run the series backwards, forecast the first p values then use them in the original series. The transformation in the next slide will show the exact approach.

Transform Back to IID

Regression residuals, $\mathbf{y} - \mathbf{X}\beta$ have mean $\mathbf{y}$ and general variance $\mathbf{V} = [v_{ij}]$ We can split the variance matrix into two parts like we take the square root

$$\begin{array}{ccccc} \sigma^2 & & = & & \sqrt{(\sigma^2)} & & \sigma \\ \mathbf{V} & & = & & \mathbf{L} & & \mathbf{L}' \\ \begin{bmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{1n} & \cdots & v_{nn} \end{bmatrix} & & = & & \begin{bmatrix} l_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ l_{1n} & \cdots & l_{nn} \end{bmatrix} & & \begin{bmatrix} l_{11} & \cdots & l_{1n} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_{nn} \end{bmatrix} \end{array}$$

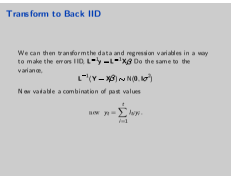
Transform to Back IID

We can then transform the data and regression variables in a way to make the errors IID, $\mathbf{L}^{-1}\mathbf{y} - \mathbf{L}^{-1}\mathbf{X}\beta$ Do the same to the variance,

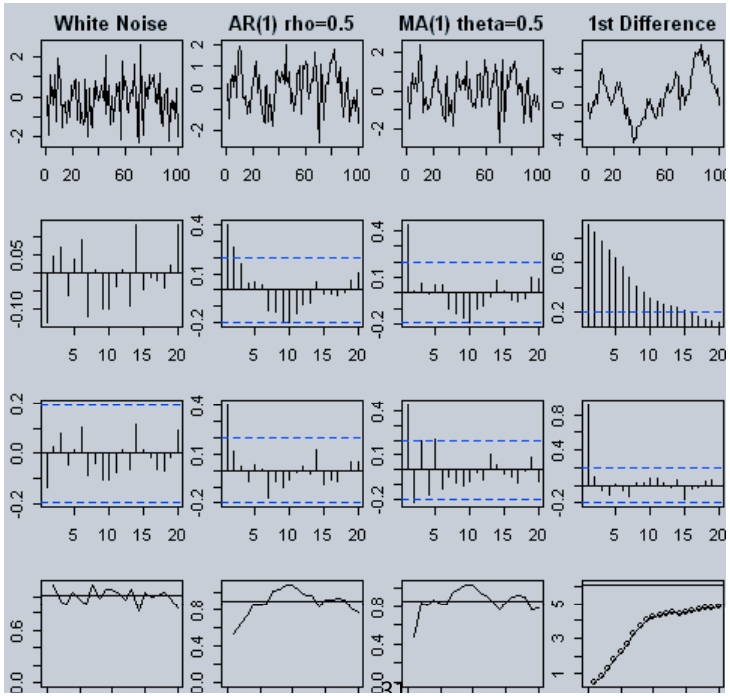
$$\mathbf{L}^{-1}(\mathbf{Y} - \mathbf{X}\beta) \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2)$$

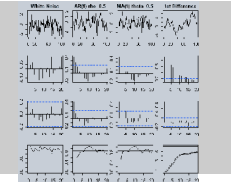
New variable a combination of past values

$$\text{new } y_t = \sum_{i=1}^{t} l_{ti} y_i.$$

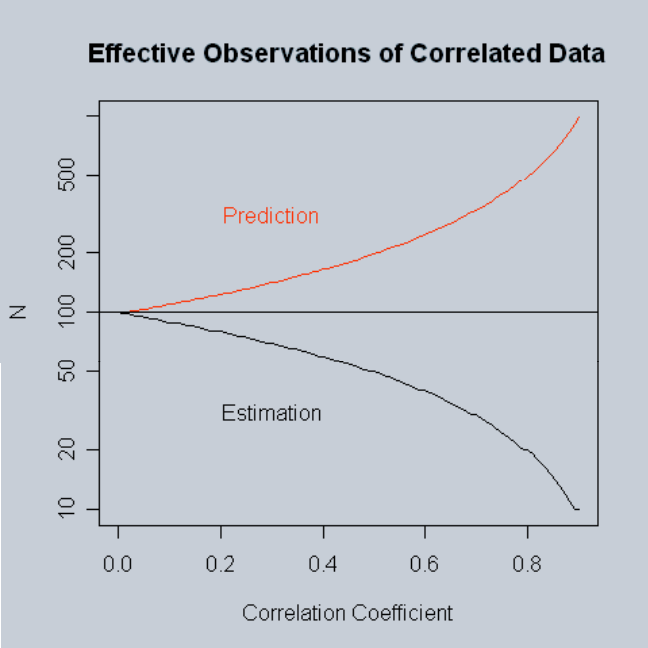


We transformed data and regression back to IID Just make new variables that are linear combinations of data and regression variables, new $y_t = \sum_{i=1}^{t} l_{ti} y_i$. They are combinations of past data. The new variables are approximately transformed back to IID The identification, estimation, and diagnostics are more complicated, but the concepts are the same.

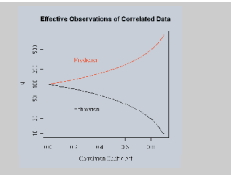




This shows four time series: white noise, autoregressive order 1 where adjacent time points data are correlated and moving average 1 where the adjacent errors are correlated and a first differenced series where the changes between one time point and the last are random.
Rows of plots are: (1)the series; (2) the autocorrelations (ACF), which are correlations between observations a given number of time lags apart; (3) the partial autocorrelations (PACF), like partial correlation coefficients. Auto-regressive-moving-average (ARIMA) models are identified using these functions. AR models by the significant ACFs, MA by the number of PACFs. (4) the variogram is how spatial correlations are identified because they give unbiased estimates and are valid for a larger range of models. The repetitions are here, the relations between observations at different time lags. For the last series, the variogram just increases. It is not stationary. There is no mean. You cannot fit models that assume stationarity to this data.



33



What are the consequences of working with correlated data. This shows the number of IID observations it would take to obtain the same standard error of the mean given the data are all correlated by the same value, ρ. This is the extreme case. Real analyzes will fall within these bounds. Note that as the correlation increases the number of effective observations drops, i.e., each observation provides less information. With prediction the situation is different; the correlation provides information to unrealized observations.

Biased Estimates of Standard Errors

OLS Regression			AR(1) phi=0.36	
Variable	Est	SE	Est	SE
Constant	97.6	12.3	80.8	15.2
72-97	-1.23	0.4	-0.6	0.6
Variance	0.036		0.032	
AIC	1159		1153	

Here is an example of a change in the Scoter breeding population due to a change in harvest regulations. In general the estimates are unbiased but the standard errors are. For positive correlations the standard error are underestimated and the reverse for negative correlations. Even though the correlations are a nuisance they need to be addressed

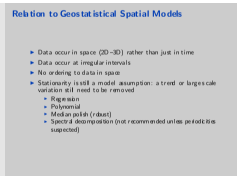
[Plot of Scoter Populations with change in Harvest Regulations]

35

Relation to Geostatistical Spatial Models

- Data occur in space (2D–3D) rather than just in time
- Data occur at irregular intervals
- No ordering to data in space
- Stationarity is still a model assumption: a trend or large scale variation still need to be removed
  - Regression
  - Polynomial
  - Median polish (robust)
  - Spectral decomposition (not recommended unless periodicities suspected)

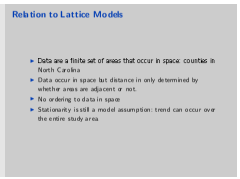
36



Since the data occur more than one dimension, the correlation structure may be different in different directions (anisotropy). Duck populations may be more similar latitudinally then longitudinally.With ARIMA models, we estimate best correlation at given lags. With spatial data, we need to model the correlation as a continuous function of distance between observations. Because there is no ordering of data in space and we are estimating a continuous correlation function, the models are jointly estimated. The correlation function only depends on distance and maybe direction. It does not depend on position in space.

Relation to Lattice Models

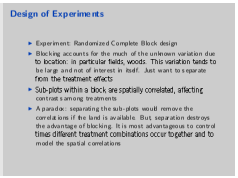
- Data are a finite set of areas that occur in space: counties in North Carolina
- Data occur in space but distance in only determined by whether areas are adjacent or not.
- No ordering to data in space
- Stationarity is still a model assumption: trend can occur over the entire study area.



Using adjacency instead of distance Makes models related to the ARIMA models. The observations or errors are correlated if they are adjacent.Because there is no ordering of data in space, it is possible to have a conditional model where each data point in conditional to all those it is adjacent to.

Design of Experiments

- Experiment: Randomized Complete Block design
- Blocking accounts for the much of the unknown variation due to location: in particular fields, woods. This variation tends to be large and not of interest in itself. Just want to separate from the treatment effects
- Sub-plots within a block are spatially correlated, affecting contrasts among treatments
- A paradox: separating the sub-plots would remove the correlations if the land is available. But, separation destroys the advantage of blocking. It is most advantageous to control times different treatment combinations occur together and to model the spatial correlations



We still need some spatial analysis to know the *range* of the correlations.

Design of Experiments

- Arrangement of treatments controlling the distance between different treatment combinations. Jun Zhu will talk about improving the efficiencies to estimate the correlations due to distance.
- Modeling the spatial correlations in the experiment
- Without modeling the spatial correlations the coverage of the tests (e.g., above time series regression example.)

42

Conclusions

- Repetition is matching pairs of data a given distance apart
- Correlated data can be tranformed back to IID
- Data must be stationary: large scale variation removed
- Correlation affects the amount of information in each observation
- Correlation affects the estimates of standard errors
- Correlation and stationarity are similar in ARIMA and geostatistical models
- Conditional and joint models are similar between ARIMA and lattice models
- Spatial correlations affect arrangement of treatments, and test coverages

43

Geostatistical Data

Matt Kramer

kramerm@ba.ars.usda.gov

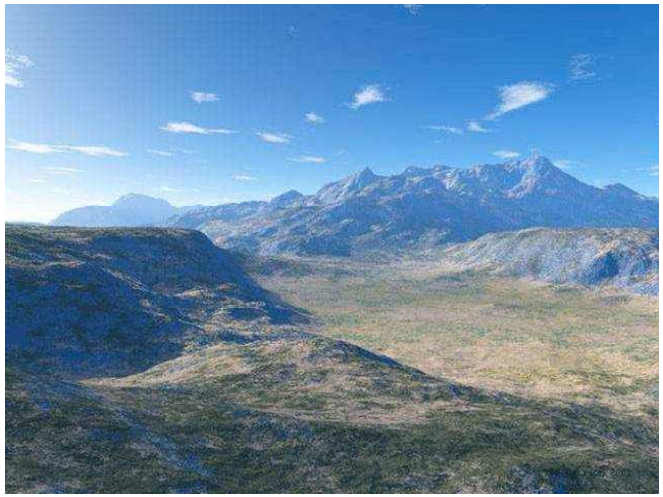
Biometrical Consulting Service, ARS/BARC/USDA

Outline

- Spatial landscapes
- Realizations
- Decomposition of the landscape
- Stationarity
- Variograms
- Ordinary kriging
- Prediction
- Universal kriging
- Important concepts not covered

Spatial landscapes

Fractal terrain by Rolf Lakaemper

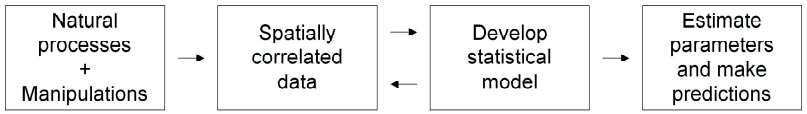


Generating fractal terrain

- The key concept behind fractals is self-similarity
- When a small region of a fractal is magnified, it looks similar to the whole region from which it was taken
- Terrain has this property (loosely defined), which is why fractal algorithms are commonly used to generate “realistic” landscapes
- The property of scale is important for field work, spatial correlation occurs at all scales and how we choose to describe it will depend on the organism (or process) being studied and the crudeness of the tools available.
- We typically classify the variation in the landscape we see into large scale variation, which we might try to explain with regression type variables (e.g. elevation), and small scale variation, which we try to explain using a model of spatial dependency (e.g. kriging).

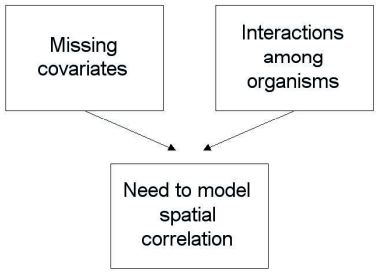
Spatial data as a process

- We observe data generated from some underlying process we are trying to understand
- These data may be observational (e.g. bird counts in a forest) or the researcher may have had a hand in the outcome (e.g. designed experiment where different treatments were applied to various locations)
- We decide on a statistical model that we believe captures the effects we are interested in
- We estimate its parameters and possibly try to interpret them



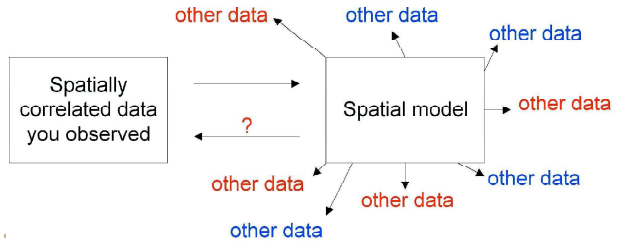
Causes of spatial correlation

- The spatial correlation in the data may be partly (or completely) due to our not having suitable variables to explain why observations closer together are more similar
- Sometimes the spatial correlation is due to interactions among the organisms themselves (e.g. root competition, aggregation), so additional covariates (predictor variables) would not help

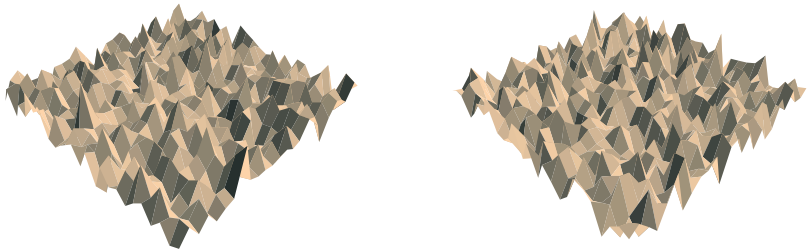


Realizations

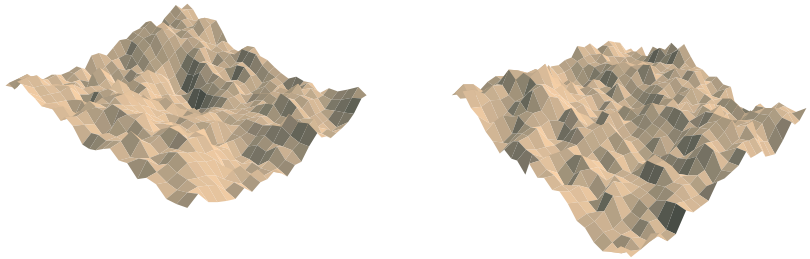
- We have formal models for describing spatial correlation
- We choose one consistant with the spatial pattern of our observations
- The data observed are not unique to that statistical model
- The data are one realization of this statistical model
- Looking at many realizations helps to better understand what kinds of sample data this model can generate



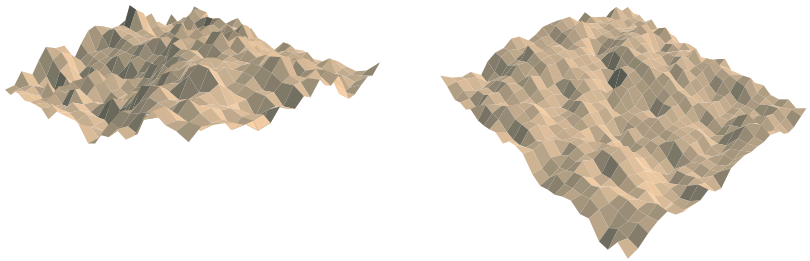
Two realizations of spacial random noise $\sim N(0,1)$



Two realizations of moderately spacially correlated data



Two realizations of strongly spacially correlated data



Realizations—what have we learned?

- spatially correlated observations look “smoother”, some of this is due to scaling
- There are regions of high and low observations with spatial correlation, this pattern may be masked by covariates or treatment effects when looking at “real” data
- You cannot determine the degree of spatial correlation by looking at these plots, we use a tool called the variogram for that
- Strongly spacially correlated data is often symptomatic of a failure to adequately model the “trend” (large scale variation)

Decomposing the landscape: Large scale variation

- Typically thought of as the trend, variation on a scale much larger than distances between observations
- Important to capture all explanatory variables making up the trend, otherwise the residuals may be "non-stationary", which will make modeling small scale variation difficult
- Especially important to capture explanatory variables that vary spatially (spatially varying covariates)
- In designed experiments, blocking is used to capture some of the large scale spatial variation and randomization within the block to reduce the impact of small scale variation
- Large scale variation is typically handled using covariates (e.g. elevation, soil characteristics, latitude and longitude) and ANOVA type variables (e.g. treatments/interventions, historical land use, type of vegetation cover)

Decomposing the landscape: Small scale variation

- Sources of variation not associated with the trend, and at a smaller scale
- Typically imagined to have two components, a smooth function which describes the covariances (correlations) between neighboring observations, and random error (or noise)
- The scale of small scale variation is larger than the smallest distance between observations (typically several times larger)
- What may be considered small scale variation in one study may be large scale variation in another.
- We ignore spatial relationships that occur at scales not captured by our data.

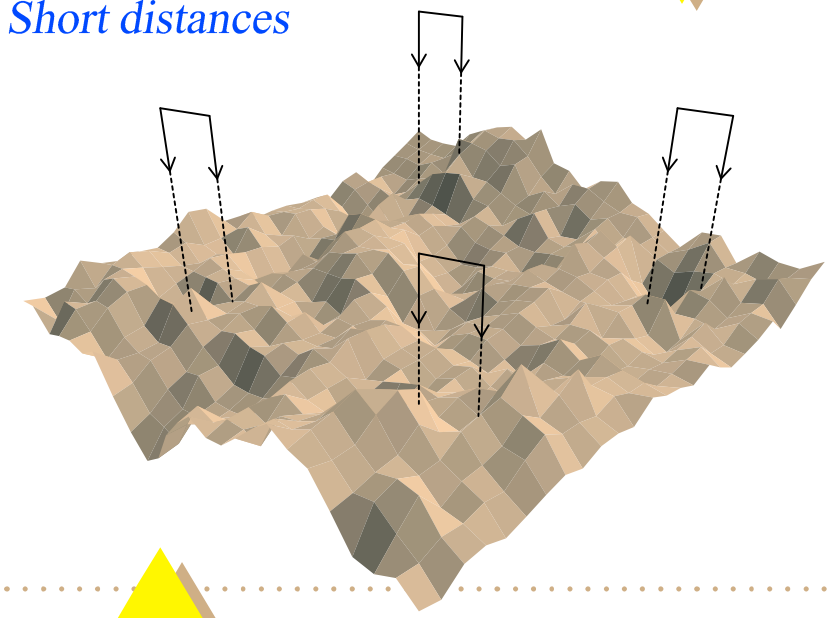
Stationarity

- We need to make simplifying assumptions to model small scale variation
- Spatial correlation necessarily involves pairs of observations
- Data sets with more than 3 observations, have more pairs of observations than observations
- We want the number of parameters in a model to be (far) less than the number of observations.
- In the simplest case, assume spatial relationships between observations are the same everywhere in the landscape, i.e. that the spatial relationships only depend on the distance between observations
- This property is stationarity

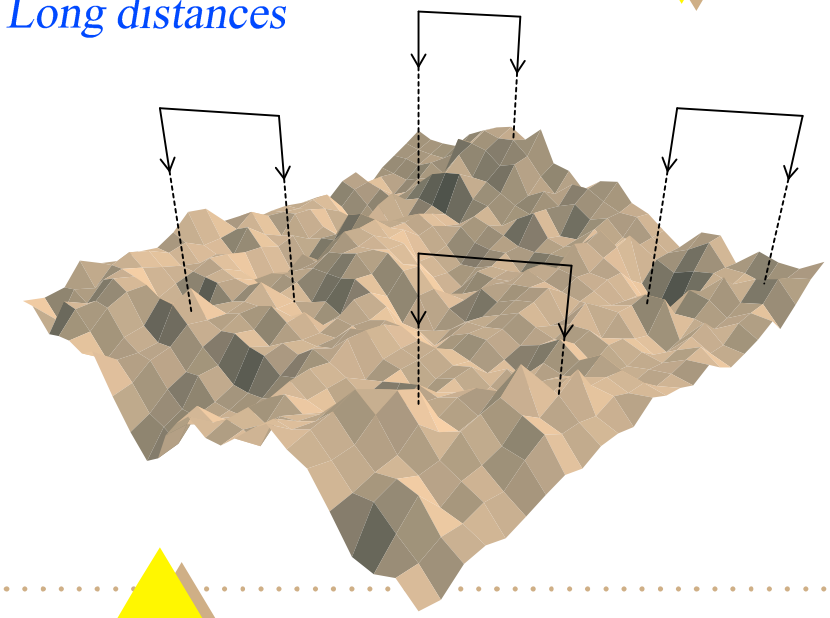
Stationarity

- Often this is not realistic, we may have to allow for spatial relationships to depend on direction (so observations may be more correlated going north to south than east to west), or for them to vary in some other way across the landscape.
- In general, raw data will not be stationary until the large scale variation is removed, so one must first deal with large scale variation before tackling small scale variation
- In the remainder of this presentation, we assume stationarity, but for real data, this would need to be verified.

Short distances



Long distances



Comparison

Comparison of short (0.09 units apart) and long (0.2) distance pairs.

distance	obs. 1	obs. 2	$2\hat{\gamma}(h)_i = (\text{obs. } 1 - \text{obs. } 2)^2$
short	-0.67	0.78	2.10
short	1.47	1.52	0.00
short	-0.82	-0.00	0.67
short	-1.12	-0.38	0.54
long	-0.67	1.40	4.27
long	1.47	2.20	0.54
long	-0.82	0.28	1.22
long	-1.12	-0.40	0.52

$2\hat{\gamma}(h)$ is the classical estimator of the variogram; h is the distance separating the observations

Variogram

If there is small scale spatial autocorrelation, we expect observations near each other to be more similar than ones further away

- This was seen in our example, $2\hat{\gamma}(0.09) = 0.83 < 2\hat{\gamma}(0.2) = 1.64$
- The pattern that emerges, if we plot distance (h) on the x-axis and $2\hat{\gamma}(h)$ (or $\hat{\gamma}(h)$, the semivariogram) on the y-axis, should tell us something about small scale variation
- $\hat{\gamma}(h)$ should be small when the distance h is small, $\hat{\gamma}(h)$ should be larger as the distance h increases
- What is the best way to do this?

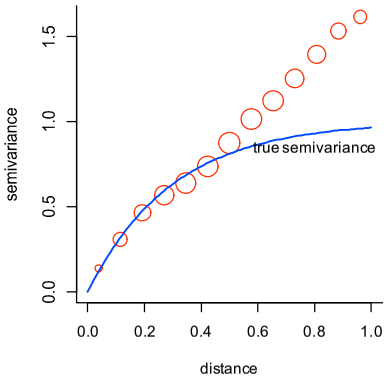
Variogram

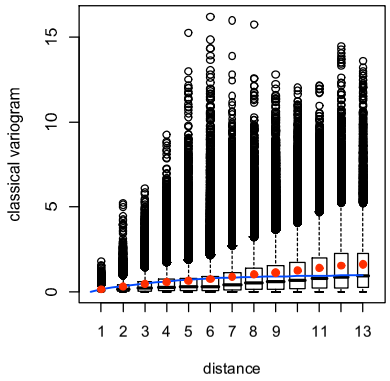
- If we look at the distribution of pairs of observations by distance apart, we find that there are far fewer pairs of observations separated by large distances
- Thus, our estimates $\hat{\gamma}(h)$ for h large will not be as good as $\hat{\gamma}(h)$ for h smaller
- If our data are not evenly spaced, we may find the same problem for h very small, there may only be a few pairs that represent the smallest distances
- This means that some regions of the semivariogram have better support than others

Variogram

- To create the semivariogram, we break h up into many distance groups (e.g. 0–0.2, 0.2–0.4, 0.4–0.6, etc.) and calculate $\hat{\gamma}(h)$ for each distance group.
- Then we can plot the average value of h for that distance group against $\hat{\gamma}(h)$
- We can also plot $\hat{\gamma}(h)_i$ for each pair of observations, this may help us decide if the average value for each h is a reasonable estimate of what the "mean" should be
- In practice, we have software that does this, though we may make decisions about how large an interval each distance group should be, and what our largest h should be (since beyond a certain h results will be rather flaky as there aren't many pairs of observations for very large h)

Variogram ($\hat{\gamma}(h)$ vs. h)





Variogram: What have we learned?

- The variogram nicely displays the similarity of neighboring observations, and how differences between observations increase with increasing distance
- Even with $n = 676$ observations, the empirical semivariances do not follow the true semivariances beyond $h = 0.5$ units (distance between the two furthest observations is 1.4 units)
- These data were generated from a known model (where we know the true parameters), yet there are still problems with the variogram
- We could regenerate data sets from this model until we created one that produced a nice variogram, but one cannot do that for "real" data

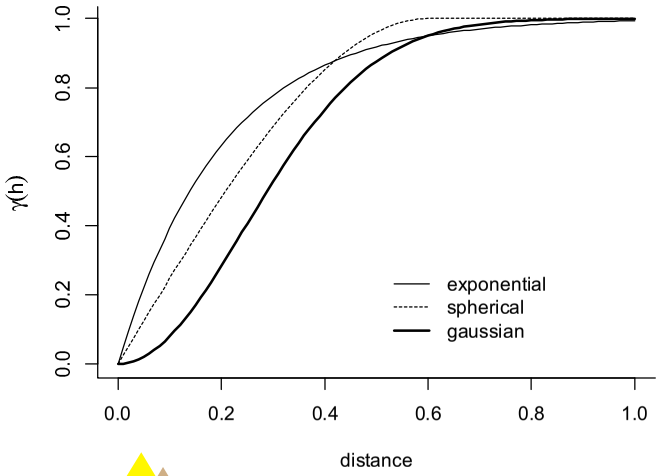
Variogram: What have we learned?

- The box plots show how variable the individual semivariance estimates are for each distance class
- The variogram is an imperfect tool, but in practice it works well
- There are robust procedures for estimating the variogram

Variogram—what model to use?

- Software for modeling spatial data will have many different models that one can use to capture the spatial autocorrelation
- These models differ in how the strength of the correlation between observations diminishes as distance between them increases
- The data for this example were generated using an exponential model
- Many of the models produce very similar results (and you might need a lot of data to be able to discriminate between models)
- It is more important to try to capture the spatial dependencies with some model, even if you aren't sure it is the "right" model, then to ignore the spatial dependencies completely.

Three common variogram models



Variogram—estimation of model parameters

- Once we have decided on a model for the data, we need to estimate its parameters
- Many variogram models have parameters (or combinations of parameters) that can be interpreted as the range, sill, and nugget (these terms show geostatistics' mining origin)
  - The range is the minimumn distance separating observations that are (nearly) spatially independent
  - The sill is the value of $\gamma(h)$ when $h =$ range
  - A nugget effect occurs if, as h (the distance between observations) goes to zero, $\gamma(h)$ does not approach zero
  - The partial sill $=$ sill $-$ nugget

Variogram models

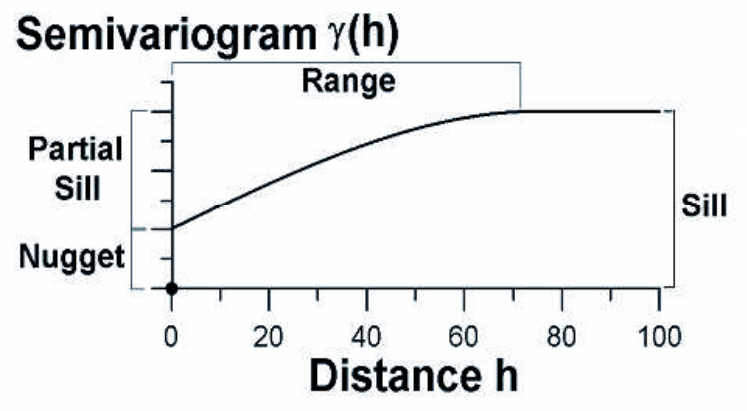
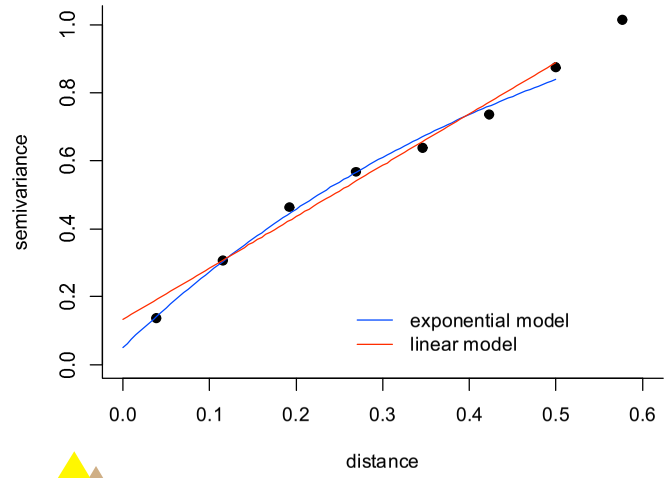


Image by Jay Ver Hoef

Variogram—estimation of model parameters

- A least squares approach (i.e. regresson equation) is common
- The least squares approach is usually modified so that it gives more weight to small h (where it is most important to have a good fit) and to areas of the variogram that have the most pairs of observations
- Robust methods have also been developed
- The software typically does this fitting, you only select the model you want to use and options for how to do the fit
- You then plot the graph against the variogram estimates (the averaged or "binned" estimates, one for each distance category) to check the fit visually

Variogram model parameters



Variogram model parameters

- Two models were fit, exponential and linear, to the data up to $h = 0.5$.
- Note: These fits look good only because the distance was cut off at $h = 0.5$!
- Estimates for the variogram model parameters, nugget, partial sill, range:
  - Exponential: τ^2 = 0.12, σ^2 = 1.28, ϕ = 0.52
  - Linear: τ^2 = 0.13, σ^2 = 1.51, ϕ = 1.00

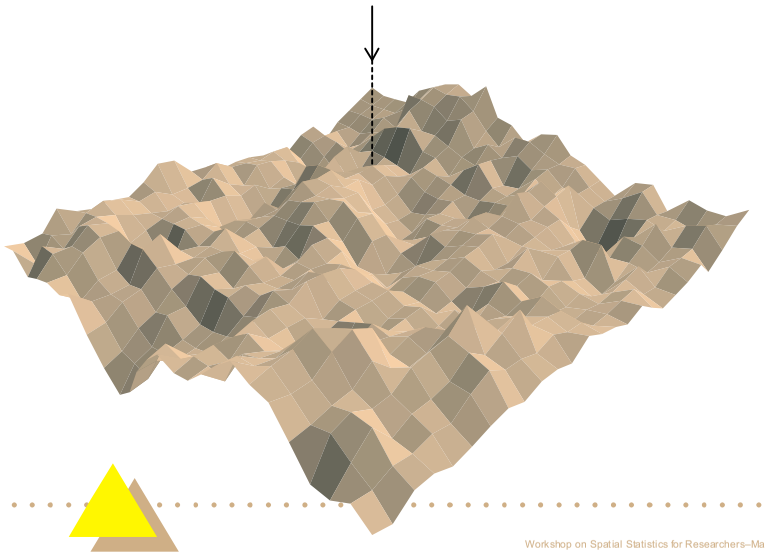
Ordinary Kriging

We now have a model for the spatial dependencies in our data.

- We can estimate a value at a particular location (which should be within the general area in which the data were collected!)
- In this case, the uncertainty associated with the estimate will depend on how far the location is from real observations and how much spatial correlation exists
- If the location is further from any real observations than the range, we get no "special" information from nearby observations and the best estimate will be the mean
- Unlike, e.g. regression, a prediction at a location where we have an observation just gives us back the value of the observation
- This is a technique that can be used for observations that are unequally spaced as well regularly spaced (the example used here is for regularly spaced data)

Prediction at $(x = 0.27, y = 0.27)$

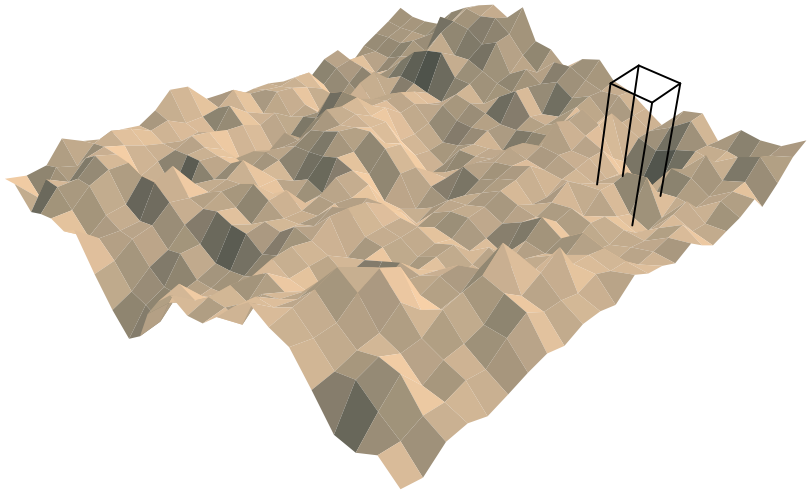
point estimate = −0.376, kriging variance = 0.044



Predict a region

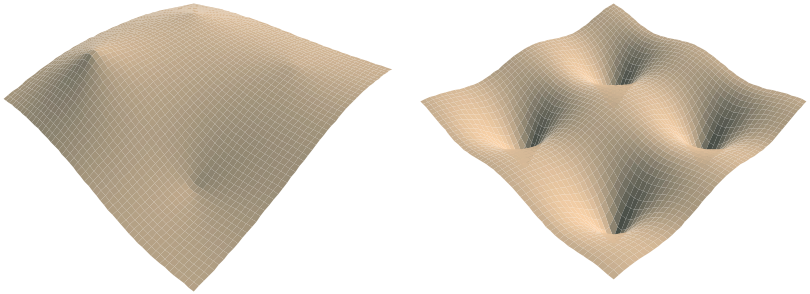
- We can also create an estimate for the region (or some subset of the region) in which the data were collected, e.g. the average value
- The uncertainty associated with this estimate will depend on the density of real observations in the region and how much spatial correlation exists
- These kinds of estimates are performed by software, we need to specify the model and what output we want

Predict a region



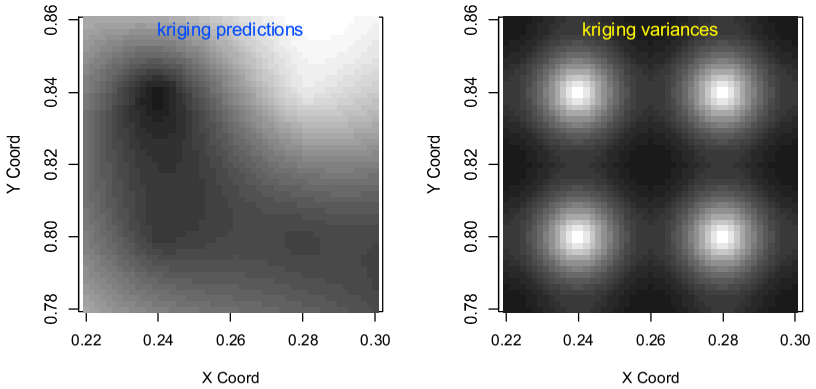
Predictions & variances—perspective view

- Left plot: krigged surface (note how smooth it is!)
- Right plot: kriging variances (variance is zero where data were taken unless there is a nugget effect)



Predictions & variances—typical output

Relative prediction and variance values coded by intensity (black = large values, white = low values)

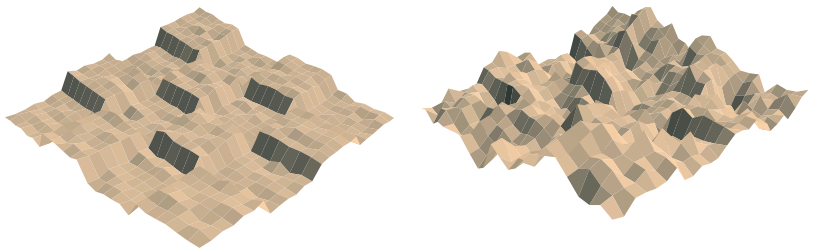


Universal Kriging—estimation strategy

- We often have other information about the landscape we are modeling, such as covariates or factors (e.g. treatment effects), in which case we have a mixed model
- If we can subtract out these effects, then we can use the strategy just discussed to model the spatially correlated residuals
- For the most common geostatistical models, mixed models software can estimate all the parameters of the model (covariates, factors, spatial covariance parameters)
- Unfortunately, there are deficiencies in the software (limited spatial models, lacking good diagnostics)

Universal Kriging—trend and noise

- Left plot: trend (covariate + two-level factor) (note: covariate effect not easy to see because it, in part, tilts the plane surface)
- Right plot: trend + noise (noise = spatially correlated residuals)



Universal Kriging—estimation strategy

- Added a covariate and factor effect to the spatially correlated observations
- We assume the spatial correlation is unrelated to these effects
- If we had no idea of the pattern of spatial correlation (of the residuals), we might start out by
  - assuming that residuals are uncorrelated and estimate the covariate and factor effect using a linear model
  - subtract out their effects from the data
  - determine if the residuals are stationary, and if so
  - use a variogram to determine their pattern of spatial covariance
  - re-estimate the model using mixed models software

Universal Kriging—estimate trend

- Although we already know the function to use for spatial correlation of the residuals, we'll pretend we don't
- First estimate the trend assuming uncorrelated residuals.

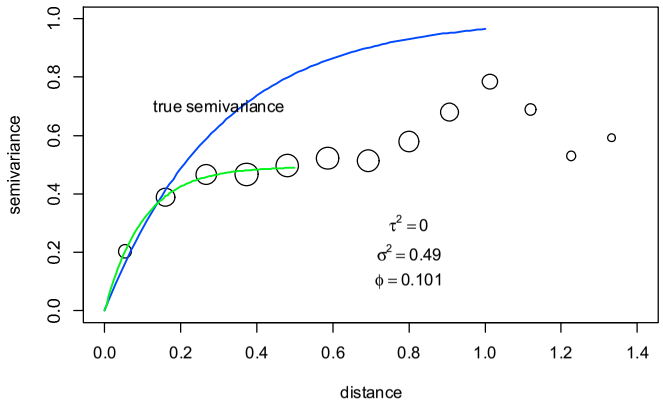
```
> fit1 <- lm (dat1 ~ as.factor(f1) + covar1 - 1)
> summary(fit1)

               Estimate Std. Error t value Pr(>|t|)
as.factor(f1)0 -0.59867    0.05293 -11.310  < 2e-16 ***
as.factor(f1)1  0.22169    0.05497   4.033 6.13e-05 ***
covar1          1.75290    0.03131  55.988  < 2e-16 ***

Residual standard error: 0.7184 on 673 degrees of freedom
```


Universal Kriging—model noise

Semivariogram of the residuals



Universal Kriging—estimate full model

R software, *geoR* package

```
gdat2 <- as.geodata(cbind(x,y,dat1))

ts1 <- trend.spatial(trend= ~ as.factor(f1) + covar1 - 1)

fit2REML <- likfit (gdat2, trend=ts1, ini.cov.pars=expfit2$cov.pars,
            fix.nugget = FALSE, cov.model="exp", method.lik = "REML")
```

Universal Kriging—estimation results

```
 beta0  beta1  beta2
0.2051 1.2540 1.0864

Parameters of the spatial component:
   correlation function: exponential
      (estimated) variance parameter sigmasq (partial sill) =  1.118
      (estimated) cor. fct. parameter phi (range parameter)  =  0.3689
Parameter of the error component:      (estimated) nugget =  0

> sqrt(diag(fit2REML$beta.var))
0.5106032 0.5107846 0.1098927
```

Estimates ignoring spatial correlation:

```
                Estimate Std. Error t value Pr(>|t|)
as.factor(f1)0 -0.59867    0.05293 -11.310  < 2e-16 ***
as.factor(f1)1  0.22169    0.05497   4.033 6.13e-05 ***
covar1          1.75290    0.03131  55.988  < 2e-16 ***
```


Universal Kriging—estimation results

These results closely match those using the *nlme* R package:

```
> fit3 <- gls (dat1 ~ as.factor(f1) + covar1 - 1, corr =
               corExp(c(1,0.1), form = ~ x + y, nugget = TRUE))
> summary(fit3)
Generalized least squares fit by REML
Correlation Structure: Exponential spatial correlation
 Formula: ~x + y
 Parameter estimate(s):
       range       nugget
3.688905e-01 3.302637e-09

                    Value Std.Error  t-value p-value
as.factor(f1)0 0.2050859 0.5105995 0.401657  0.6881
as.factor(f1)1 1.2539898 0.5107810 2.455044  0.0143
covar1         1.0863683 0.1098926 9.885727  0.0000
Residual standard error: 1.057395
```


Universal Kriging—model comparison

Comparison of results from ignoring spatial correlations versus incorporating them into the model

- for the fixed part of the model (covariate + factor), parameter estimates and standard errors differ
- differences in parameter estimates are not that large once centering has been taken into account
- standard errors are much larger for model with correlated residuals, this shows that ignoring spatial autocorrelation produces incorrect tests on factors (e.g. treatment effects)
- estimation time for the linear model was < 1 sec., for the model with autocorrelated residuals, > 10 min. ($n = 676$)

Important concepts not covered

- Isotropy—anisotropy
- non-Euclidean distance measures
- Diagnostics
- Transforming data that are not normal
- Robust methods
- Variances/standard errors for kriged estimates

THE END

Field-Scale Spatial Variability: Yield Response of Potatoes

Rose Shillito

Crop Systems and Global Change Lab., USDA-ARS, Beltsville, MD
Natural Resource Sciences and Landscape Architecture, University of Maryland, College Park, MD

March 15 – 16, 2006

1

First Law of Geography

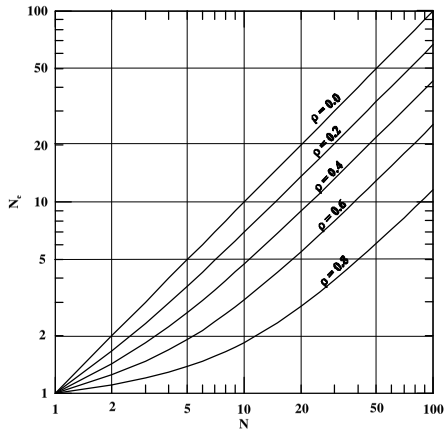
- All things are related, but nearby things are more related than distant things.

W.R. Tobler, 1970

2

Effective number of observations

- The effective number of observations, N_e, is related to the number of observations, N, as a function of autocorrelation, ρ.
- A 50-year record with ρ=0.2 contains as much information as a 33-year record with ρ=0.0.



3

Descriptive Statistics for Spatial Studies

- (auto)covariance function

$$C(h) = \mathrm{cov}\left[A_i(x), A_i(x+h)\right] = \frac{1}{N}\sum_{i=1}^{N-h}\left[A_i(x_i) - \overline{A}\right]\left[A_i(x_i+h) - \overline{A}\right]$$

- autocorrelation function

$$\rho(h) = \frac{\mathrm{cov}\left[A_i(x), A_i(x+h)\right]}{\sqrt{\mathrm{var}\left[A_i(x)\right]}\sqrt{\mathrm{var}\left[A_i(x+h)\right]}}$$

- variogram

$$\gamma(h) = \frac{1}{2N(h)}\sum_{i=1}^{N(h)}\left[A_i(x_i) - A_i(x_i+h)\right]^2$$

where

$A_i(x_i)$ = value of A_i measured at x_i

h = distance

4

Descriptive Statistics for Spatial Studies

These statistics are commonly used in spatial studies. They indicate the degree that the data at any two points are related to each other and, thus, give some indication of non-independence of the data.

They are shown here as a function of distance, h, between any two points, and are omnidirectional. Directional bounds can be specified such that only data points within a specified radius will be considered.

These terms apply to univariate spatial studies. In multivariate spatial studies, the prefix “cross” is frequently used (i.e., cross variogram).

The variogram is a fundamental metric in geostatistics and is related to the other measures.

Developments and Issues

- Geostatistics exploits fundamental autocorrelation in data (Matheron, 1970; etc.)
- Issue: pseudoreplication (Hurlbert, 1984)
- Issue: information being lost by not experimenting and measuring as a landscape continuum (Peterson et al., 1993)
- Issue: computation intensity no longer an impediment; emphasis on design of spatially efficient experiments (Edmonson, 2005)

6

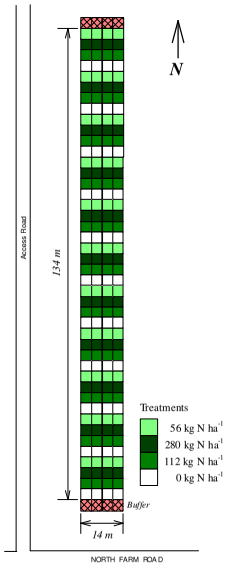
Objectives

- Simulate a more realistic, gradually varied treatment design for potato response to nitrogen.
  - continuous
  - field-scale
- Correctly test for treatment effects in the presence of spatial variability.
- Describe the effect of field properties in yield response.

7

Experimental Field

- BARC-W, Maryland
- 0.18 ha (135 m x 14 m)
- Experimental unit: 3 m x 3 m
- Transect: 44 units
- Field: 4 transects
- Potatoes planted DOY 113 (April 23, 2003; April 22, 2004)
- Planting density 3.6 plants m^{-2}
- Buffers
  - 3 m at N and S ends
  - 1 row along edges



8

Experimental Field

- Calcium nitrate applied at 22 DAE (2003) and 17 DAE (2004)
- 4 levels
  - 0 kg N ha^{-1}
  - 112 kg N ha^{-1}
  - 280 kg N ha^{-1}
  - 56 kg N ha^{-1}
- Constant across field width
- Sinusoidal pattern along field length
- No irrigation
- Potatoes harvested 118 days after planting

9

Experimental Field

10

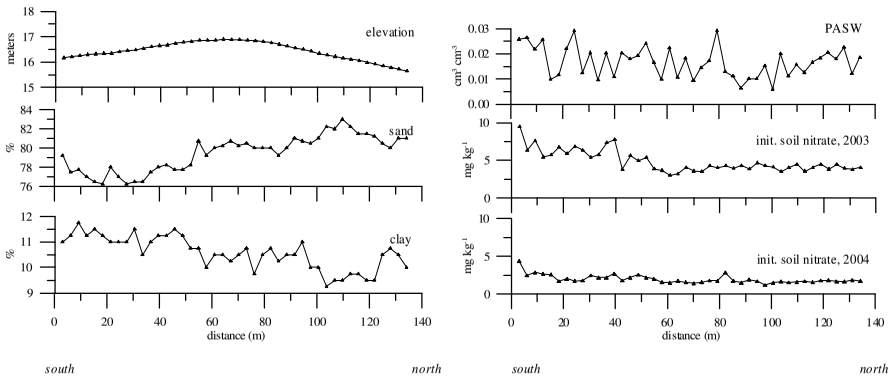
Experimental Field

Image of potato field taken in July, 2003.

There were no noticeable disease impacts, and pests and weeds were controlled throughout the 2003 and 2004 growing seasons.

A rye cover crop was planted in the field prior to both the 2003 and 2004 experiments. The rye was mechanically plowed under while the field was chiseled and disked during field preparation prior to planting.

Transects of Field Properties



12

Transects of Field Properties

Because the field was long and narrow, data gathered over the field were averaged into a transect for analysis.

Field topography was sampled via a real-time kinematic GPS survey at an approximate spacing of 1 point per 2.7 meters.

A soil probe was used to extract a 15-cm sample of the surface soil from the center of each of the 176 plots for particle size analysis and to determine initial pre-application soil NO3-N.

Undisturbed soil cores (5.4 cm dia. x 6.0 cm len.) were collected from the center of each unit of one field transect (44 units) to determine plant available soil water capacity (PASW). PASW was determined as the difference between volumetric water contents at matric potentials of -0.01 MPa and -1.5 MPa.

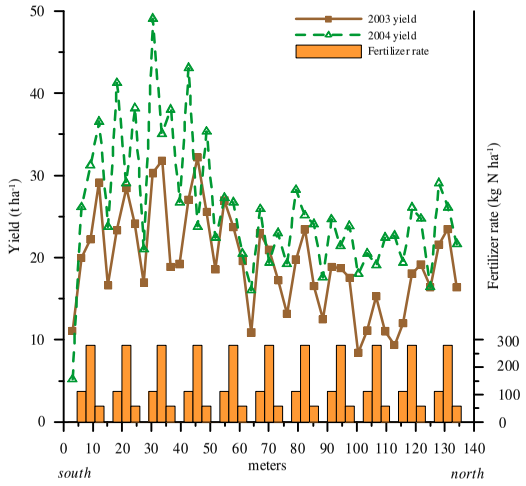
Correlation of Field Properties

	Elev	Sand	Clay	PASW	InitN03	InitN04
Elev	1.00	-0.24	0.24	-0.10	-0.10	-0.02
Sand		1.00	-0.81**	-0.19	-0.65**	-0.44**
Clay			1.00	0.20	0.59**	0.54**
WHC				1.00	0.22	0.27
InitN03					1.00	0.73**
InitN04						1.00

**Significant at the 0.01 probability level.

14

Transects of Yield



15

Mixed Model Analysis

General linear mixed model

$$\mathbf{y} = \mathbf{x}\beta + \mathbf{Zu} + \mathbf{e}$$

$$Var(\mathbf{e}) = \mathbf{R}$$

Spatial definition of **R** (SAS)

$$Cov(e_i, e_j) = f\left(\sigma^2, h, \rho\right)$$

$$\sigma^2 = \sigma_p^2 + \sigma_n^2$$

σ^2 = variance; h = distance between e_i and e_j; ρ = range; σ_p^2 = partial sill; σ_n^2 = nugget

16

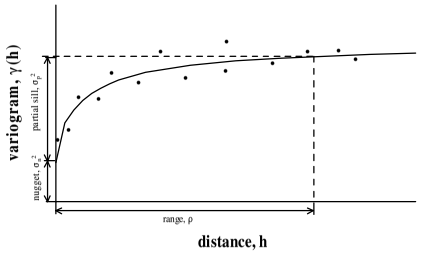
Mixed Model Analysis - 1

General Linear Mixed Model:

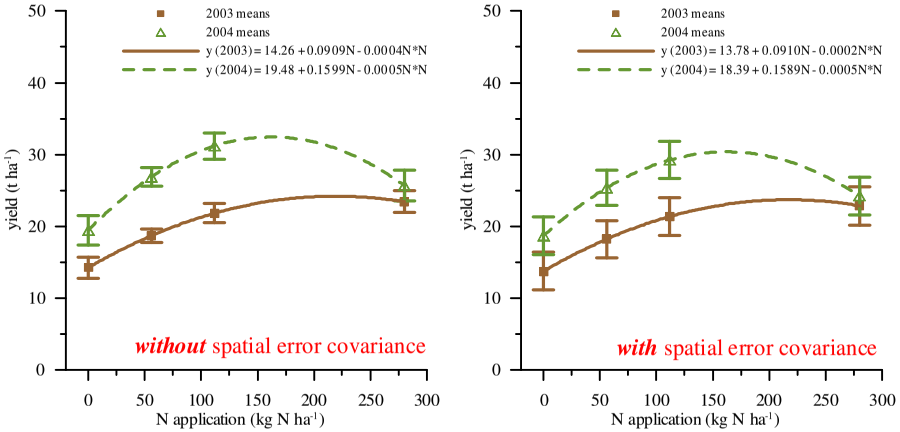
xβ = fixed effect(s)
Zu = random effect(s)
e = random errors

Since the data were effectively contiguous, no blocking was necessary and the random effects were not considered. The data for 2003 and 2004 could have been considered a random (year) effect, but two years of data does not allow for reasonable variance calculations.

In SAS, the components of the covariance matrix are output in terms of the variogram. But the data considered in the covariance are the residuals after the fixed effects have been taken into account.



Nitrogen Treatment Response – Means and Standard Errors



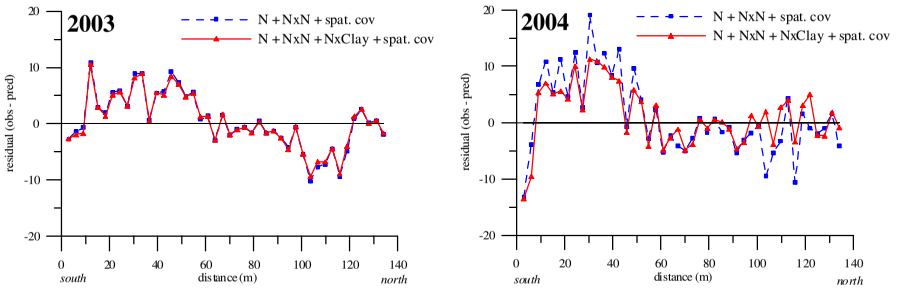
18

Mixed Model Analysis

Model	Non-spatial R^2	Spatial R^2
2003		
N treat + N treat²	0.34	0.68
N treat + N treat² + N treat x Clay	0.40	0.70
2004		
N treat + N treat²	0.28	0.54
N treat + N treat² + N treat x Clay	0.59	0.69

19

Analysis of Residuals



20

Soil Type

- KmA
  Keyport and Matawan Soils, 0 – 2% slopes
  sandy loam, silt loam
- D1B
  Downer-Ingleside Loamy Sands, 2 – 5% slopes
  loamy sand
- MwkB
  Matawan and Keyport Soils, 2 – 5% slopes
  loamy sand, silt loam

Special Soil Report, 1995

21

Mixed Model Analysis

Model	Non-spatial R^2	Spatial R^2
2003		
N treat + N treat^2	0.34	0.68
N treat + N treat^2 + N treat x Clay	0.40	0.70
N treat + N treat^2 + N treat x Soil Type	0.66	0.74
N treat + N treat^2 + N treat x Clay x Soil Type	0.67	0.74
2004		
N treat + N treat^2	0.28	0.54
N treat + N treat^2 + N treat x Clay	0.59	0.69
N treat + N treat^2 + N treat x Soil Type	0.53	0.64
N treat + N treat^2 + N treat x Clay x Soil Type	0.63	0.71

22

Mixed Model Analysis - 2

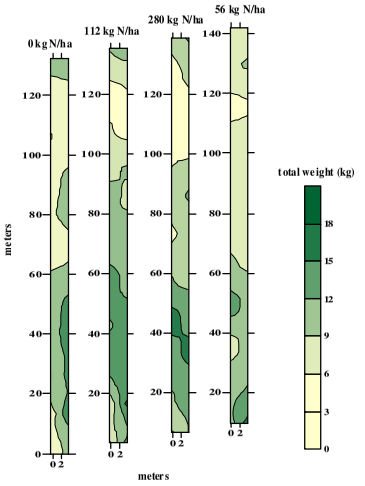
The coefficient of determination (R^2) was calculated for various mixed models developed for the data. The quadratic model (N treatments + N treatments^2) was considered the base model—the nitrogen response curve. The only significant field variable (as determined by backward elimination regression analysis) was clay.

The yield residuals (observed yield – predicted yield) exhibit some spatial patterning. Including the N treatment x clay interaction decreases the residual variability, especially at the north end of the field in the 2004 data.

Other interactions (e.g., soil type—a classification variable) were tested although not developed through significance testing or AIC minimization.

Yield Estimation

- Interpolated yield estimates using kriging
- High yields at one end of field; low yields at other end
- Poor yield response to fertilizer where clay and init. soil nitrate low



24

Is there more?

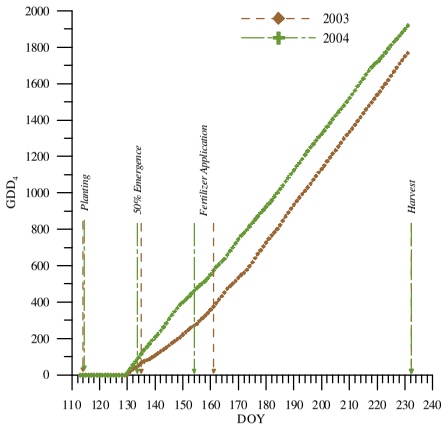
Weather Data for Growing Season

	Avg. Daily Temp. (ºC)	Total Precip. (mm)
2003	20.6	522
2004	21.8	509

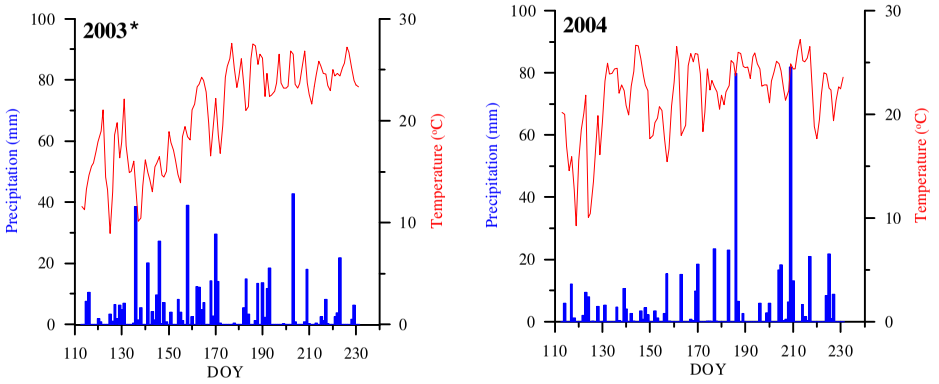
Growing Degree Days

$$GDD = \sum_{i=1}^{n} (T_i - T_b), \quad i = 1,...,n$$

T = avg. daily temp.
b = base temp. = 4℃



Growing Season Weather



*wettest year on record!

Summary and Conclusions - 1

- Field properties varied throughout field.
- Yield response varied throughout field.
- Yield response to treatments varied throughout field.
- Spatially correlated errors made treatment means less distinct.
- The linear association between yield and treatments increased if spatially correlated errors were considered.

Summary and Conclusions - 2

- The effect of field properties (continuous and classed) was tested; clay content and soil type class both proved significantly related to yield.
- Residuals still exhibited spatial variability throughout field.
- Pattern of yield response similar both years; magnitude of yield will require management and climatic inputs.
- Treatment application pattern allowed for systematic testing of all treatments throughout field, effectively increasing experimental design by four.

References

- ***Cited Works***

  Edmonson, RN. 2005. Past developments and future opportunities in the design and analysis of crop experiments. J. Agric. Sci.

  Hurlbert, SH. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Mongr.

  Matheron, GF. 1970. La théorie des variables régionalisées et ses applications. Ecole des Mines de Paris, Fontainebleau.

  Peterson, GA, DG Westfall, and CV Cole. 1993. Agroecosystem approach to soil and crop management research. SSSAJ.

- ***Spatial Variability and Experimentation***

  Zimmerman, DL, and DA Harville. 1991. A random field approach to the analysis of field-plot experiments and other spatial experiments. Biometrics.

  van Es, HM, and CL van Es. 1993. Spatial nature of randomization and its effect on the outcome of field experiments. Agron. J.

  Hoosbeck, MR, A Stein, H van Reuler, and BH Janssen. 1998. Interpolation of agronomic data from plot to field scale: Using a clustered versus a spatially randomized block design. Geoderma.

  Hong, N, JG White, ML Gumperz, and R White. 2005. Spatial analysis of precision agriculture treatments in randomized complete blocks: Guidelines for covariance model selection. Agron. J.

- ***Spatial Mixed Models***

  Littell, RC, GA Milliken, WW Stroup, and RD Wolfinger. 1996. SAS System for Mixed Models. SAS Institute Inc., Cary, NC.

Acknowledgements

- *Farm crew at USDA-ARS Beltsville Area Research Center-West, Beltsville, Maryland*
- *Technicians, scientists, and staff of the USDA-ARS Crop Systems and Global Change Laboratory, Beltsville, Maryland*
- *Students and faculty of the University of Maryland, College Park, Maryland*
- *Staff of the Biometrical Consulting Service, USDA-ARS, Beltsville, Maryland*

Lattice Models with Spatial Dependencies - An Introduction

Mary C. Christman
Univ. of Florida
Department of Statistics - IFAS

Lattice Models

- Area of interest is subdivided into mutually exclusive and exhaustive plots, strata, or subareas
- Data are aggregated or summary values for each subarea

EXAMPLE: Sudden Infant Death Syndrome statistics for counties in North Carolina in the 1970s

Additional Comments

- Note that unlike geostatistical modeling, in lattice models there is no concept of interpolating between plots or subareas.
- As a result, we are less interested in mapping and more interested is modeling such as regression with correlated data or mixed models with covariance matrices that are not diagonal

Questions

- Classic models are used to test hypotheses about explanatory variables (factors, covariates, etc)
  - Q: Should we worry about spatial autocorrelation?
  - If so, how should the spatially-explicit aspect be incorporated into our modeling effort?
- When planning a study, need to address:
  - Spatial arrangement of treatments if planned experiment
  - Spatial arrangement of plots when observational study

Additional Comments

- Traditionally, the spatial autocorrelation that was presumed to be a potential problem was handled in experimental designs using such techniques as blocking
  - E.g. the Average Distance Balanced Design in which treatments are arranged spatially so that the average distnace between plots of different treatments is approximately constant over all treatments

Classic Model Assumptions

- For General Linear Models
  - Error terms are Normally distributed with constant mean ($\mu = 0$) and variance (σ^2) and
  - Error terms (and hence the responses) are independent
- For Generalized Linear Models
  - Response Variable is distributed appropriately (usually Binomial, Poisson or similar) with a mean that is a function of covariates ($\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$) and variance that depends on the mean.
  - The responses are independent

Additional Comments

- Even with restricted randomization methods to account for spatial arrangement of locations,
  - there may still be spatial autocorrelation and hence the error terms/response variables are not independent
  - and so classical assumptions fail.

Failure of the Independence Assumption

- Due to non-spatial issues such as sampling design
  - E.g. blocking, clustering or temporal effects
- Due to Spatial autocorrelation
  - Correlation between 2 values of the response variable, $Y(s_i)$ and $Y(s_j)$ at locations s_i and s_j, is non-zero and a function of distance
  - How does it arise?

Additional Comments

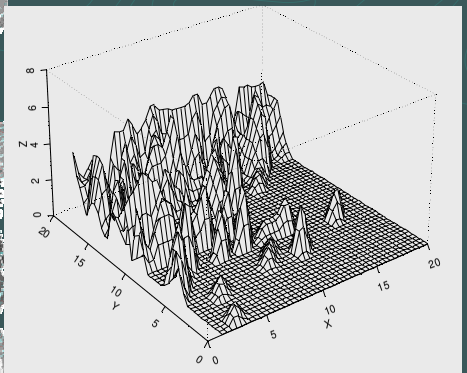
- Non-spatial lack of independence is handled as usual, e.g. random blocks or time series.
- Spatial lack of independence is handled using autocorrelation covariance matrices that require additional information
  - form of the non-independence (as a function of distance),
  - neighborhood structures, etc.
- Note: I assume that distance is Euclidean unless otherwise specified

Sources of Spatial Autocorrelation in Y

- Induced
  - Values close in space could be similar due to an important explanatory variable that varies smoothly in space
    1. E.g. The spatial distribution of bell pepper fungus in a field
       - could be due to spatial distribution of soil moisture
       - could be due to geography (e.g. elevation changes)

Example - Bell Pepper fungus

Leaf Disk Assay



See Larkin et al. 1995. Phytopath 85:191-203

Field plot was subdivided into 400 1x1 m subplots.

In each subplot, 5 leaf disks were assayed for presence of fungus.

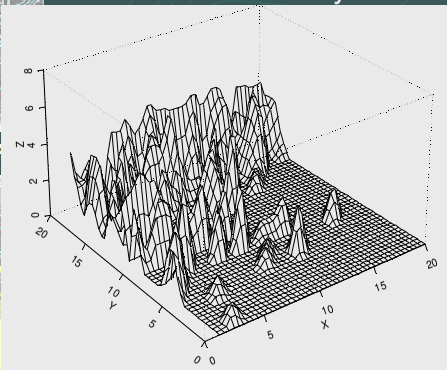
Recorded number that tested positive.

Additional Comments

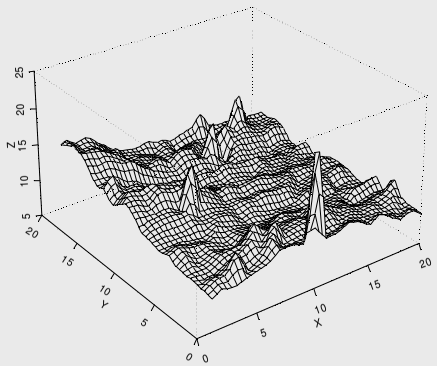
- Graph shows the number of leaf disks assays that tested positive for fungus (out of 5) for each 1x1 m plot within the study area.
- Note the trend (low in SE corner, high in NW corner) as well as grouping of similar values spatially.
- The next slide shows that the pattern may be related to soil moisture, i.e. the spatial patterns show similarity. Is it possible that moisture is a partial predictor for fungus presence?

Example - Bell Pepper fungus

Leaf Disk Assay



Soil Moisture



Sources of Spatial Autocorrelation in Y

- True
  - Intrinsic, underlying covariance that is a function of distance
    - E.g. for the spatial distribution of soil moisture, it could be due to soil characteristics that allow water movement into and through adjacent plots
  - Causal interaction among nearby locations
    - E.g. The spatial distribution of leaf fungus could be due to dispersal mechanism
      - Leaves touching vs. air dispersal

Sources of Spatial Autocorrelation in Y

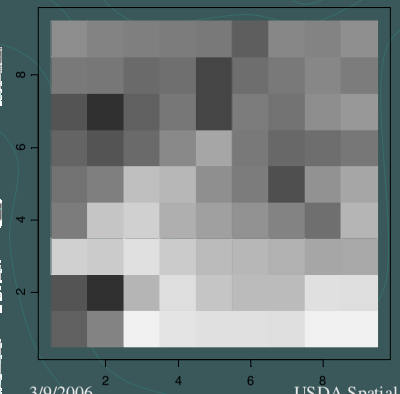
- Spurious
  - Values close in space could be similar due to chance
    - E.g. due to the spatial arrangement of the sampling locations
    - E.g. due to smoothing of the data during preliminary data management
    - E.g. due to the scale at which the data have been aggregated

Additional Comments

- Spurious autocorrelation is unlikely for the bell pepper fungus dataset since plots are small and there is no data manipulation prior to analysis.
- Spurious autocorrelation is the hardest to capture and identify.
  - An example would be in precision agriculture due to the slight delay in recording soil attributes. The recording device often has a delay of 3-4 seconds but the location is recoded not where the data were collected but where the recorder reports the value.
  - See this sometimes in satellite images as well due to interpolation for pixel data

Example

Reflectance Values From An Areal Survey of Pollution Levels Due To Pumping Of Waste Material Into The English Channel



Darker areas represent more polluted spots. This location is closest to the source of the pollution.

Values in any one grid cell are averages over the cell and, due to location error, possibly include values in neighboring cells as well.

Autoregressive Lattice Models

$$Y(s_i) = \mu(s_i) + \sum_{s_j \in N_i} \varpi_{ij}[Y(s_j) - \mu(s_j)] + \varepsilon(s_i)$$

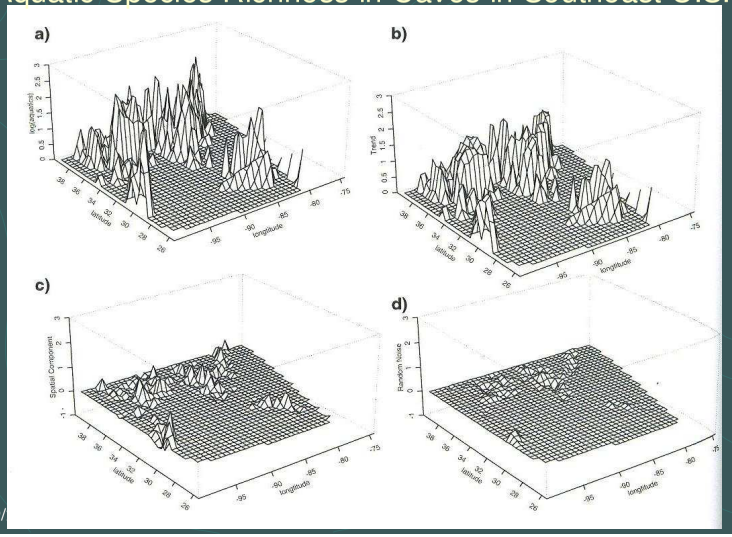
- $Y(s_i)$ is the response variable at location s_i
- $\mu(s_i)$ is the large-scale trend or mean for location s_i
  - may depend on explanatory variables or treatments
- $\sum \varpi_{ij}[Y(s_j) - \mu(s_j)]$ small-scale variation at location s_i
  - Depends on the values in the neighborhood N_i and weights ϖ_{ij}
- $\varepsilon(s_i)$ is the error term, conditionally independent with zero mean and constant variance

Additional Comments

- The large-scale mean is usually dependent on explanatory variables such as covariates or treatment levels or even location (such as a trend surface that is a polynomial in space.
- The small scale variation can be used to calculate the conditional mean, that is the predicted value at a location using the covariates at a location and the values of observations around that location. The conditional mean is the sum of several parts: 1) the mean of the individual subplot, $\mu(s_i)$; 2) the weighted average of the error terms for all of the neighboring subplots.

Example of the Decomposition

Aquatic Species Richness in Caves in Southeast U.S.



Additional Comments

- These perspective plots show the decomposition of species richness values in counties throughout the southeast US.
- $Y(s_i)$ are shown in (a), a plot of the observed values of log (aquatic species richness) in counties in the southeast US.
- $\hat{\mu}(s_i)$ are shown in (b) a plot of the estimated county means of log (aquatic species richness) predicted by the explanatory variable, X=number of caves found in the county.
- $\sum \varpi_{ij}[Y(s_j) - \hat{\mu}(s_j)]$ are shown in (c), a plot of the estimated small-scale variation in each county based on observations of log (aquatic species richness) in contiguous counties. The weights were $w_{ij} = 1$ if counties i and j were contiguous and $w_{ij} = 0$ if they were not.
- $\hat{\varepsilon}(s_i)$ are shown in (d), a plot of the unexplained or residual noise.
- Note that the values in (b), (c) and (d) add up to the observed values shown in (a).

Large-scale Variation $\mu(s_i)$

- Could be a function of factors being manipulated in a planned experiment
  - E.g. a split-plot design with a whole plot factor of crop rotation schedule and a subplot factor of nitrogen source
- Could be explanatory variables being observed
  - E.g. soil moisture in the bell pepper fungus study
  - E.g. the number of caves in a county to predict the species richness of aquatic subterranean animals

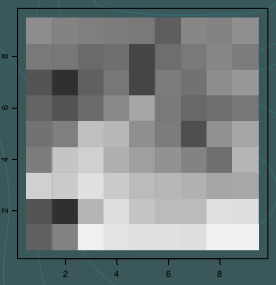
Small scale Variation

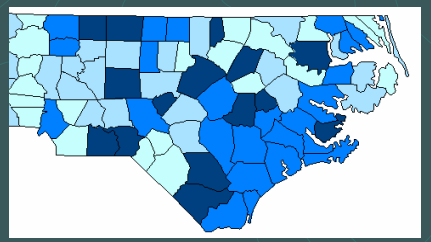
$$\sum_{s_j \in N_i} \varpi_{ij}[Y(s_j) - \mu(s_j)]$$

- Two parts
  - Neighborhood structure N_i
  - Weighting scheme ϖ_{ij}

Constructing Neighborhoods

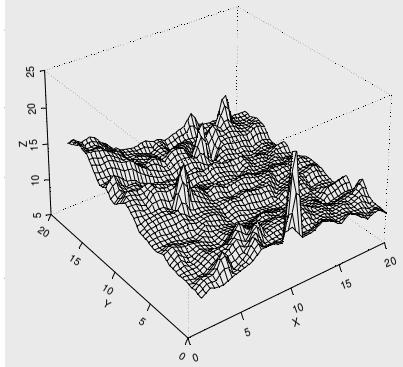
- Depends on whether layout is regular or irregular

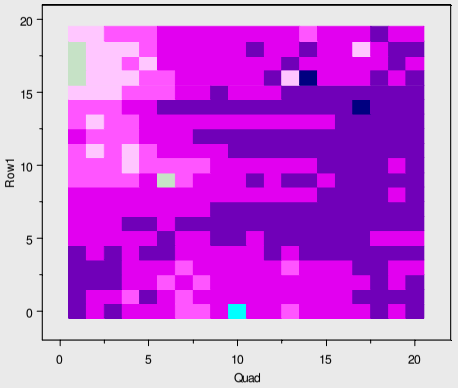




- Every cell (plot, county) must have a defined neighborhood

Example – Bell Pepper Fungus



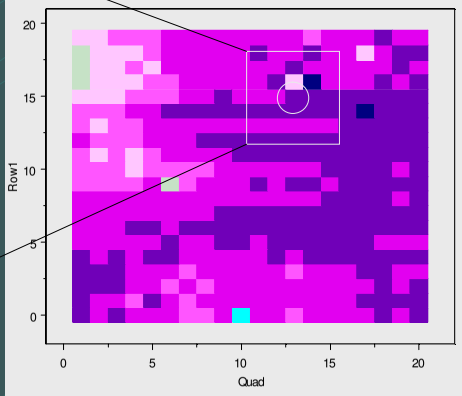


Additional Comments

- The bell pepper fungus data was collected on a regular grid layout with 20 rows (“row1”) and 20 columns (“quad”)–
  - data for each of the 400 cells in the field plot.
  - For example, while soil moisture may in fact vary over a 1x1 m square plot, only a single number is reported for each 1x1 m plot and so represents the value for that plot.
- Showing two graphics here
  - the left one is a perspective plot which shows the variation in soil moisture values
  - The right one shows the same information as color gradations for each cell for which we have data

Examples: Neighborhoods for Square Lattices

10.8	10.9	9.7	10.2	12.0
11.5	9.8	15.0	6.1	10.3
9.7	9.0	9.2	10.4	10.7
8.6	8.8	8.6	8.4	9.5
10.2	11.0	10.3	10.1	10.2



First-order NB

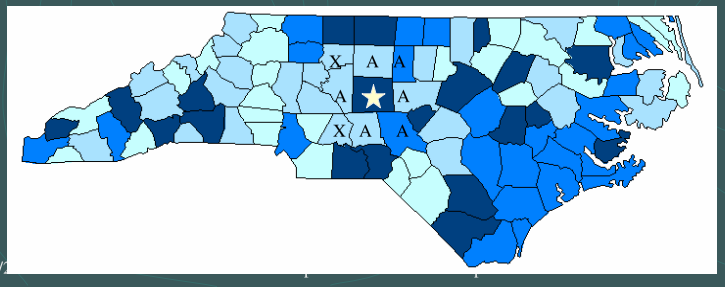
+ Second-order NB

Additional Comments

- These neighborhoods are two of many possible examples – one can further change them or even use different setups.
  - For example, in the case of the water moisture, one might expect that autocorrelation would be higher in the within row direction rather than across rows. This could be due to watering the field by flooding of the pathways between rows or of the beds are raised. In that case, the neighborhood might only be the plots adjacent and within the same row, say the N-S plots only.

Examples: Neighborhoods for Non-Square Lattices

- N_i = *{cells labeled A}* is a neighborhood whose boundaries touch the boundary of the i^{th} cell
- N_i = *{cells labeled A or X}* is a neighborhood whose centroids are within a specified distance from the centroid of the i^{th} cell



Weighting Scheme ϖ_{ij}

- The larger the weight the more that neighboring plot contributes
- Common approaches
  - As a function of Euclidean distance
  - As a function of contiguity
  - Directional weighting (certain directions contribute more than others)
  - As a function of the length of the common boundary
  - Weighting to correct for heterogeneity of variance

Additional Comments

- Weighting can involve some combination of these approaches and is clearly integrally related to the definition of the neighborhood.
- Weights are usually standardized so that they sum to a constant, e.g. $\sum_{j \in N_i} w_{ij} = \eta$
- Negative weights (which imply a negative correlation) are usually avoided but there are times when they are appropriate.

Weighting Scheme ϖ_{ij}

- Crucial to identify appropriate weighting method
- Should have some idea of
  - The range of likely autocorrelation
  - How fast autocorrelation decays as distance increases
  - The direction of likely autocorrelation
    1. The directionality is influenced by both the choice of neighborhood as well as differential weighting by direction.

Weighting Scheme ϖ_{ij}

- Methods for exploring likely form of autocorrelation:
  - Calculate some common autocorrelation statistics such as Moran's I or Geary's C
    - Validity depends on the neighborhood and weighting scheme
    - Try different neighborhoods and weights
  - Do variography using the centroids or nodes of a lattice as the point locations

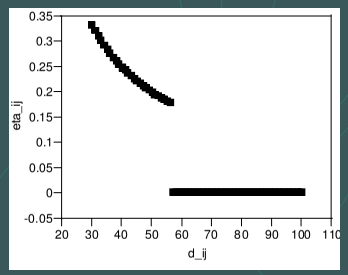
Simple Weighting

- Bell Pepper Fungus
  - Let N_i be the 1x1 m plots having boundaries with the i^{th} plot (first-order NB)
  - Define the weights to be $\varpi_{ij} = \eta \text{ if } j^{th} \text{ plot in } N_i$
  - These weights imply
    - no directionality
    - each neighboring plot is equally autocorrelated with the i^{th} plot
    - The autocorrelation is the same regardless of the location of the i^{th} plot

More Complex Weighting

- Aquatic Cave Species in SE US
  - Defined the neighborhood to be counties with county seats within 56 km of the i^{th} county
  - Uses Euclidean distance to weight closer counties higher than farther counties

$$\eta_{ij} = \begin{cases} 0, & \text{if } d_{ij} > 56\,\text{km} \\ \dfrac{\rho\left\{\dfrac{1}{d_{ij}}\right\}}{\max\limits_{i,j}\left\{\dfrac{1}{d_{ij}} : j \in N_i\right\}}, & \text{otherwise} \end{cases}$$



Additional Comments

- The numerator is a constant times the inverse of the distance between the 2 locations (inverse so that closer neighbors weight higher than further neighbors).
- The denominator is a scaling or standardizing function so that ρ is the correlation between the ith county and its nearest neighbor.
- This approach is a type of "row standardization" and constrains the constant ρ to be less than 1.

Modeling

- So,
  - having identified the explanatory variables for the large-scale variation (trend),
  - the neighborhood structure and weighting scheme for the small-scale variation, and
  - checked for homogeneity of variance,
- the next step is
  - to do the actual model fitting to obtain estimates of the model parameters, means (and SEMs) and, if desired, predictions (and MSPE).

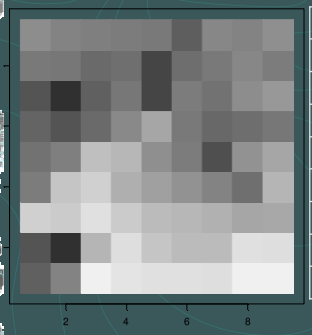
Modeling Approaches

- Two approaches
  - Simultaneous Autoregressive Models (SAR models)
  - Conditional Autoregressive Models (CAR models)
- The difference is in the variance-covariance matrix for the $\{Y(s_1), \ldots, Y(s_n)\}$
- Both can be fitted but fitting the SAR model leads to residuals that are correlated with the neighboring *Y*-values
  - CAR model does not have this problem and is generally preferred

Additional Comments

- Every SAR model can be described in terms of a CAR model but CAR models are not always easily or naturally described as SAR models. This is based on the choices of neighborhoods and variance structure and weights.

Simple Example – Reflectance Values for Pollution in the English Channel



Data Values

32	35	36	37	38	47	34	35	31
38	39	43	41	55	42	38	34	37
50	62	46	39	55	37	40	32	28
45	50	43	33	24	38	44	42	39
40	36	16	18	31	37	52	30	24
37	14	10	21	26	30	35	41	19
10	12	5	12	17	18	20	24	23
50	62	19	6	14	17	17	5	6
46	35	0	4	5	5	6	0	0

from Haining (1990)

Additional Comments

- Like the bell pepper fungus, this dataset is on a regular grid. So, the spatial coordinate system is taken to be
  - the row ID, "r", and
  - the column ID, "c".

Conditional Autoregressive Model

$$Y(s_i) = \mu(s_i) + \sum_{s_j \in N_i} \varpi_{ij}[Y(s_j) - \mu(s_j)] + \varepsilon(s_i)$$

- The error terms are conditionally independent and Normally distributed with mean 0 and constant variance σ^2
- The conditional mean of $Y(s_i)$ is

$$\mu(s_i) + \sum \varpi_{ij}[Y(s_j) - \mu(s_j)]$$

and the unconditional mean is

$$\mu(s_i)$$

Additional Comments

- The conditional mean is the predicted value for an individual observation. The unconditional (marginal) mean is the mean of the trend part only.
- The conditional mean is estimated using the BLUP and the unconditional mean by the BLUE (LSmeans).

Conditional Autoregressive Model

$$Y(s_i) = \mu(s_i) + \sum_{s_j \in N_i} \varpi_{ij}[Y(s_j) - \mu(s_j)] + \varepsilon(s_i)$$

- The conditional variance of $\{Y(s_i) : i = 1,\ldots,n\}$ is

$$\sigma^2 \mathbf{I}$$

and the unconditional variance is

$$(\mathbf{I} - \mathbf{W})^{-1}\sigma^2 \mathbf{I}$$

where $\mathbf{W} = \{w_{ij}\}$ is the matrix version of the weights for the neighborhood

Large-Scale Trend $\mu(s_i)$

$$Y(r,c) = \beta_0 + \beta_1 r + \beta_2 c + \beta_{12} rc + \beta_{11} r^2 + \beta_{22} c^2 + \ldots + \varepsilon(r,c)$$

- Haining (1990) started by ignoring the spatial autocorrelation and fit linear regression models using polynomials in (r, c) where r is the row ID, c is the column ID
- He determined that the linear model had the best fit

$$Y(r,c) = \beta_0 + \beta_1 r + \beta_2 c + \varepsilon(r,c)$$

Small-Scale Variation

- We should now test for autocorrelation in the data, so we'll use the residuals from the large-scale trend fit
- Calculate Moran's I for different neighborhood structures (see next slide) using weight = 1 if grid cell was in the neighborhood and 0 otherwise. From this we can tell
  - If there is autocorrelation
  - Which neighborhood is best (among those reviewed of course)

Small-Scale Variation

Neighborhood	Moran's I	SE	Normal Statistic	Normal p-value	Permutation p-value
Row	0.3215	0.1164	2.869	0.004	0.001
Column	0.5434	0.1164	4.775	0+	0+
Diagonal	0.2043	0.0862	2.514	0.012	0.007
First-order	0.4324	0.0814	5.468	0+	0+
Second-order	0.3251	0.0577	5.848	0+	0+

The highest Moran's I value occurs for the column neighborhood and the second highest for the first-order neighborhood.

Additional Comments

- Under the null hypothesis of no autocorrelation, the expected value of Moran's I is E{I} = -1/(n-1). The stronger the correlation, the closer I is to 1.
- Two approaches for testing autocorrelation using Moran's I are:
- 1) approximate normality holds assuming the number of cells is sufficiently large (also depends on the extent and manner in which the cells are connected by the weights). The usual rule of thumb is at least 20 locations.
- 2) permutation or randomization test in which the Z data are randomly permuted (assigned to different locations) repeatedly and the observed results compared against the expected results.

Fit the CAR model

Model was fit with a linear trend and with weights $w_{ij}=\rho$ if plot j was in the neighborhood and = 0 otherwise.

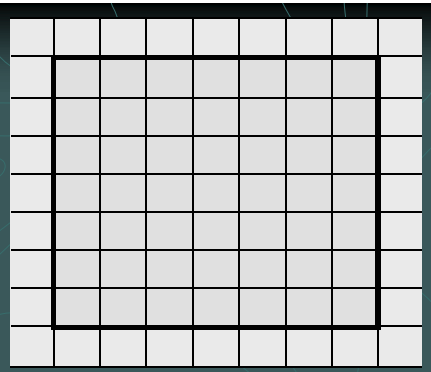
Model	β_0 Estimate	β_1 Estimate	β_2 Estimate	ρ Estimate	Root MSE	Log Likel.
1	50.966**	-2.733**	-2.131*	0.256**	9.02	-362
2	53.755**	-2.966**	-1.757**	0.442**	8.85	-364
1a	60.289**	-3.467**	-2.050*	0.251**	10.00	-211

Models: (1) first-order neighborhood with 9x9 area
(2) column neighborhood with 9x9 area
(1a) first-order neighborhood with 8x8 interior area

Additional Comments

- There is very little difference in the models with the two different neighborhood structures, so for parsimony choose the model using the column neighborhood
- Note that the estimated spatial weight is 0.256 for the first-order neighborhood and 0.44 for the column neighborhood. The difference in values has more to do with the number of neighbors in the neighborhood than with any estimate of autocorrelation.
- The final model is the result of adjusting for boundary effects (next).

Boundary Effects



- The neighborhoods of the cells on the edges are halved
  - Standard errors of predictions at the edges very high
  - Introduces possible estimation bias
- One way to avoid is to analyze only that part of the study region completely within the entire region

Additional Comments

- Choose the subregion within the study area so that every cell in the subregion has a complete neighborhood that can be used in the modeling
- In the bell pepper example, that would be a subregion 19x19 (rather than 20x20) that would be modeled (Y-values on the left side of the model). The remaining cells would appear only on the right side of the model in the small-scale variation.

Fit the CAR model

Model was fit with a linear trend and with weights $w_{ij}=\rho$ if plot j was in the neighborhood and = 0 otherwise.

Model	β_0 Estimate	β_1 Estimate	β_2 Estimate	ρ Estimate	Root MSE	Log Likel.
1	50.966**	-2.733**	-2.131*	0.256**	9.02	-362
2	53.755**	-2.966**	-1.757**	0.442**	8.85	-364
1a	60.289**	-3.467**	-2.050*	0.251**	10.00	-211

Models: (1) first-order neighborhood with 9x9 area
(2) column neighborhood with 9x9 area
(1a) first-order neighborhood with 8x8 interior area

Additional Comments

- Note the difference between model 1 and model 1a in the estimates of the model coefficients. Due to
  - smaller number of observations (64 vs. 81)
  - Better estimation of the spatial autocorrelation since every observations has a full neighborhood

Summary and Conclusions

- When data are collected in aggregate for non-overlapping subregions of the study area and
- The spatial arrangement is such that there are effects due to space (or to spatial covariates that were not measured)
- Then consider models that incorporate an effect due to spatial correlation

Advantages

- Accounts for some additional sources of variation
- Increases understanding of the process of interest
- Overall lattice models are excellent approaches for incorporating spatial correlation and for providing improved predictions

Caveats When Fitting Lattice Models

- If covariates are available that explain the seeming spatial correlation, then these are more appropriately used
- Choice of neighborhood and weighting scheme are critical to good model fitting
- Sample sizes could be too small to adequately estimate the spatial correlation
- Modeling might require a lot of exploratory analyses. Note that this means that the conclusions are only tentative and should be independently tested with a new experiment.

Colorado potato beetle infestation in plots on a lattice design

Matt Kramer and Don Weber

kramerm@ba.ars.usda.gov, weberd@ba.ars.usda.gov

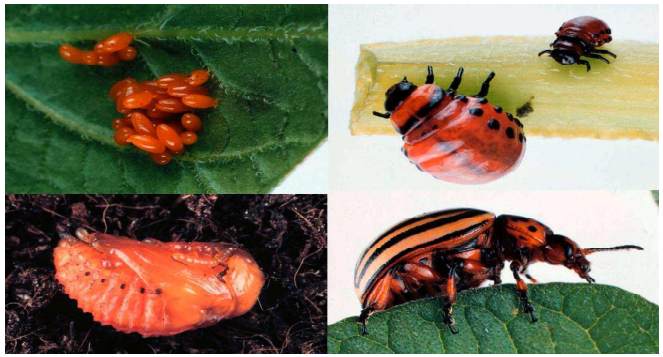
Biometrics (MK) and Insect Biocontrol (DW), ARS/BARC/USDA

Outline

- Introduction to Colorado potato beetles
- Experimental design and plots
- *R* software—*spdep* package
- Spatial weights
- Models and analyses
  - Ignore spatial dependencies
  - SAR
  - CAR
- Diagnostics
- Conclusion

Introduction to Colorado potato beetles

There are 4 life stages: egg, larva, pupa, adult. Larvae and adults feed on leaves.



photographs by Doro Röthlisberger, Zoological Museum, University of Zurich

Introduction to Colorado potato beetles

- Colorado potato beetles (*Leptinotarsa decemlineata*) overwinter as adults and can pass through 2–3 generations in Maryland.
- The data were taken (mid May) when all life stages were present
- CPB is a pest in North America (where it is native) but has also been introduced into Europe, which now suffers damage from it comparable to that in North America.
- CPB attacks plants in the nightshade family (potatoes, eggplants, tomatoes, and their wild relatives).
- Colorado potato beetles have developed resistance to a long succession of different insecticides, and its natural enemies do not reliably control it in current farming practices.
- New practices, in combination with natural enemies, show promise to maintain CPB populations below economic thresholds, reducing the need for pesticide applications.

Introduction to Colorado potato beetles

- In this experiment, tillage practice, planting date, and mulch cover were manipulated.
- We chose these data for a lattice example because the plots are laid out on a lattice, and it is a reasonably small data set. At the onset, we knew there were treatment effects but did not know if there were spatial dependencies.
- The goal of the project is to determine which combination of treatments best reduces CPB infestation
- In addition to treatment effects, we thought there might be block and border effects (and spatial correlation among neigboring plots)
- Sampling occurred in the interior of the plots
- Spatial correlation was suspected because adults and larva are mobile, both walk and adults can fly

Cooperators and administrators

Left to right: Matt Greenstone, Phyllis Johnson, Don Weber, Ron Korcak, John Teasdale, Aref Abdul-Baki, Vinod Kumar

Field team

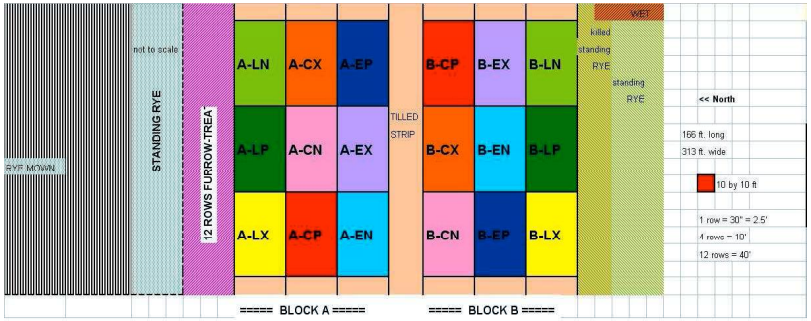
Left to right: Jenn Curtis, Jon Curtis, Eddie Bender, Michael Donovan, Mike Athanas, Greg Benedict

Experimental design and plots

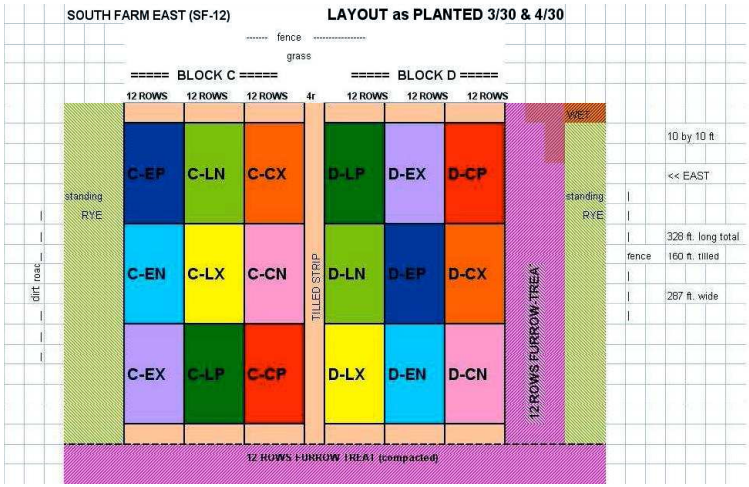
- Treatments were cultivation (whole plot effect)
  - E = early planting, no till
  - L = late planting, no till
  - C = late planting, till

  and amount of mulch used (split plot effect)
  - N = rye cover crop only, none added
  - P = rye cover crop + 1x mulch (straw from rye cover crop)
  - X = rye cover crop + 2x mulch
- The measure of infestation is CPB equivalents per plant stalk = number of adults + $\frac{2}{3}$ of the number of large larvae + $\frac{1}{4}$ of the number of small larvae, averaged over 20 plants per plot
- Split plot design (though not analyzed that way here)
- Four blocks (in two spatially distant sets), nine treatment combinations per plot, so 36 total observations

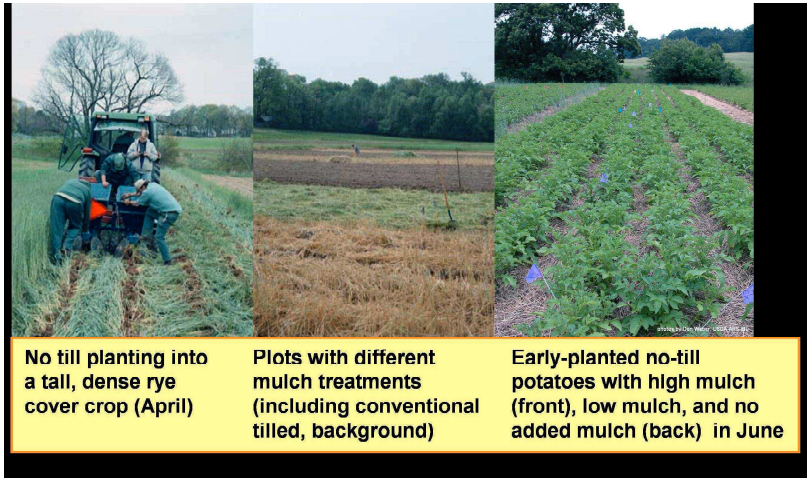
Experimental design and plots



Experimental design and plots



Experimental design and plots

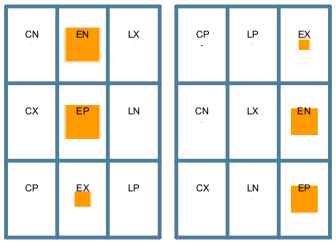


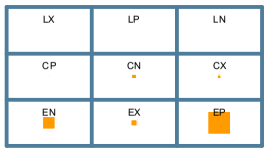
Data collected for Plot A

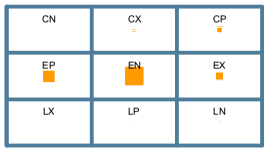
trt	X	Y	till	plant	CPB	borders	mulch
LN	130	185	no	late	0.12	N, E	none
LP	80	185	no	late	0.00	N	+1x mulch
LX	30	185	no	late	0.02	N, W	+2x mulch
CX	130	155	yes	late	0.33	E	+2x mulch
CN	80	155	yes	late	0.32	-	none
CP	30	155	yes	late	0.19	W	+1x mulch
EP	130	125	no	early	4.10	E, block B	+1x mulch
EX	80	125	no	early	0.67	block B	+2x mulch
EN	30	125	no	early	1.28	W, block B	none

CPB equivalent incidence on plots

South Farm Plots







East Line Road Plots

R *software—*spdep *package*

- *R* software (`http://www.R-project.org`) was used for the analysis
- *spdep* package (main author: Roger Bivand) which has functions for creating spatial weights, tests for spatial autocorrelation (e.g. Moran's I), estimating spatial simultaneous autoregressive (SAR) lag and error models, conditional autoregressive (CAR) models (in a preliminary stage), and includes routines for using sparse matrices
- Installation (on Linux and Windows) of the *spdep* package requires some other *R* packages. For Linux, some of these require compiling C and Fortran code.

Spatial weights

- There are several important decisions to make, e.g. what is a neighbor and how should neighbors be weighted
- *Spdep* can be given the xy coordinates of the middle of each plot and then use a distance cutoff to determine neighbors (weight of 1 for neighbor, 0 if not a neighbor)
- This was tried for various distance cutoffs, and spatial dependence was smaller with a bigger cutoff (bigger neighborhood)
- One can also input a matrix of spatial weights, which could depend on characteristics not directly related to distance (e.g. if plots share a common border). This could be binary (1 if a neighbor, 0 if not a neighbor) or scaled to represent the relationship between neighbors (e.g., length of common border)

Spatial weights

- We used spatial weights that depended on the length of the common border, scaled so the sum of the weights = 36.
- For the same set of residuals, this weighting scheme produced higher estimated spatial dependencies (i.e. seemed to capture more of the spatial correlation)
- Another alternative is to try geostatistical models (e.g. exponential decay, spherical, etc.), these would be based on the distances between the centers of the plots.

Spatial weights for lengths of common border for Block A (not scaled)

plot ID	LN	LP	LX	CX	CN	CP	EP	EX	EN
LN	0	3	0	5	0	0	0	0	0
LP	3	0	3	0	5	0	0	0	0
LX	0	3	0	0	0	5	0	0	0
CX	5	0	0	0	3	0	5	0	0
CN	0	5	0	3	0	3	0	5	0
CP	0	0	5	0	3	0	0	0	5
EP	0	0	0	5	0	0	0	3	0
EX	0	0	0	0	5	0	3	0	3
EN	0	0	0	0	0	5	0	3	0

More on spatial weights

- spatial weights can be symmetric (as in the last example) or asymmetric
- asymmetric weights occur when the spatial weight of the effect of A on B differs from that of B on A. This is reasonable in many circumstances, e.g.,
  - prevailing wind is mostly from one direction
  - the number of neighbors of A is less than that of B, and since B is influenced by many neighbors, the effect of A on B is diluted
- row standardization (i.e. for each observation, the sum of the weights of the neighbors is one) is often suggested, this will lead to asymmetric weights (weights of neighbors will be larger if an observation has fewer neighbors).

Ignore spatial dependencies

- An analysis to determine which treatment, block, and border effects to include in fixed part of model using stepwise regression (based on minimizing AIC)—this is because there were a large number of candidate regressors and only 36 observations to support their estimation.
- Effects were coded as zero-one dummy variables, including some interaction effects
- Since the data were based on counts, a square root transformation was performed. Diagnostics also suggested that this transformation was better than a log or no transformation
- Model: $\sqrt{\mathbf{y}} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where
  - $\sqrt{\mathbf{y}} =$ square root of Colorado pototo beetle equivalents
  - $\boldsymbol{X\beta} =$ fixed effects
  - $\boldsymbol{\epsilon} =$ uncorrelated random error (noise)

Another look at the data

Data for Block A—border effects (b1–b4) differ depending on block, m1 and m2 represent mulch levels, format for stepwise regression

trt	X	Y	till	pl	CPB	b1	b2	b3	b4	m1	m2
LN	130	185	0	0	0.12	1	0	1	0	0	0
LP	80	185	0	0	0.00	1	0	0	0	1	0
LX	30	185	0	0	0.02	1	0	0	1	0	1
CX	130	155	1	0	0.33	0	0	1	0	0	1
CN	80	155	1	0	0.32	0	0	0	0	0	0
CP	30	155	1	0	0.19	0	0	0	1	1	0
EP	130	125	0	1	4.10	0	1	1	0	1	0
EX	80	125	0	1	0.67	0	1	0	0	0	1
EN	30	125	0	1	1.28	0	1	0	1	0	0

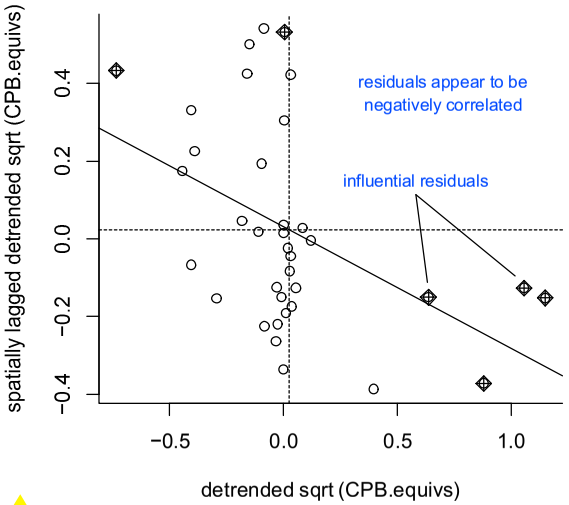
Model from Stepwise regression

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.1704     0.0830   2.053 0.051146 .
TILLAGE               0.6132     0.1114   5.507 1.16e-05 ***
PLANT.TIME.LATE       1.4389     0.1114  12.921 2.67e-12 ***
mulch2                0.1054     0.1301   0.810 0.426118
b.AB.west            -0.3111     0.1111  -2.800 0.009925 **
block.D               0.5530     0.1150   4.810 6.73e-05 ***
b.D.west             -0.8014     0.1870  -4.286 0.000255 ***
b.CD                 -0.4682     0.1150  -4.072 0.000439 ***
b.C.east              0.3205     0.1428   2.245 0.034268 *
b.AB.east             0.1351     0.1004   1.345 0.191332
PLANT.TIME.LATE:mulch2 -0.7913   0.1860  -4.255 0.000276 ***
TILLAGE:mulch2        -0.3239     0.1860  -1.741 0.094438 .

Residual standard error: 0.1938 on 24 degrees of freedom
Multiple R-Squared: 0.9438,     Adjusted R-squared: 0.9181
F-statistic: 36.66 on 11 and 24 DF,  p-value: 2.656e-12
```


Predicted (green) vs. Data (orange)

Moran's I on detrended observations



Which spatial dependency model (for SAR models)?

```
Lagrange multiplier diagnostics for spatial dependence

LMerr =  3.9604, df = 1, p-value = 0.04658
RLMerr = 0.6708, df = 1, p-value = 0.4128
LMlag =  5.7218, df = 1, p-value = 0.01676
RLMlag = 2.4321, df = 1, p-value = 0.1189
SARMA =  6.3925, df = 2, p-value = 0.04092
```

Suggests the lag model might be better than the error model

Simultaneous autoregressive spatial models

There are two basic models fit by *spdep*

- spatial simultaneous autoregressive *error* models

$$y = X\beta + u, u = \lambda W u + \epsilon$$

where

- $y =$ square root of Colorado pototo beetle equivalents
- $X\beta =$ fixed effects
- $u =$ correlated errors with two components
- $\lambda =$ autoregressive error parameter
- $Wu =$ weighted vector of neighboring *residuals* (describes which *residuals* of the neighbors the *residual* of the observation is correlated with and how they are weighted)
- $\epsilon =$ uncorrelated random error (noise)

Simultaneous autoregressive spatial models

- spatial simultaneous autoregressive *lag* models

$$y = \rho W y + X\beta + \epsilon$$

where (for the new terms)

- $\rho =$ autoregressive lag parameter
- $Wy =$ weighted vector of neighbors (describes which neighbors the observation is correlated with and how they are weighted)

Comparison of fixed effects estimates

effect	linear model (SE)	error (SE)	lag (SE)
(Intercept)	0.17 (0.08)	0.18 (0.06)	0.36 (0.10)
TILLAGE	0.61 (0.11)	0.55 (0.08)	0.61 (0.08)
PLANT.LATE	1.44 (0.11)	1.43 (0.08)	1.45 (0.08)
mulch2	0.11 (0.13)	0.08 (0.10)	0.13 (0.10)
b.AB.west	-0.31 (0.11)	-0.31 (0.07)	-0.38 (0.09)
block.D	0.55 (0.12)	0.52 (0.07)	0.65 (0.10)
b.D.west	-0.80 (0.19)	-0.71 (0.13)	-0.85 (0.14)
b.CD	-0.47 (0.12)	-0.41 (0.08)	-0.58 (0.10)
b.C.east	0.32 (0.14)	0.33 (0.09)	0.30 (0.11)
b.AB.east	0.14 (0.10)	0.10 (0.06)	0.09 (0.08)
PLANT.LATE:mulch2	-0.79 (0.19)	-0.73 (0.13)	-0.79 (0.14)
TILLAGE:mulch2	-0.32 (0.19)	-0.20 (0.14)	-0.27 (0.14)

Errorsarlm vs. Lagsarlm

Error model:

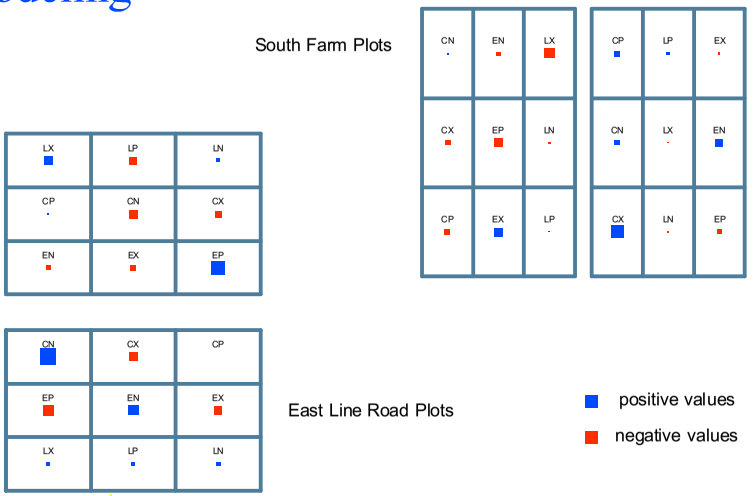
```
Lambda: -0.46916 LR test value: 5.7618 p-value: 0.016379
Asymptotic standard error: 0.14247 z-value: -3.293 p-value: 0.00099118
Log likelihood: 18.17689 for error model
ML residual variance (sigma squared): 0.019354, (sigma: 0.13912)
Number of parameters estimated: 14
AIC: -8.3538, (AIC for lm: -4.592)
```

Lag model:

```
Rho: -0.24002 LR test value: 6.1499 p-value: 0.013142
Asymptotic standard error: 0.091073 z-value: -2.6355 p-value: 0.008401
Log likelihood: 18.37094 for lag model
ML residual variance (sigma squared): 0.020609, (sigma: 0.14356)
Number of parameters estimated: 14
AIC: -8.7419, (AIC for lm: -4.592)
```

Lag: predicted (green), data (orange)

Data minus fixed effects: What $\rho W y$ is modeling



Conditional Spatial Autoregressive Model

In CAR models, an observation's value is conditioned on neighboring values. This is one representation for the model:

$$E(y_i|\mathbf{y}_{*i}) = \mathbf{X}\beta + \lambda\mathbf{W}(\mathbf{y}_{*i} - \mu_{*i})$$

where

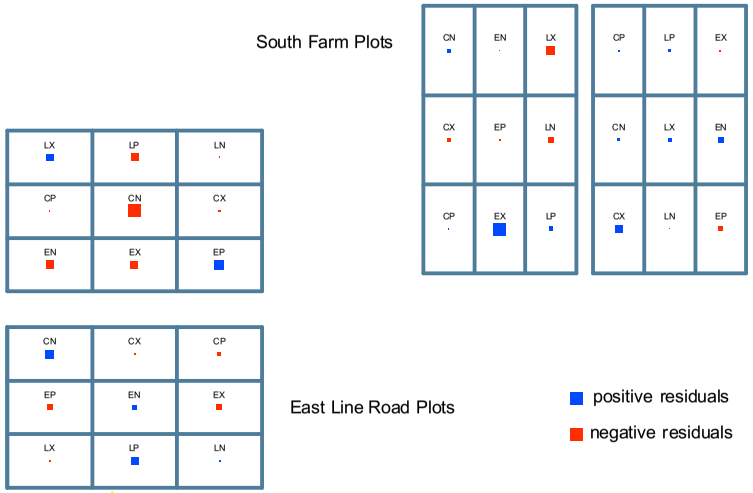
- y_i = square root of Colorado pototo beetle equivalents
- $\mathbf{X}\beta$ = fixed effects for y_i
- $\mathbf{y}_{*i}$ = neighbors of y_i ($*i$ = not including observation i)
- λ = autoregressive parameter
- $\mathbf{W}(\mathbf{y}_{*i} - \mu_{*i})$ = weighted vector of mean adjusted neighbors

Estimation results from CAR model

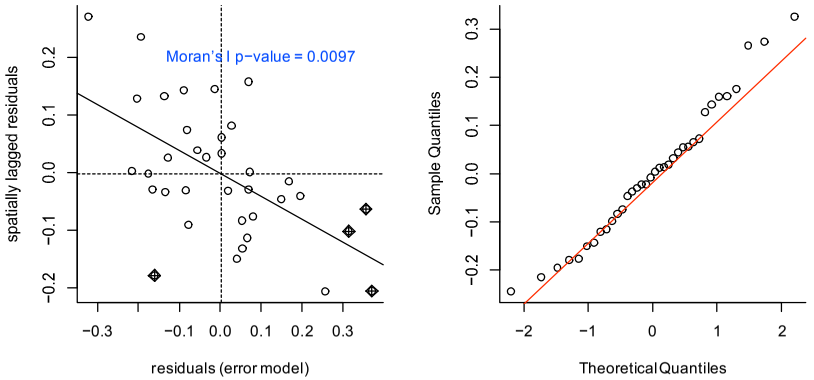
```
                           Estimate Std. Error z value  Pr(>|z|)
(Intercept)                0.176337   0.056978  3.0948 0.0019694 **
TILLAGE                    0.560195   0.080908  6.9238 4.395e-12 ***
PLANT.TIME.LATE            1.430348   0.079894 17.9030 < 2.2e-16 ***
mulch2                     0.079565   0.100107  0.7948 0.4267288
b.AB.west                 -0.309661   0.071309 -4.3426 1.408e-05 ***
block.D                    0.523969   0.071477  7.3306 2.292e-13 ***
b.D.west                  -0.728253   0.133909 -5.4384 5.375e-08 ***
b.CD                      -0.427340   0.081758 -5.2269 1.724e-07 ***
b.C.east                   0.326350   0.095643  3.4122 0.0006445 ***
b.AB.east                  0.106542   0.064655  1.6479 0.0993820 .
PLANT.TIME.LATE:mulch2    -0.732866   0.137060 -5.3470 8.941e-08 ***
TILLAGE:mulch2            -0.210189   0.142252 -1.4776 0.1395201

Lambda: -0.70764 LR test value: 5.3146 p-value: 0.021147
Log likelihood: 17.95331    AIC: -7.9066
ML residual variance (sigma squared): 0.018909, (sigma: 0.13751)
```

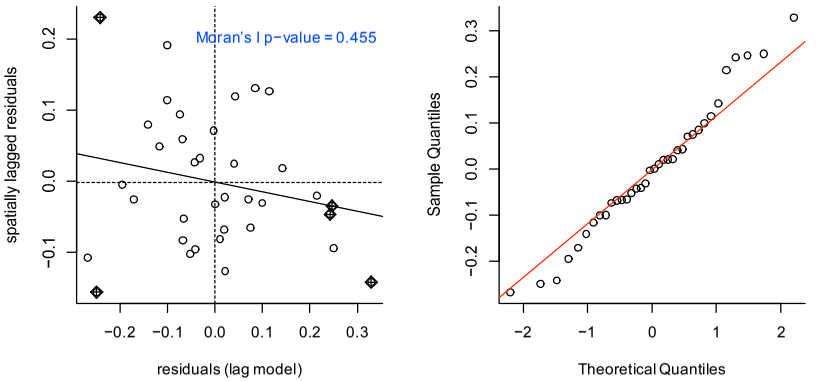
Residuals of CAR model



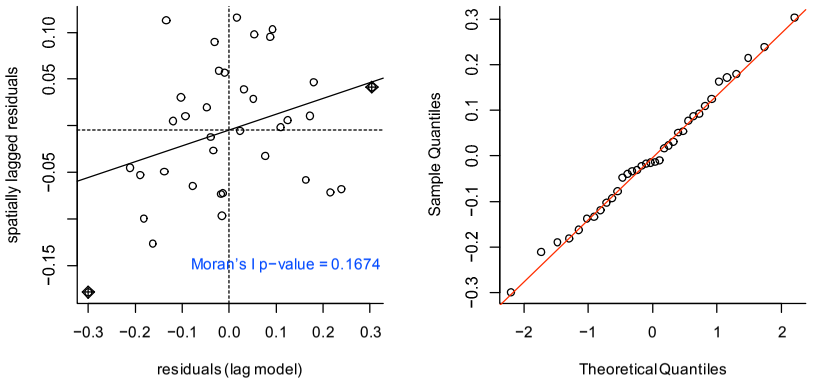
Diagnostics, SAR error model: Moran and QQnorm plots



Diagnostics, SAR lag model: Moran and QQnorm plots



Diagnostics, CAR model: Moran and QQnorm plots



Conclusions I.

- Estimates of fixed effects parameters were similar for all models
- Standard errors of fixed effects parameters were smaller when spatial dependencies were taken into account
- For these data, judging by AIC, the spatial dependencies appeared to be captured adequately by all spatial models discussed, and there is a substantial improvement over the model that ignores spatial dependencies
- The CAR model seems to have better behaved residuals

Conclusions II.

- Why a negative correlation between neighboring plots? Our best guess is that the beetle population is locally redistributing to favorable plots after departing unfavorable ones. So, the relative accumulation of beetle numbers on a particular treatment combination depends on which neighbors it has.
- In field season 2006 we will be looking at individual beetle behavior including arrival and residence time in different treatments, which should yield insight into this spatial pattern.

Spatial Sampling Design and Strategies

Jun Zhu

Department of Statistics and Department of Soil Science
University of Wisconsin - Madison
Madison, Wisconsin

Outline

Outline

Spatial sampling design

- Example: a study of old-growth northern hardwood forests (Miller et al., 2002).
  - Consideration of biodiversity in natural resource management.
  - Spatial patterns of forest understory vegetation (herbs, shrubs, tree seedlings, saplings).
  - Different species exhibit different spatial patterns within a given environment?
  - Biotic and abiotic factors in the environment are related to a species' spatial pattern?
- An important question: *where* should data be collected?
- The purpose is to design a sampling scheme that ensures scientific objectivity.

Spatial sampling design

- Suppose the study area of interest is D.
- Suppose measurements of Z will be taken at locations $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n$ in D, where $\boldsymbol{s} = (x, y)$ and n is the sample size. Where should they be?
- It depends!
- Possible objectives
  - Estimation of mean (e.g. average soil P in a field)
  - Estimation of variogram (e.g. map of soil P in the field)
  - Comparison of treatments (e.g. effect of a new fertilizer)
- Possible prior information
  - Accessible study area and sampling locations
  - Affordable sample size
  - Condition of a study area

Related subjects

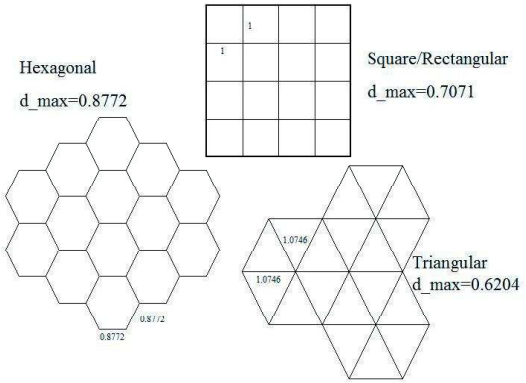
- Survey sampling: design-based sampling versus model-based sampling (Gruijter and Braak, 1990; Särndal et al., 1992)
- Design of experiment and optimal design (Mead et al., 1993)
- Spatial sampling design and optimal sampling (Webster and Oliver, 2001)
- An excellent review article: Stein and Christien (2003)

Outline

Regular grids

- Triangular or isometric grid: tiling plane regularly with equilateral triangles.
- Rectangular grid: tiling plane regularly with squares.
- Hexagonal grid: tiling plane regularly with hexagons.

Regular grids



$d_{\max}$ = maximum distance from any point in D to the nearest grid point.

A plausible scenario

- The goal is to estimate the overall mean

$$\mu = E(Z).$$

- Assume a regular spatial sampling grid with a fixed sampling density.
- Assume an exponential semivariogram for the spatial correlation structure.

And the winner is...

- A triangular grid is the most efficient design with the smallest $d_{\max}$.
- That is, for the same sampling intensity, it places the sampling locations as far apart as possible while minimizing the area that is under-represented.
- A triangular grid is most efficient for most bounded variograms that have finite ranges.

Remarks

- Under some other assumptions, a hexagonal grid is the most efficient design.
- For convenience, a rectangular grid is often the preferred design in field work.
- Major drawbacks of a regular grid include poor variogram estimates at short distances and the potential problems of systematic design (as versus randomized design).

Outline

Main idea

- To compensate for poor variogram estimates using regular grid designs, an improved method was proposed by Clayton and Hudelson (1995).
- The main idea is to use a regular grid system, but sample at unequal spacings.
- In one dimension (1D), the design allows the estimation of variogram at all multiples of the smallest lag with a minimum number of sampling locations.

1D transect

- Let × = sample; ○ = skip (sampling).
- A 3/7 cyclic sampling design with 2 repeats looks like:

× × ○ × ○ ○ ○ × × ○ × ○ ○ ○

with lag distances

×	1	-	3	-	-	-	7	-	-	-	-	-	-
-	×	-	2	-	-	-	6	7	-	-	-	-	-
-	-	-	×	-	-	-	4	5	-	7	-	-	-

1D transect

- Choice of specific sampling pattern is important.
- Why not

× × × ○ ○ ○ ○ × × × ○ ○ ○ ○

with lag distances

×	1	2	-	-	-	-	7	-	-	-	-	-	-
-	×	1	-	-	-	-	6	7	-	-	-	-	-
-	-	×	-	-	-	-	5	6	7	-	-	-	-

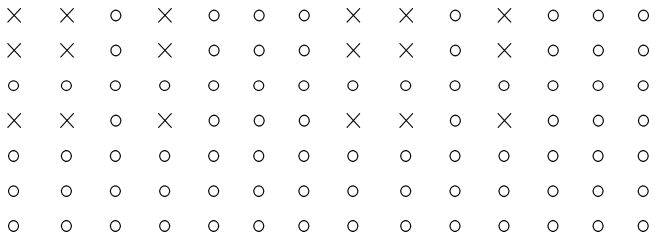
- Lag distances 3 and 4 are missed.

Remarks

- For each lag distance, the proposed 3/7 design gives enough data for making the confidence intervals of the variogram small.
- There are more 7-lag distances than others in a 3/7 design, which cannot be avoided.
- Other possible cyclic sampling designs are: 4/13, 5/21, 6/31, etc. (Clinger and Ness, 1976).

2D region

- Extension to a 2D region is straightforward, but only approximately optimal.
- A 3/7 design for both the x-axis and y-axis:



- One can have different cyclic sampling designs for rows and columns.
- See Miller et al. (2002) for more details of the understory vegetation example.

Sampling design in practice

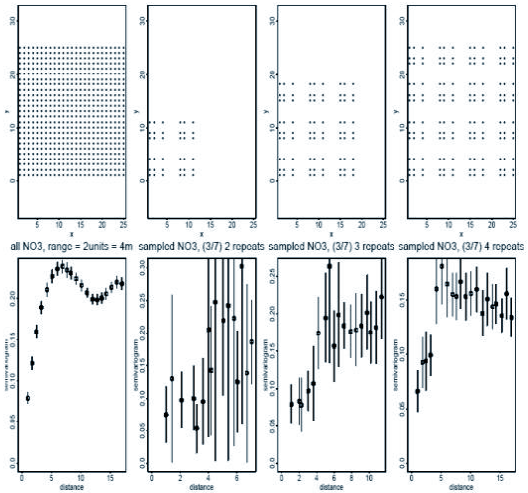
In practice, how to choose a particular cyclic sampling design and hence the sample size?

1. Conduct a pilot study to obtain a rough estimate of the range, sill, and nugget.
2. Simulate data on a grid with the finest grain scale possible for sampling, based on the estimated range, sill, and nugget.
3. Sample from the simulated data according to different sampling designs.
4. For each sample, compute the fitted range, sill, and nugget, and the confidence intervals of the variogram.
5. Evaluate the effect of different designs on the confidence interval width.
6. Consult a statistician!

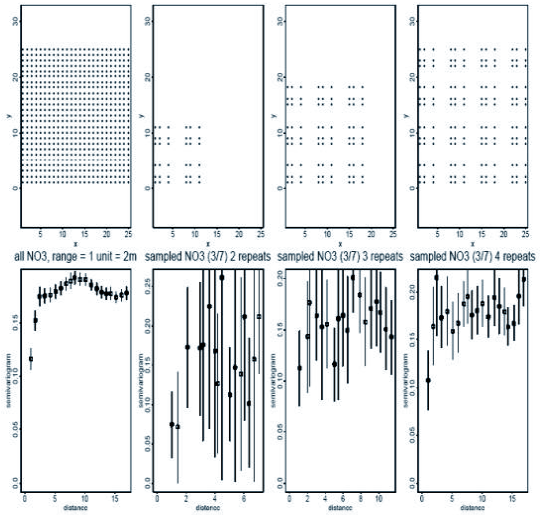
Example: Nitrogen cycling

- Assume exponential variogram model with
  - $r = 2$ (i.e. 95% effective $r = 6$).
  - $r = 1$ (i.e. 95% effective $r = 3$).
- Assume a 25×25 grid structure at a 2-m increment.
- Compare the use of 2D 3/7 cyclic sampling design with 1, 2, or 3 repeats.

$r = 2$



$r = 1$



Outline

Main idea

- In many field experiments, blocking is used to account for experimental unit (EU) heterogeneity, assuming that EUs within block homogeneous.
- Often there is spatial correlation within a block.
- If the goal is to have equal precision for the tests of treatment differences, it would make sense to design the experiment so that all treatments are equally near each other.

Example

Block	Treatment			
1	C	A	B	D
2	B	D	C	A
3	A	B	D	C
4	D	B	C	A

	Distance between plots				
Contrast	Block 1	Block 2	Block 3	Block 4	Average
A vs B	1	3	1	2	1.75
A vs C	1	1	3	1	1.50
A vs D	2	2	2	3	2.25
B vs C	2	2	2	1	1.75
B vs D	1	1	1	1	1.00
C vs D	3	1	1	2	1.75
Average					1.67

Average distance balanced design

- Not a balanced design since some treatments are on average closer than others.
- Simple switch in block 4 to DACB would result in much closer average distances.
- A strategy may be to strive for an average distance balanced design.

Nearest neighbor approach

- Instead of distance, look at neighbors of each treatment:
  A vs B: 2 A vs C: 3 A vs D: 0
  B vs C: 1 B vs D: 4 C vs D: 2
- Similar problem as before. While switching block 4 would help, we can do better.
- There are 12 neighbor pairs and 6 trt pairs:

Block	Treatment			
1	C	A	B	D
2	B	D	A	C
3	A	D	C	B
4	D	C	B	A

- Arrangement above is balanced for nearest neighbors and distance.
- Often correlation in both directions (2D). Similar approaches apply.

References

M. K. Clayton and B. D. Hudelson. Confidence intervals for autocorrelations based on cyclic samples. *Journal of the American Statistical Association*, 90:753–757, 1995.

W. Clinger and J. W. Van Ness. On unequally spaced time points in time series. *Annals of Statistics*, 4:736–745, 1976.

J. J. De Gruijter and C. J. F. Ter Braak. Model free estimation from spatial samples: a reappraisal of classical sampling theory. *Mathematical Geology*, 22:407–415, 1990.

R. Mead, R. N. Curnow, and A. M. Hasted. *Statistical Methods in Agriculture and Experimental Biology, 2nd Edition*. Chapman & Hall, London, 1993.

T. F. Miller, D. J. Mladenoff, and M. K. Clayton. Old-growth northern hardwood forests: Spatial autocorrelation and patterns of understory vegetation. *Ecological Monographs*, 72:487–503, 2002.

C. E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer, New York, 1992.

A. Stein and E. Christien. An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons. *Agriculture, Ecosystems and Environment*, 94:31–47, 2003.

R. Webster and M. A. Oliver. *Geostatistics for Environmental Scientists*. Wiley, West Sussex, 2001.

I would like to thank. . .

- Bruce Craig, Purdue University
- Larry Douglass, University of Maryland - College Park
- Murray Clayton, University of Wisconsin - Madison
- Monica Turner, University of Wisconsin - Madison

GIS Basics

SPATIAL STATISTICS WORKSHOP

March 15 – 16, 2006

Presenter: D. Alan Davenport, GIS Coordinator
Division of Migratory Bird Management
U. S. Fish and Wildlife Service, Laurel, MD

Acknowledgement:

Some of the slides used in this presentation were adapted from the course **TEC7112 – *GIS Introduction for Conservation Professionals*** taught at the National Conservation Training Center, Sheperdstown, WV.

What is a Geographic Information System?

A GIS is

- A computer-based system designed for the collection, storage, and analysis of phenomena where geographic (spatial) location is an important characteristic or critical to the analysis.

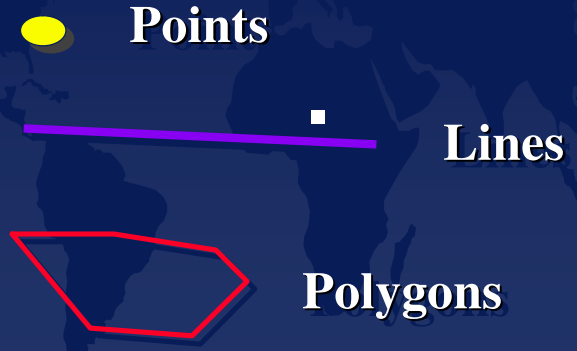
Components

- Spatial Data
- Attributes

Spatial Data

- Landscape elements that have physical dimensions and geographical location. These elements can be represented in two different ways:
- Vectors
- Rasters

Vector Data



Vector Data

Points

- locations of buildings
- wood duck boxes
- water control structures

Vector Data

Lines

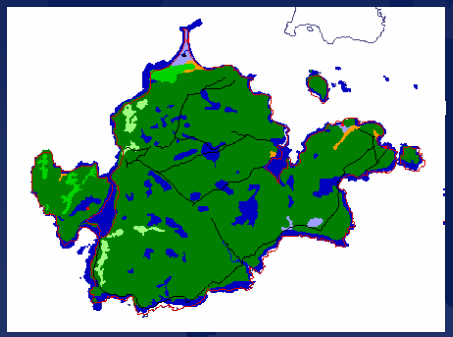
- roads
- boundaries
- streams
- power lines

Vector Data

Polygons

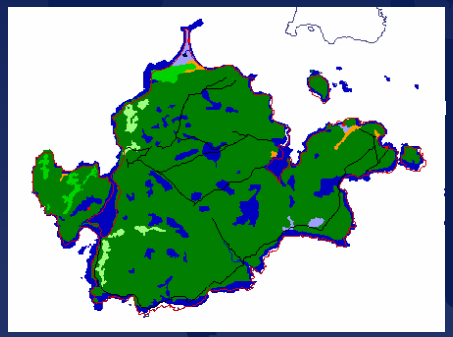
- lakes
- cover types
- timber stands

Vector Data



Overview

Vector Data



Overview

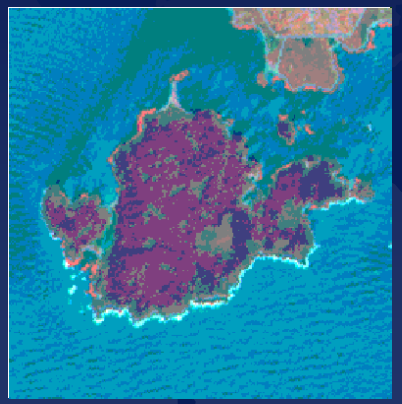
Close-up

Raster Data

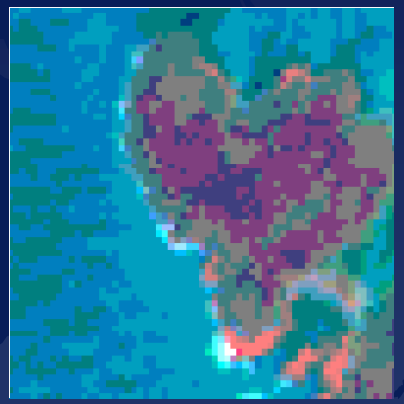
Cells or Pixels

- Landscape elements represented as rows and columns of continuous cells
- Each cell has a location
- Each cell has a value or attribute

Raster Data



Overview



Close-up

Raster Data

- Considerations:

Each cell is a rectangle or square of a constant size. The size of the cells determines the resolution of the map. As the cell size decreases the map resolution increases, but so does the storage requirement in the computer.

Raster Data

Overview

Attributes

- The number of eggs in wood duck box number 27.
- The level of water at Lake Sepik on 27 June 1994.
- The name of a road.
- The volume of red oak saw logs in timber stand number 4.
- The number of black duck broods in Hayes Flowage in 1994.

How do we put it all together?

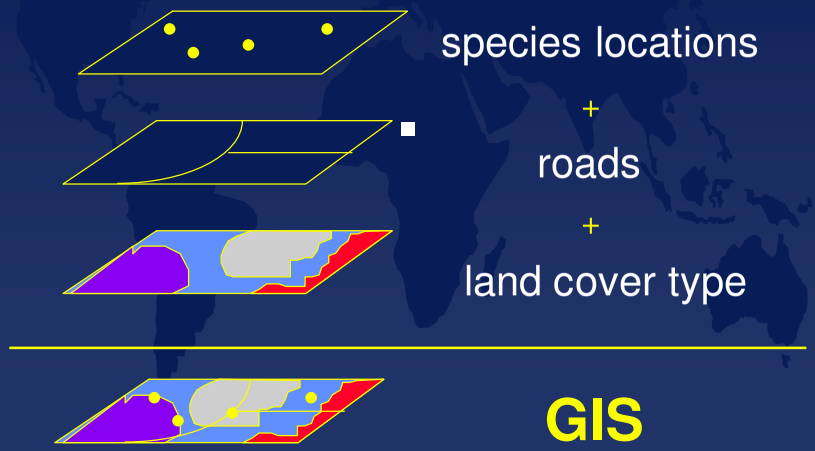
To use spatial data in a GIS you need to know:

- Where each feature is located ***(Coordinates)***
  Geographical Coordinates, X and Y
- What each feature represents ***(Attributes)***
  Can be any number of descriptive characteristics, but there must be at least one.
- Relationships among features ***(Topology)***
  The logic that connects the features to each other, for example, how the location of a wood duck box relates to the location of the nearest wetland. Topology is internally managed by the GIS software.

Spatial data and its attributes must be arranged in a logical order to create a GIS.

This arrangement is a series of layers, or ***THEMES***, each which share a common coordinate system.

A GIS consists of *Data Layers or Themes*



The ultimate purpose of a GIS is to answer spatial questions...

...NOT necessarily to make 'PRETTY' maps!

Typical questions include:

- What is at .....?
- Where is .....?
- What has changed since .....?
- What spatial patterns exist?
- What if .....?

An important thing to remember...

The questions must be asked before the data are developed.

GIS Software

http://esri.com/

http://imgs.intergraph.com/

http://www.mapinfo.com/location/integration

http://www.genaware.com/products/genamap/

Who is ESRI ?

•Environmental Systems Research Institute, Redlands, CA

http://www.esri.com/index.html

ArcGis

What is ArcGIS?

- *An integrated collection of GIS software products* for building a complete GIS. The ArcGIS framework enables you to deploy GIS functionality—in desktops, servers (including the Web), or mobile devices

Why ArcGIS?

- The defacto GIS software standard within the FWS

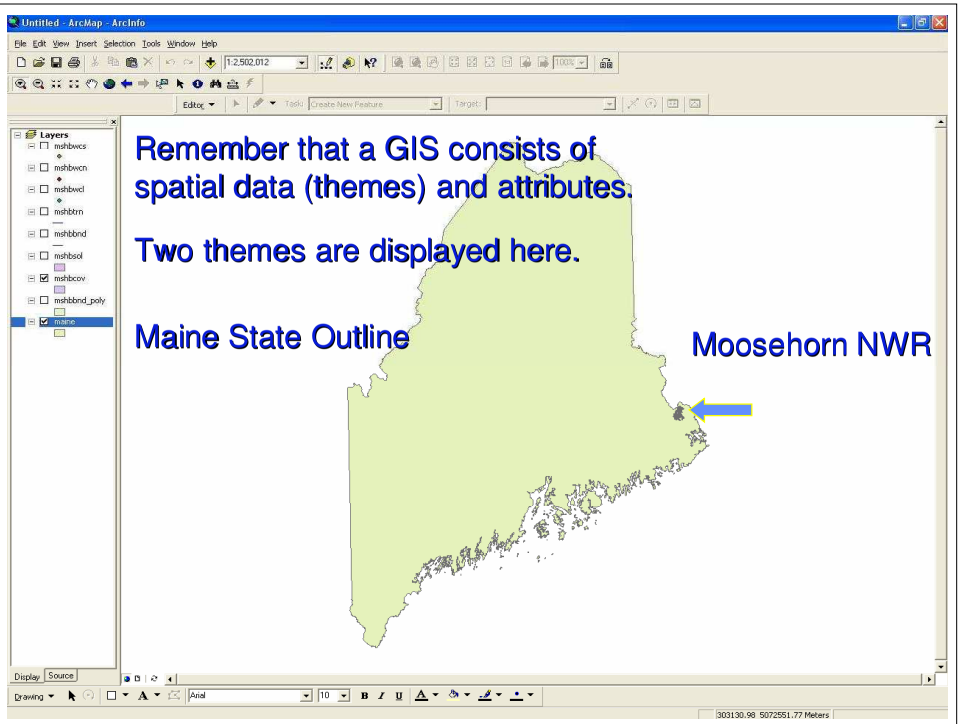
Who else uses ArcGIS?

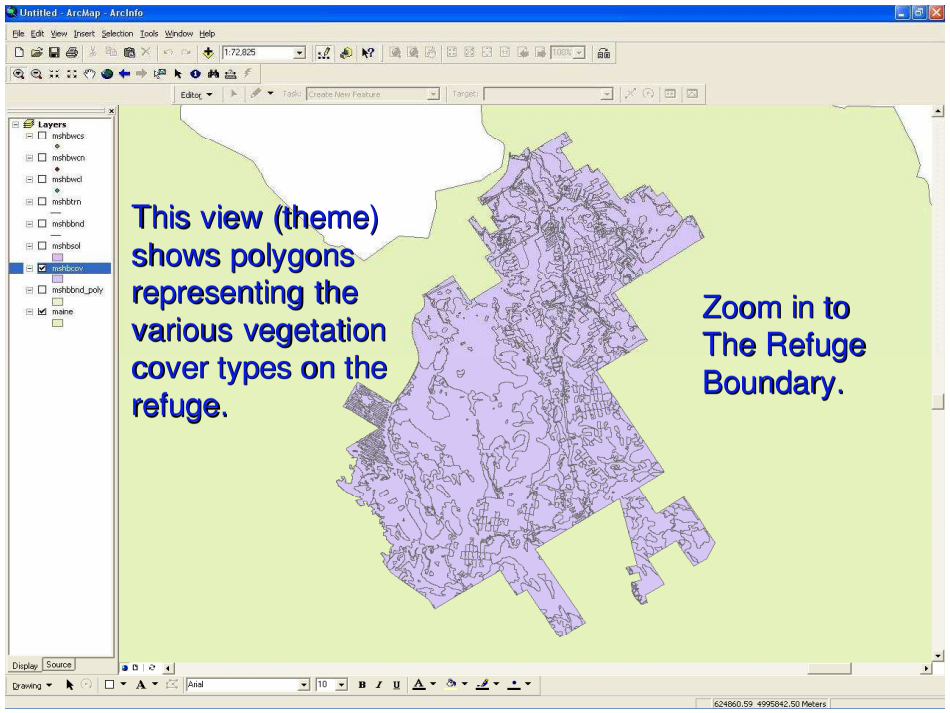
- Most Federal & State Land Management Agencies
- USGS, Forest Service, NPS, BLM, FWS

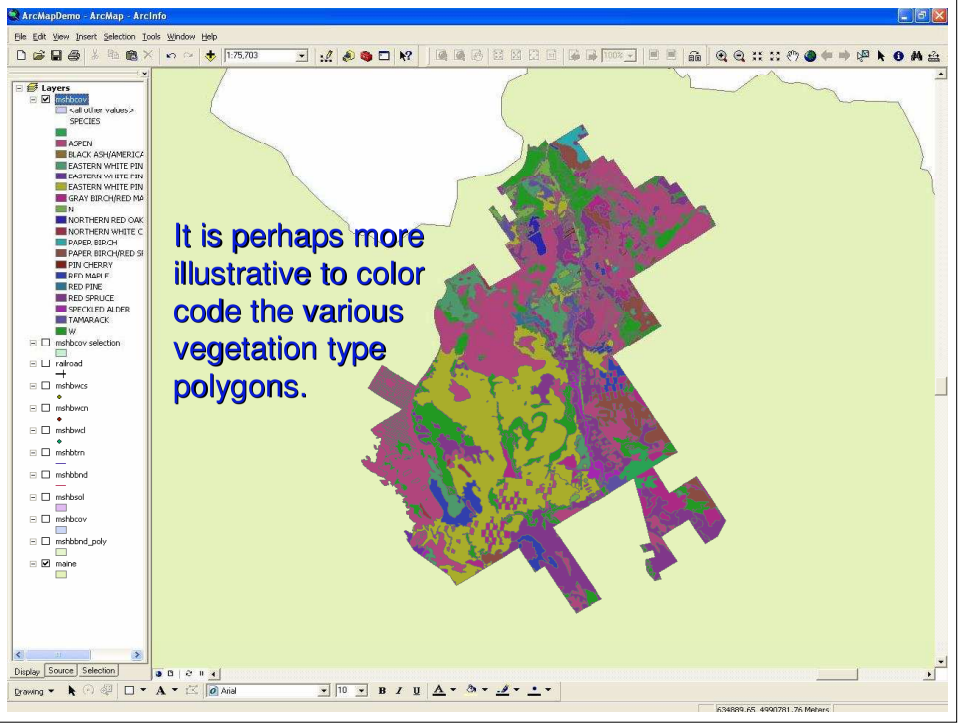
Spatial Analysis Exercises

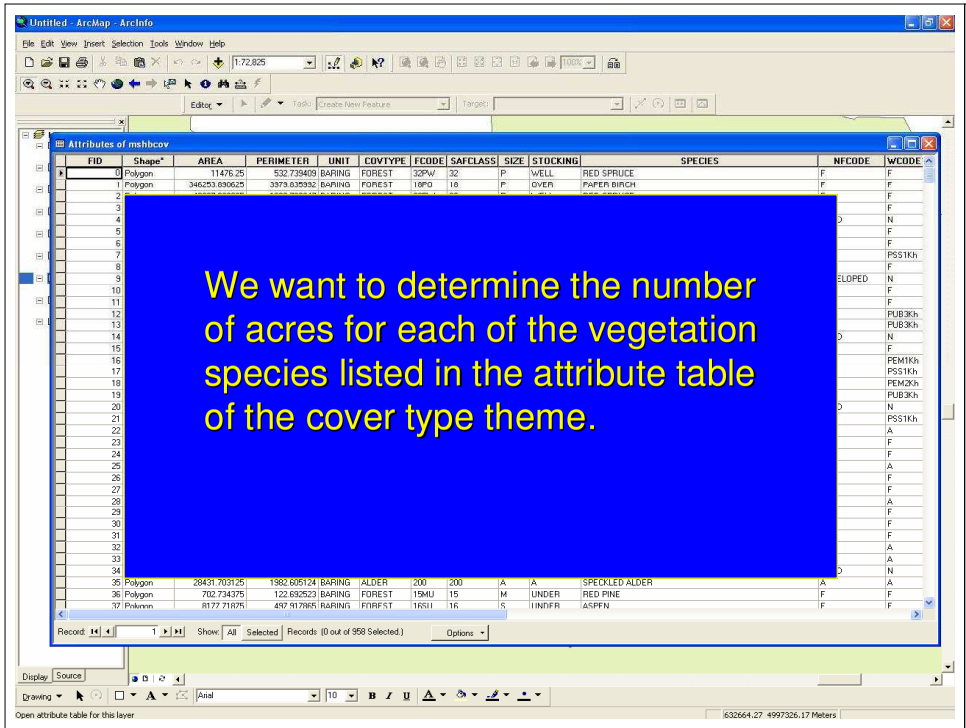
Using ArcMap

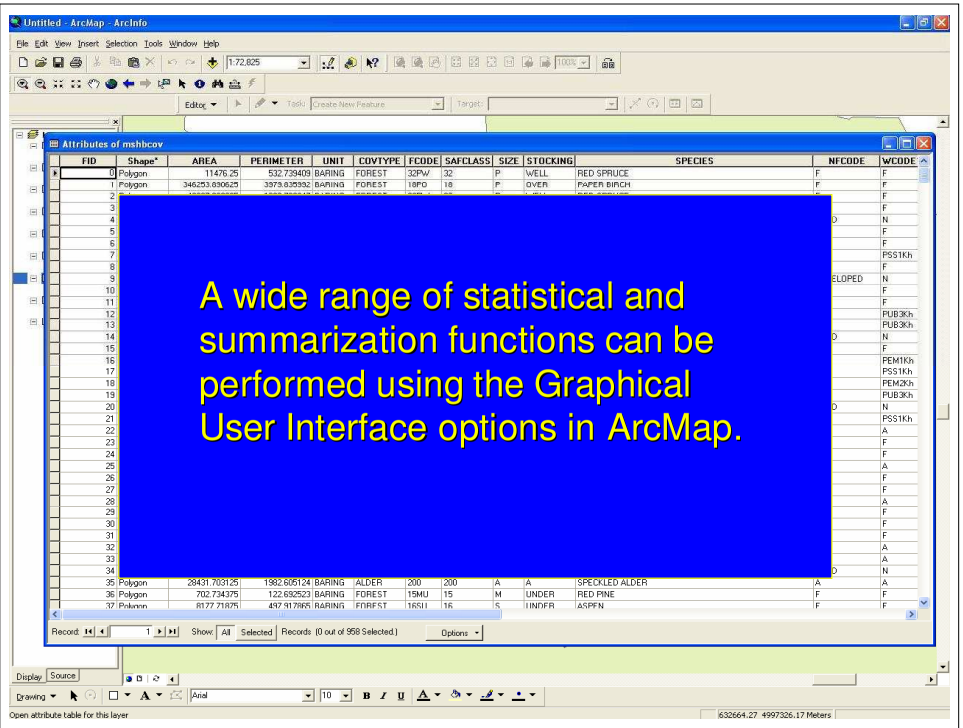
- Perform simple Descriptive Statistical Analyses
- Conduct Complex Spatial Analyses

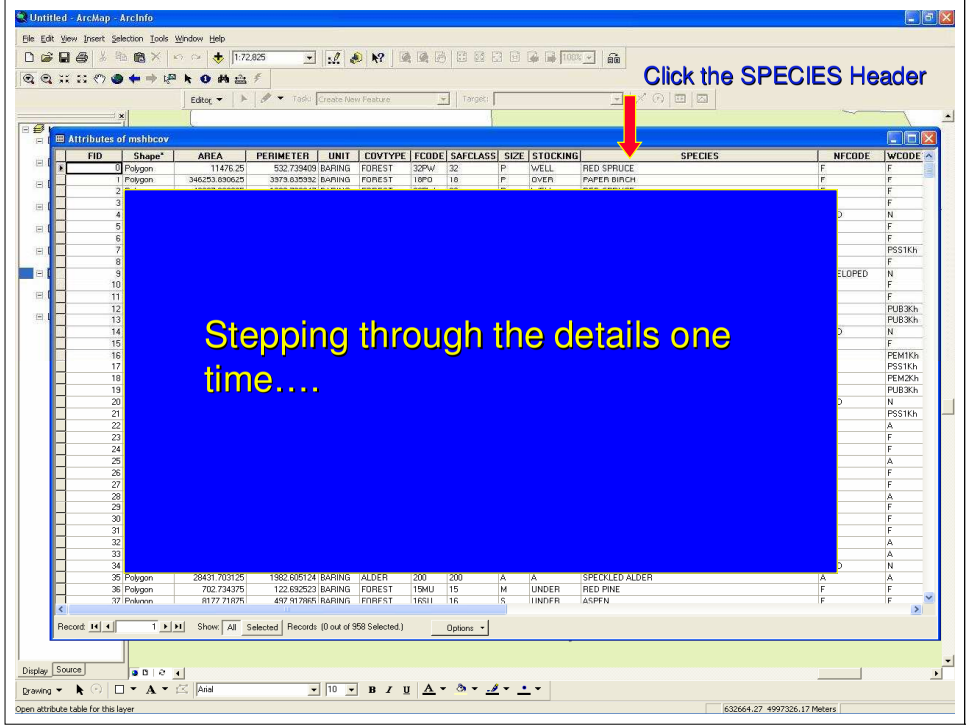


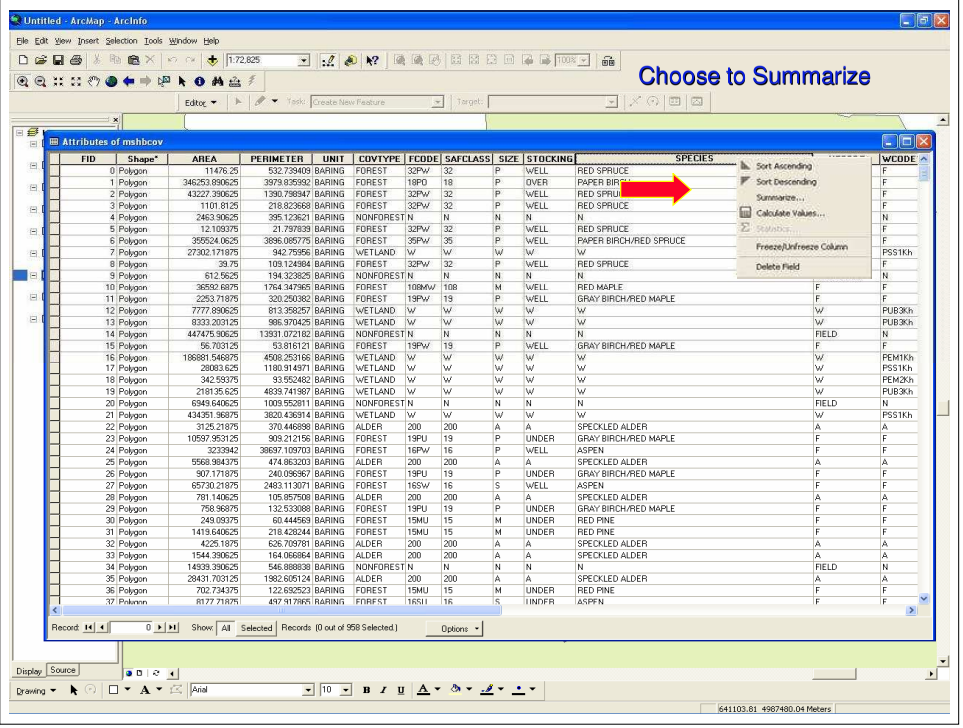


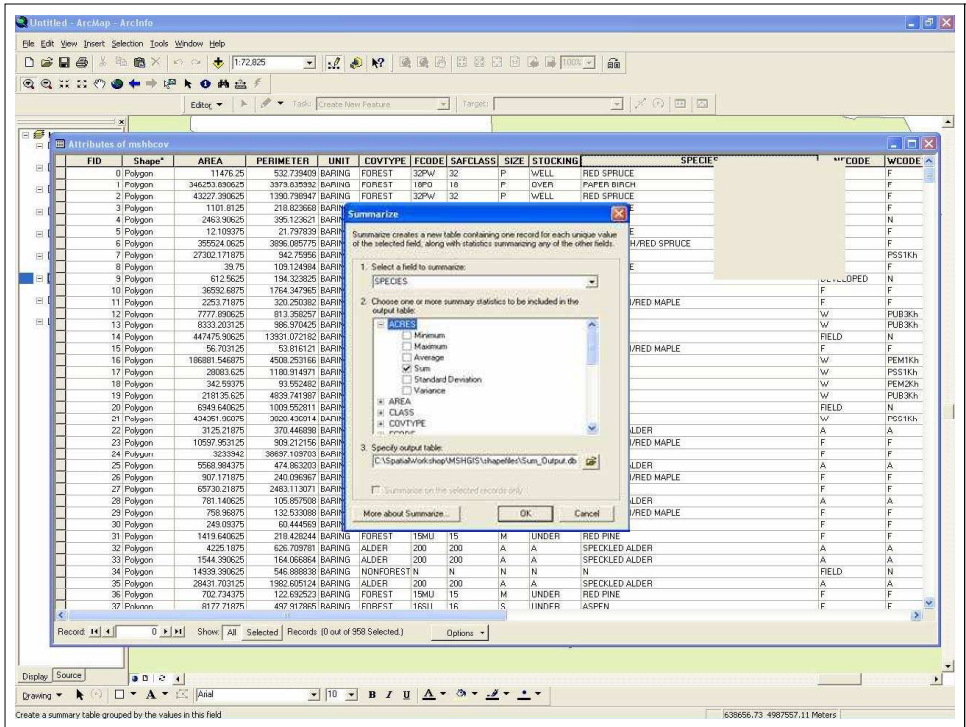


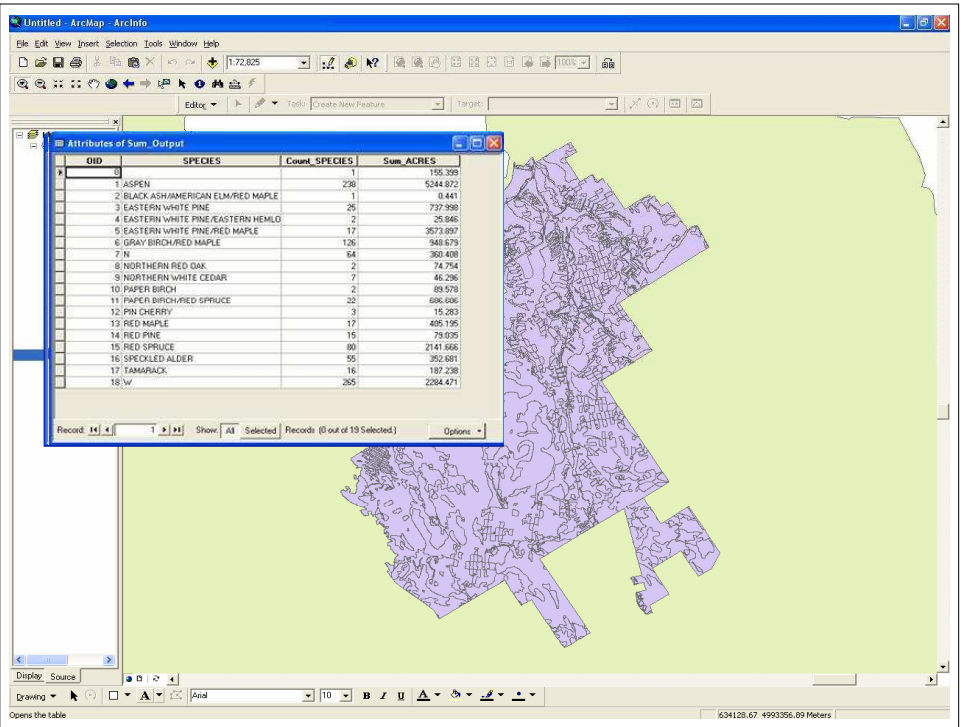

We want to determine the number of acres for each of the vegetation species listed in the attribute table of the cover type theme.

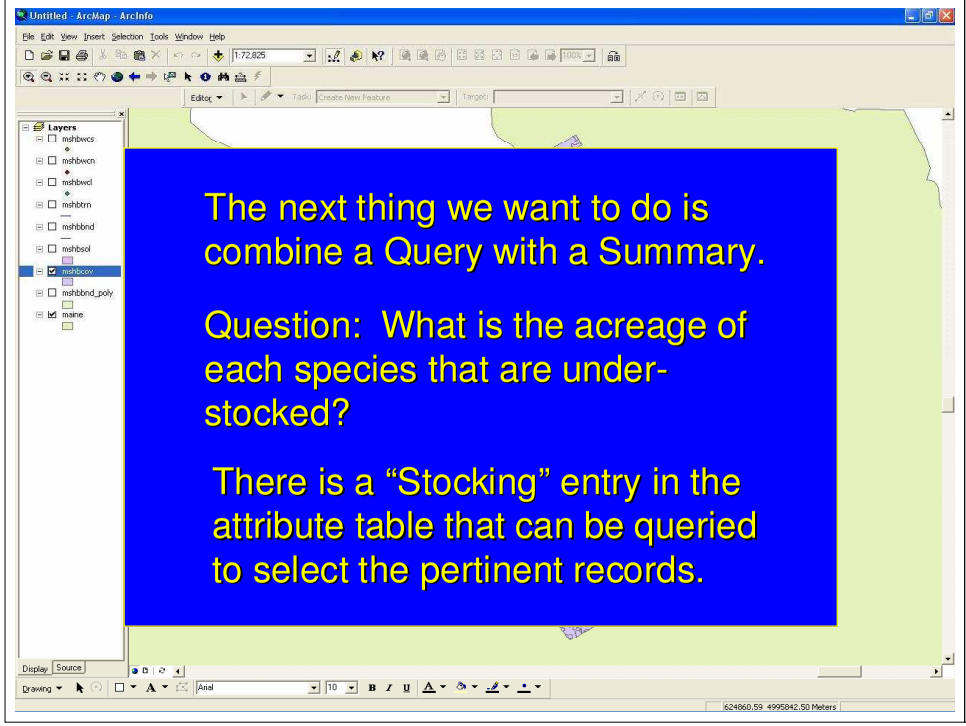

A wide range of statistical and summarization functions can be performed using the Graphical User Interface options in ArcMap.

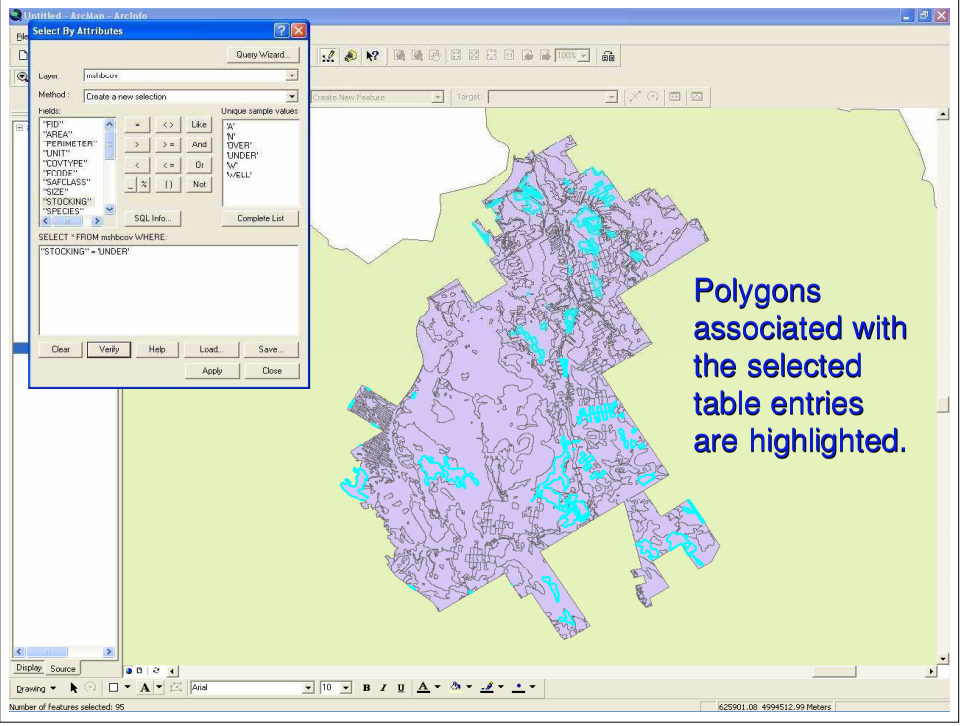

Click the SPECIES Header
Stepping through the details one time….

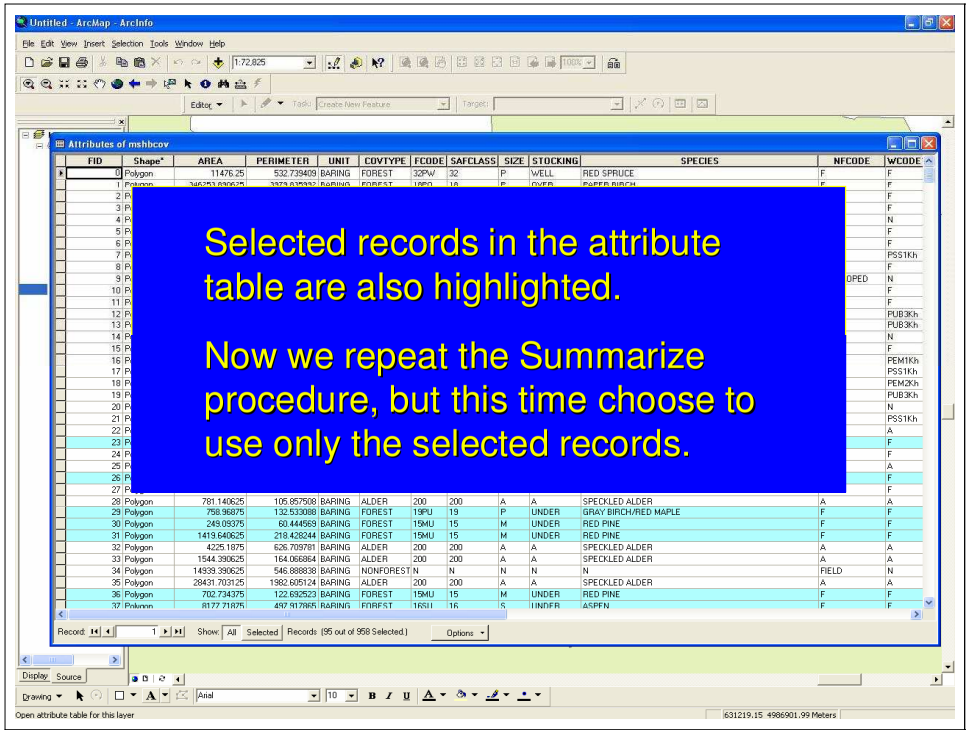

Choose to Summarize

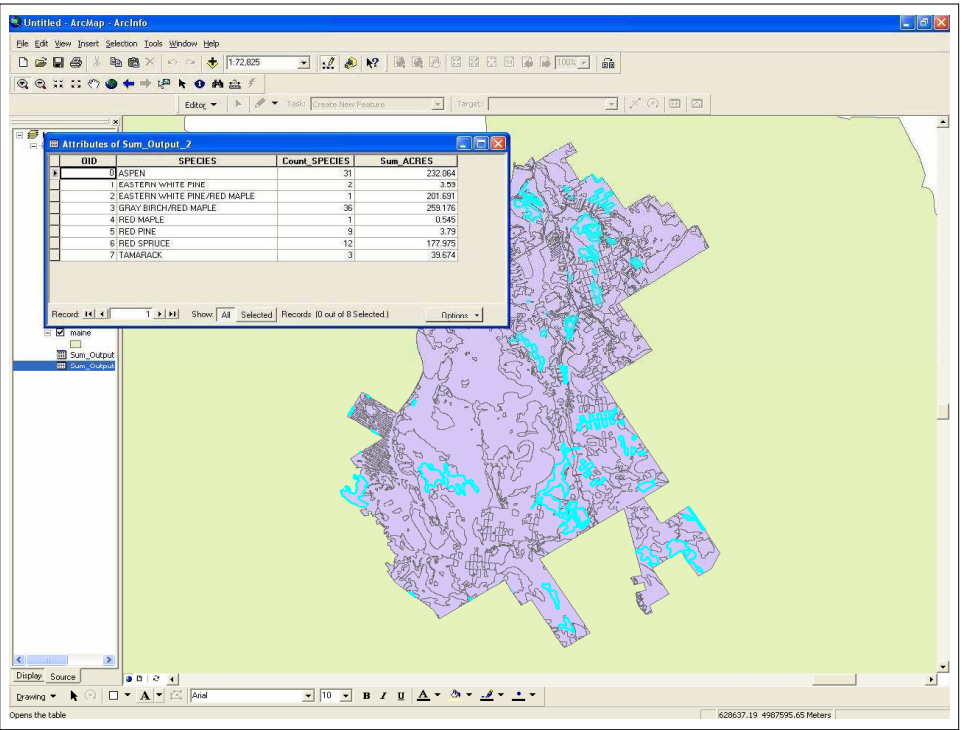

Untitled - ArcMap - ArcInfo
Attributes of mshbcov
Summarize
Summarize creates a new table containing one record for each unique value of the selected field, along with statistics summarizing any of the other fields.
1. Select a field to summarize:
SPECIES
2. Choose one or more summary statistics to be included in the output table:
ACRES
Minimum
Maximum
Average
Sum
Standard Deviation
Variance
AREA
CLASS
COVTYPE
3. Specify output table:
C:\SpatialWorkshop\MSHGIS\shapefiles\Sum_Output.db
More about Summarize...
OK
Cancel

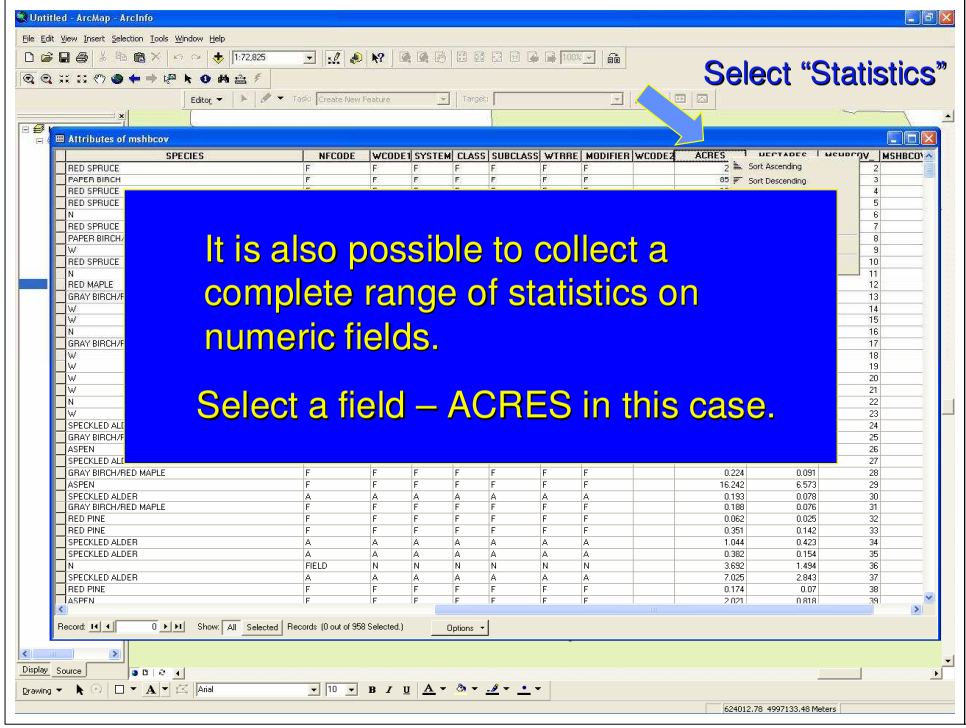

Untitled - ArcMap - ArcInfo
Attributes of Sum_Output
OID
SPECIES
Count_SPECIES
Sum_ACRES
1 ASPEN 238 5244.872
2 BLACK ASH/AMERICAN ELM/RED MAPLE 1 0.441
3 EASTERN WHITE PINE 25 737.998
4 EASTERN WHITE PINE/EASTERN HEMLO 2 25.846
5 EASTERN WHITE PINE/RED MAPLE 17 3573.897
6 GRAY BIRCH/RED MAPLE 126 948.579
7 N 64 360.408
8 NORTHERN RED OAK 2 74.754
9 NORTHERN WHITE CEDAR 7 46.296
10 PAPER BIRCH 2 89.579
11 PAPER BIRCH/RED SPRUCE 22 686.606
12 PIN CHERRY 3 15.283
13 RED MAPLE 17 405.195
14 RED PINE 15 79.035
15 RED SPRUCE 80 2141.666
16 SPECKLED ALDER 55 352.681
17 TAMARACK 16 187.238
18 W 265 2284.471

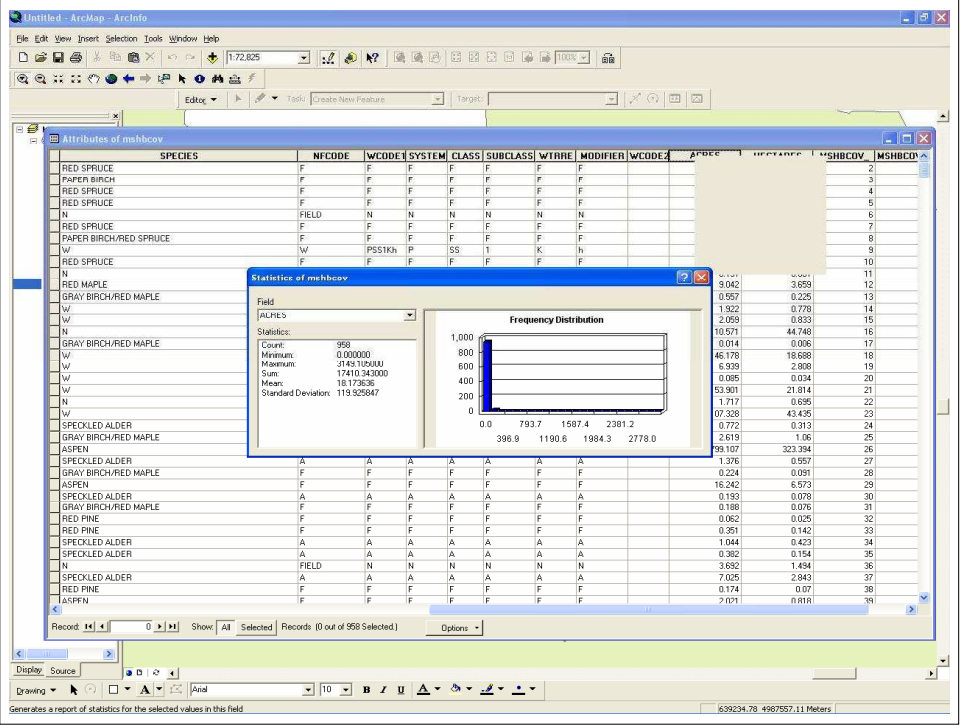

Untitled - ArcMap - ArcInfo
The next thing we want to do is combine a Query with a Summary.
Question: What is the acreage of each species that are under-stocked?
There is a “Stocking” entry in the attribute table that can be queried to select the pertinent records.


Untitled - ArcMap - ArcInfo
Select By Attributes
SELECT * FROM mshbcov WHERE:
"STOCKING" = 'UNDER'
Polygons associated with the selected table entries are highlighted.


Selected records in the attribute table are also highlighted.
Now we repeat the Summarize procedure, but this time choose to use only the selected records.




Select "Statistics"
It is also possible to collect a complete range of statistics on numeric fields.
Select a field – ACRES in this case.

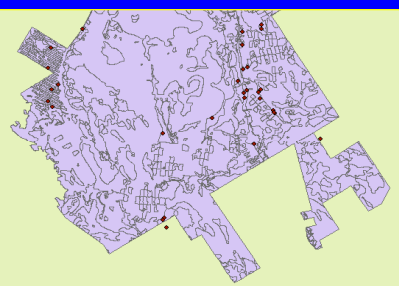


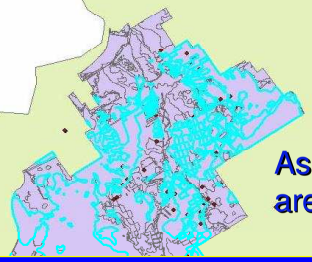
The next example is to find point features that are completely within polygon features.

Task: Find woodcock nest locations that are completely within aspen stands at Moosehorn NWR.

We can display the point theme showing woodcock nest locations.

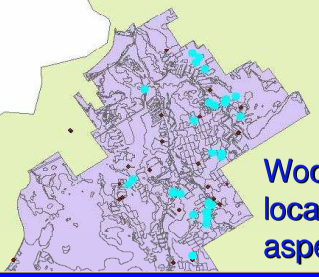
The next step is to select the aspen stands by the same method used to find the under-stocked polygons earlier.





Aspen stands are highlighted

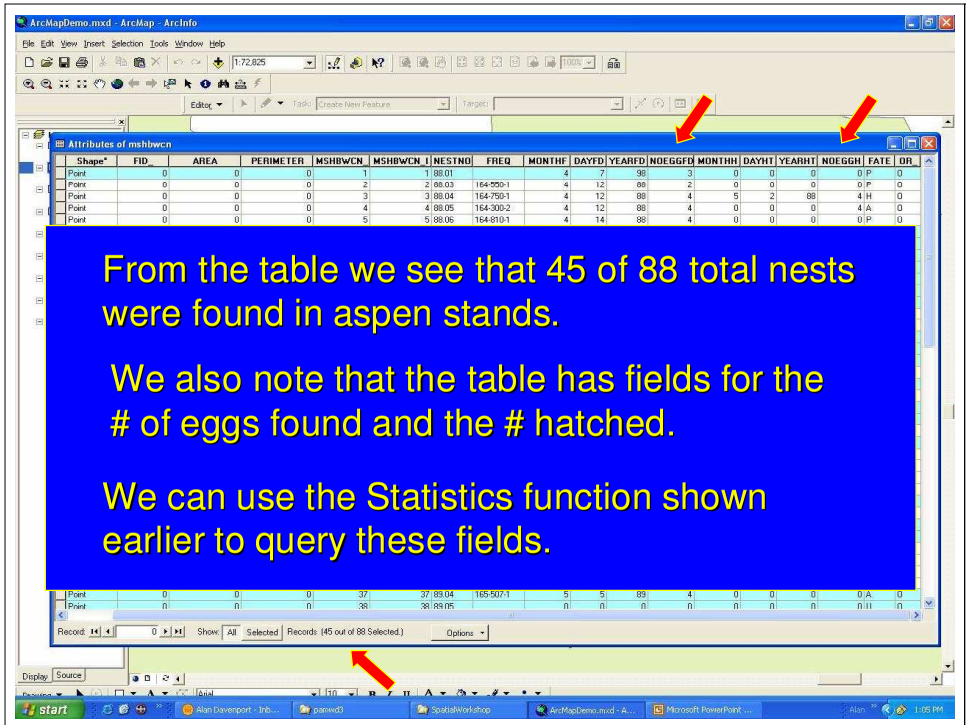
Using the query interface in ArcMap, we can ask for all woodcock nests that occur within the highlighted polygons (Aspen stands).

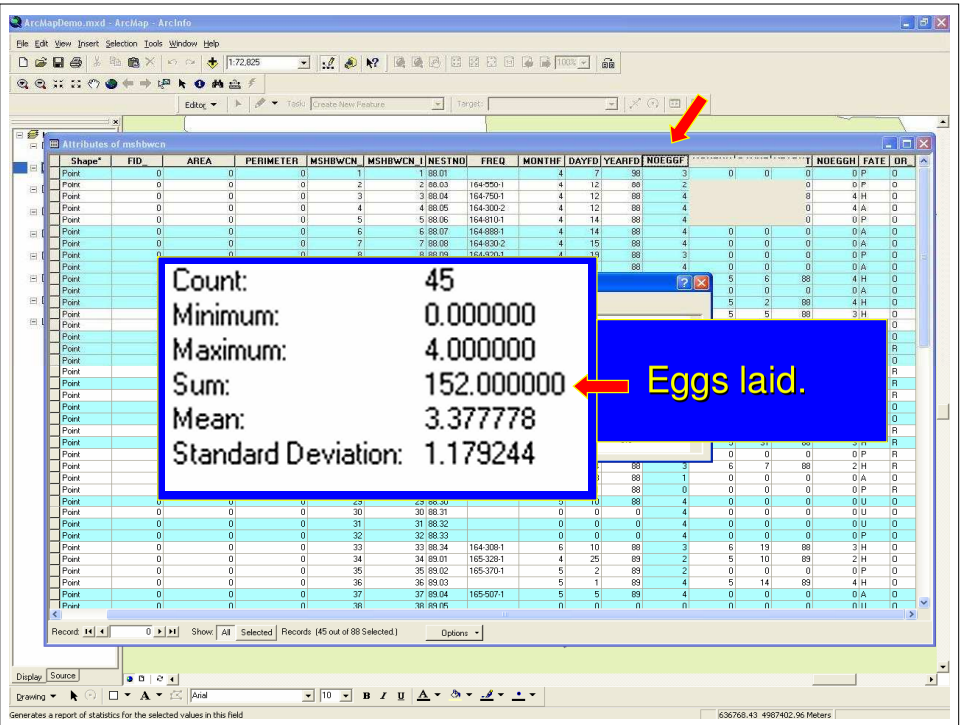


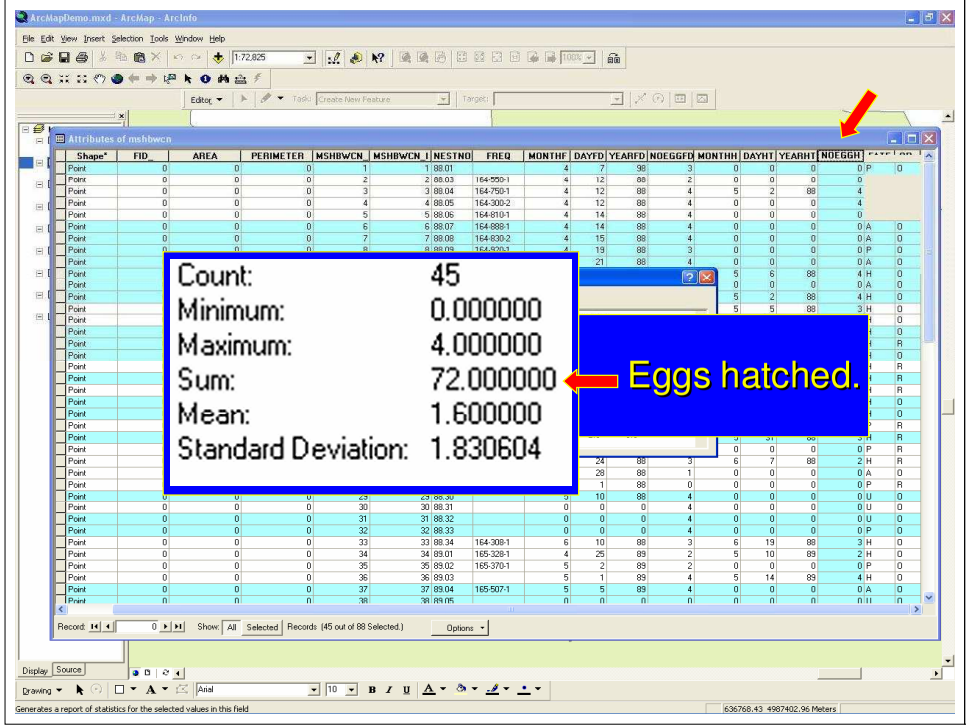
Woodcock Nest locations within aspen stands.

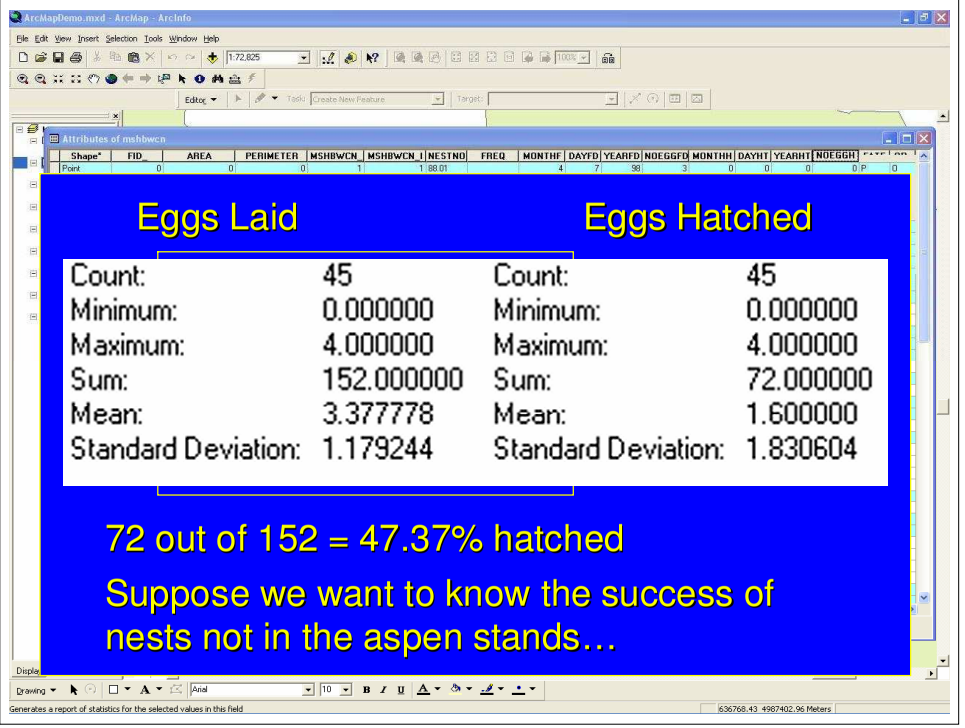
Suppose we wanted to know how many eggs hatched from these nests….

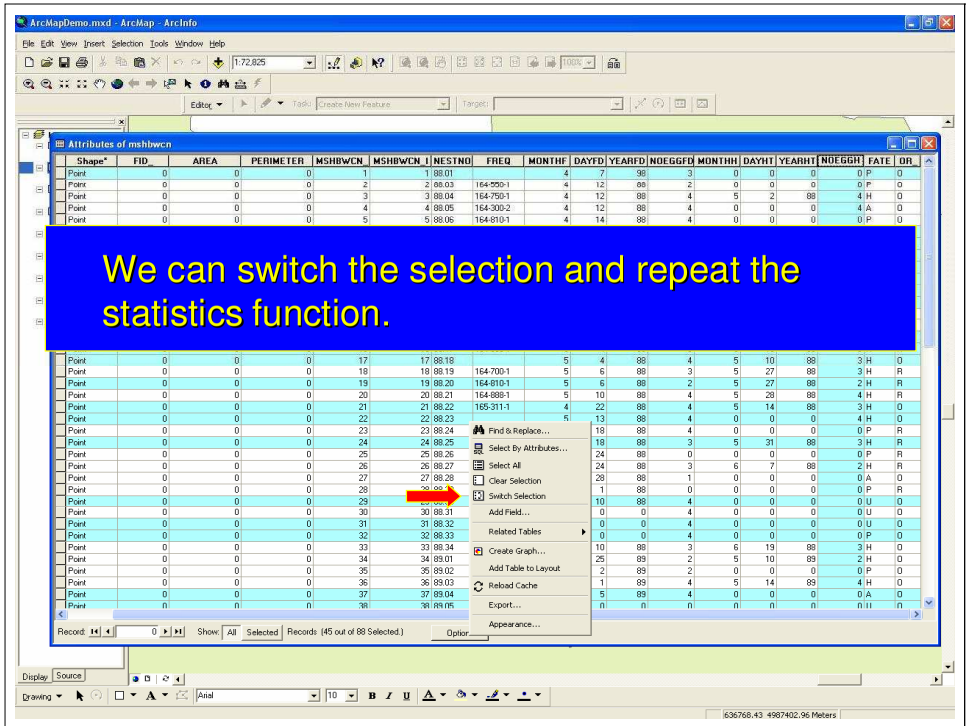
We would need to query the selected records in the nest attribute table.

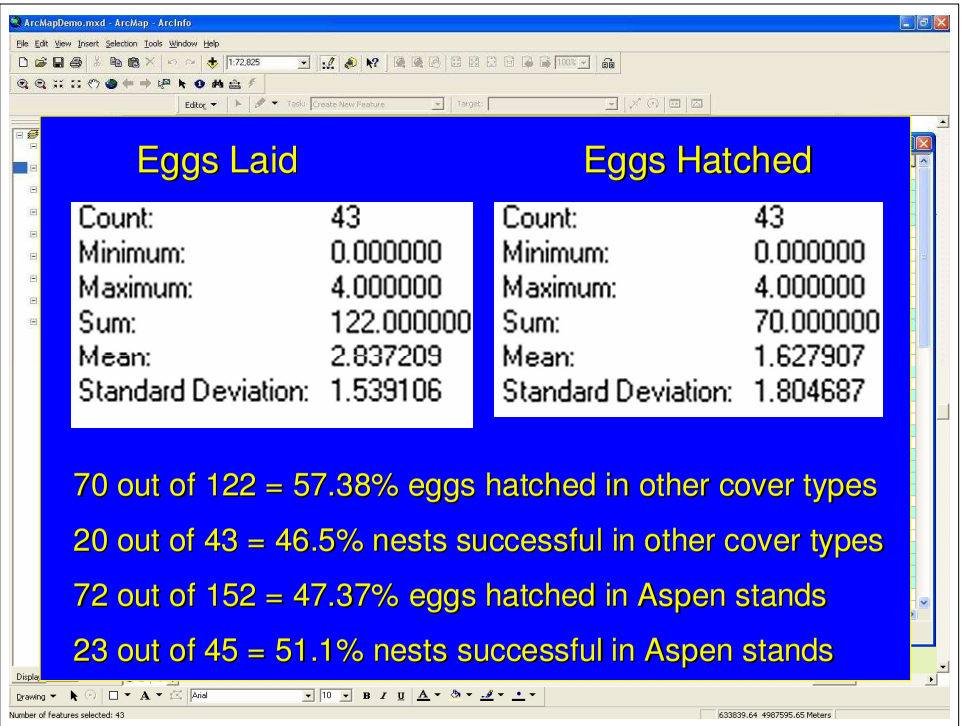

From the table we see that 45 of 88 total nests were found in aspen stands.
We also note that the table has fields for the # of eggs found and the # hatched.
We can use the Statistics function shown earlier to query these fields.

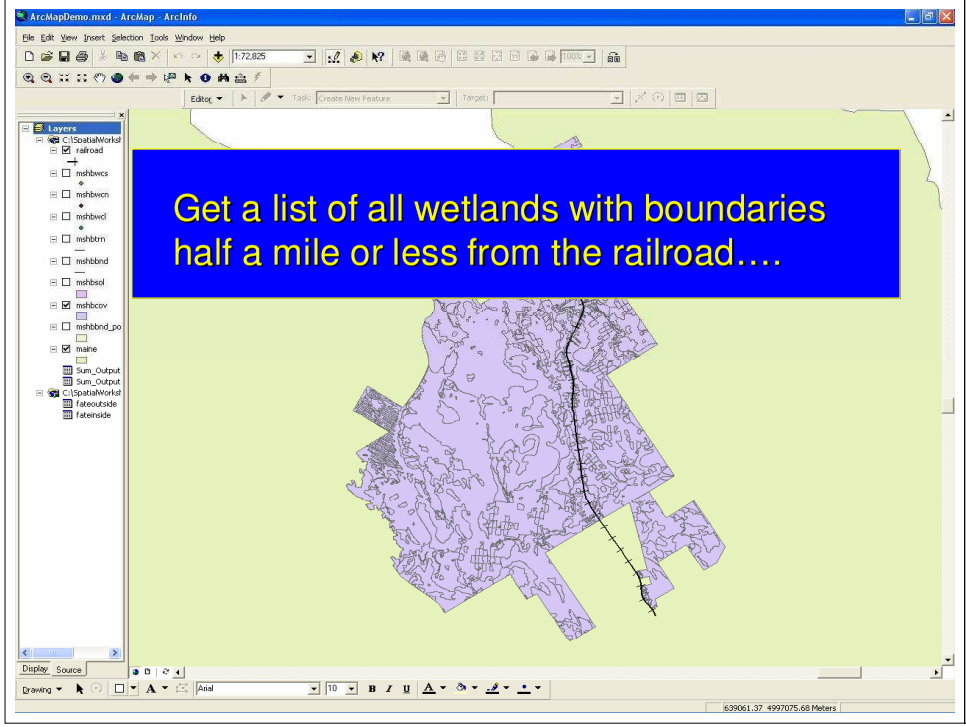

Count: 45
Minimum: 0.000000
Maximum: 4.000000
Sum: 152.000000
Mean: 3.377778
Standard Deviation: 1.179244
Eggs laid.

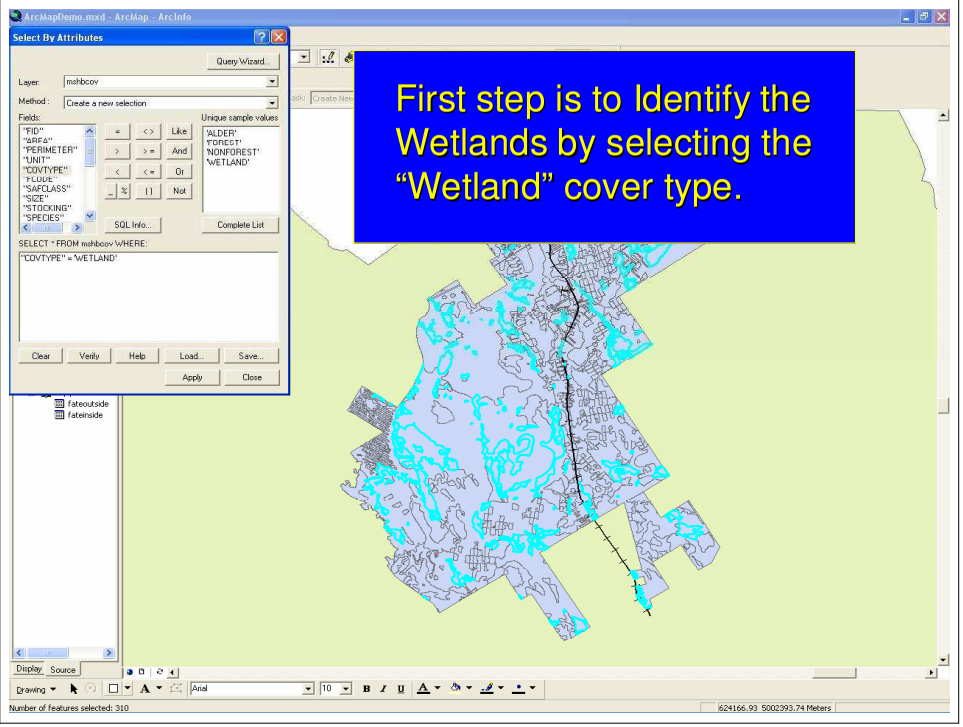

Count: 45
Minimum: 0.000000
Maximum: 4.000000
Sum: 72.000000
Mean: 1.600000
Standard Deviation: 1.830604
Eggs hatched.

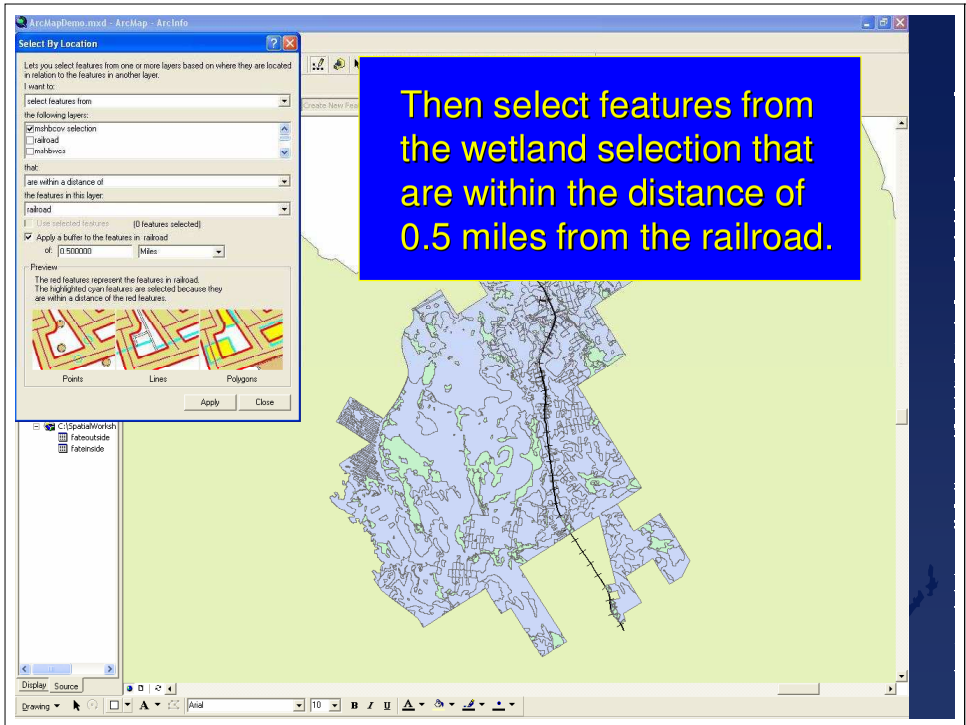

Eggs Laid
Count: 45
Minimum: 0.000000
Maximum: 4.000000
Sum: 152.000000
Mean: 3.377778
Standard Deviation: 1.179244
Eggs Hatched
Count: 45
Minimum: 0.000000
Maximum: 4.000000
Sum: 72.000000
Mean: 1.600000
Standard Deviation: 1.830604
72 out of 152 = 47.37% hatched
Suppose we want to know the success of nests not in the aspen stands…

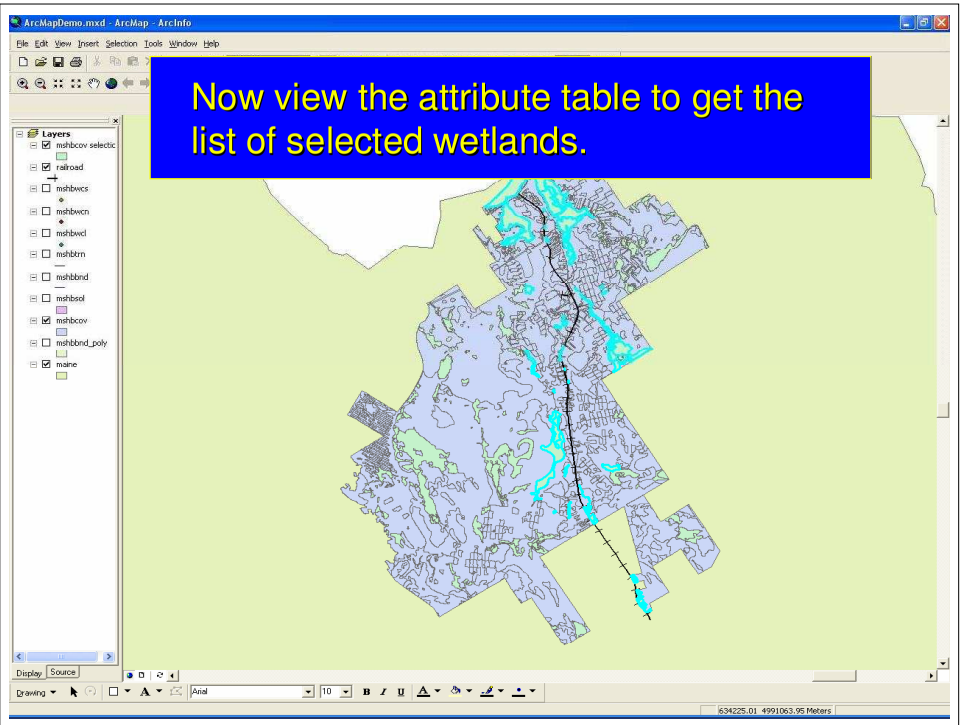

ArcMapDemo.mxd - ArcMap - ArcInfo
Attributes of mshbwcn
We can switch the selection and repeat the statistics function.

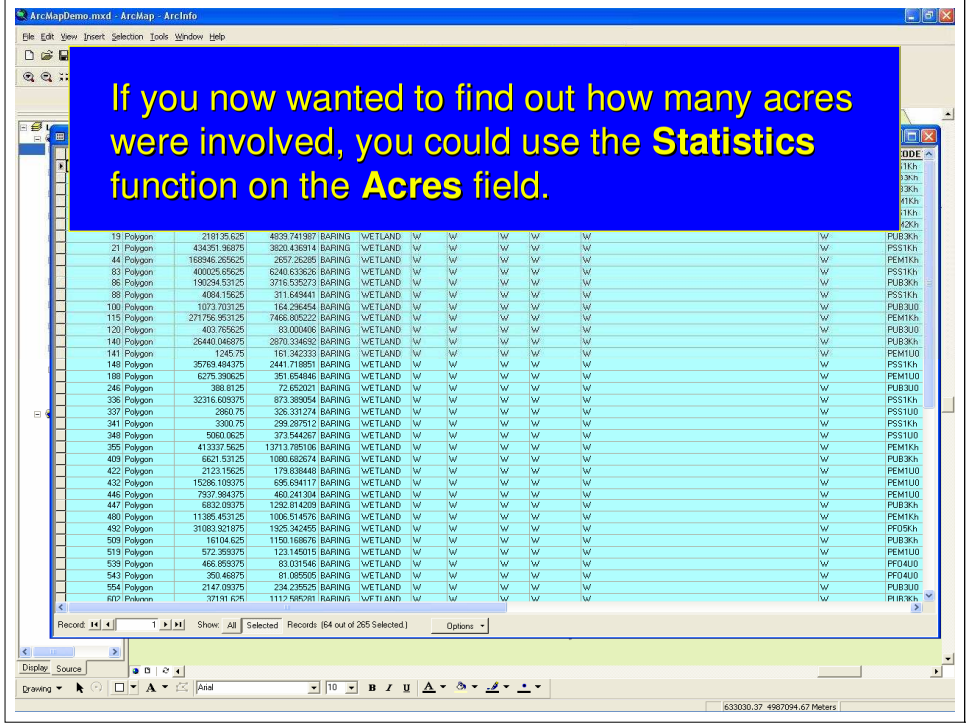

ArcMapDemo.mxd - ArcMap - ArcInfo
Eggs Laid
Count: 43
Minimum: 0.000000
Maximum: 4.000000
Sum: 122.000000
Mean: 2.837209
Standard Deviation: 1.539106
Eggs Hatched
Count: 43
Minimum: 0.000000
Maximum: 4.000000
Sum: 70.000000
Mean: 1.627907
Standard Deviation: 1.804687
70 out of 122 = 57.38% eggs hatched in other cover types
20 out of 43 = 46.5% nests successful in other cover types
72 out of 152 = 47.37% eggs hatched in Aspen stands
23 out of 45 = 51.1% nests successful in Aspen stands

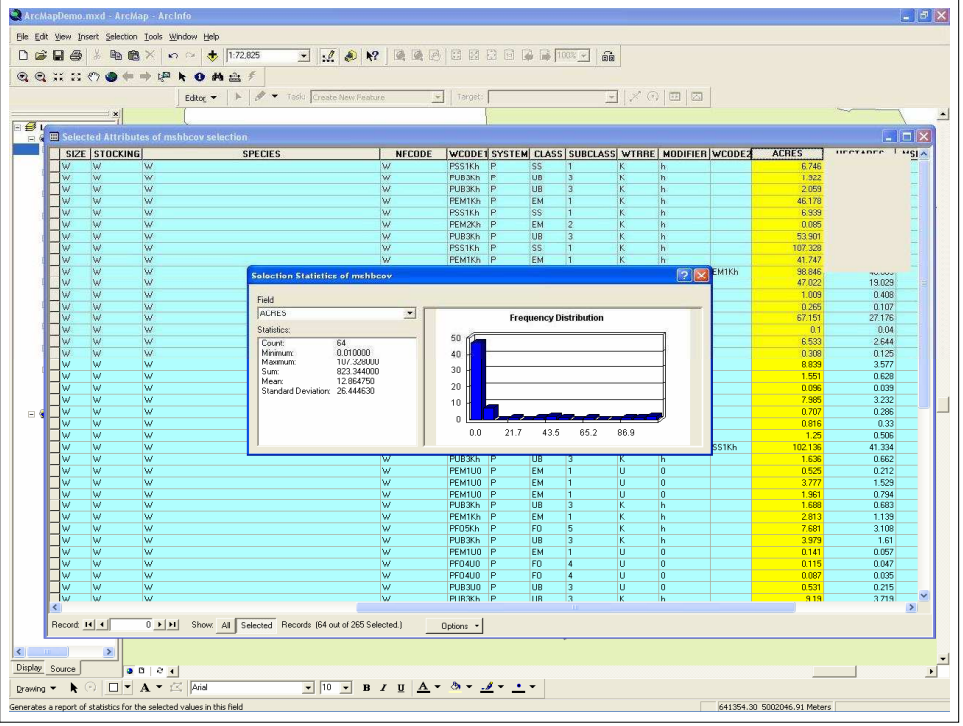

ArcMapDemo.mxd - ArcMap - ArcInfo
Get a list of all wetlands with boundaries half a mile or less from the railroad….


ArcMapDemo.mxd - ArcMap - ArcInfo
Select By Attributes
SELECT * FROM mshbcov WHERE:
"COVTYPE" = 'WETLAND'
First step is to Identify the Wetlands by selecting the "Wetland" cover type.


Then select features from the wetland selection that are within the distance of 0.5 miles from the railroad.


Now view the attribute table to get the list of selected wetlands.


If you now wanted to find out how many acres were involved, you could use the **Statistics** function on the **Acres** field.

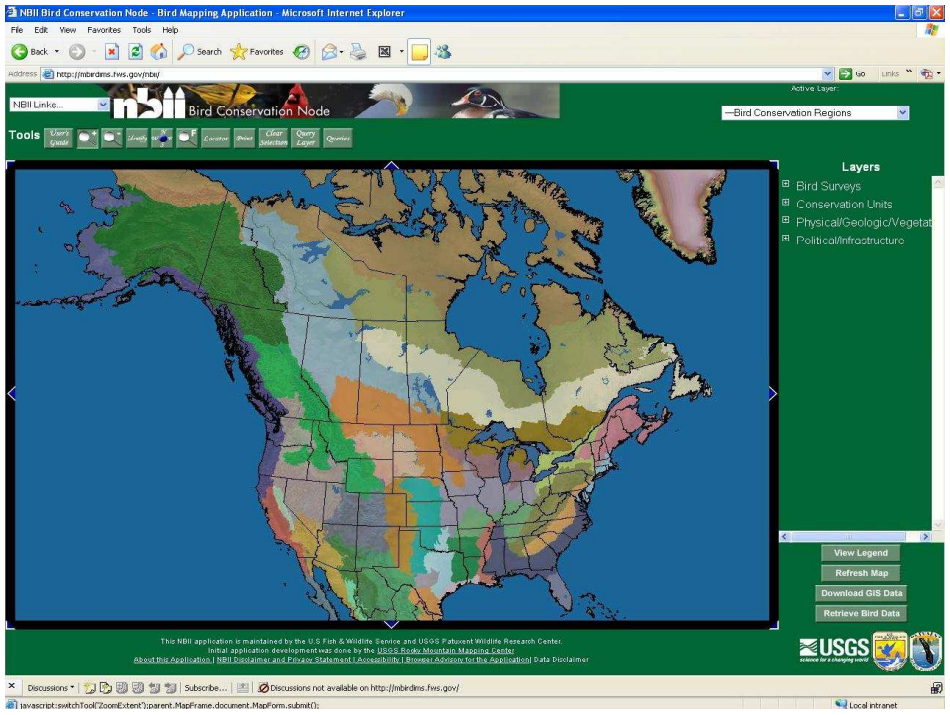


ArcIMS

- Internet Map Server
- Provides for viewing and manipulation of spatial data over the Internet.
- Our office hosts an ArcIMS application for the Bird Conservation Node of the NBII (National Biological Information Infrastructure)
- The application is reachable through the URL: http://mbirdims.fws.gov

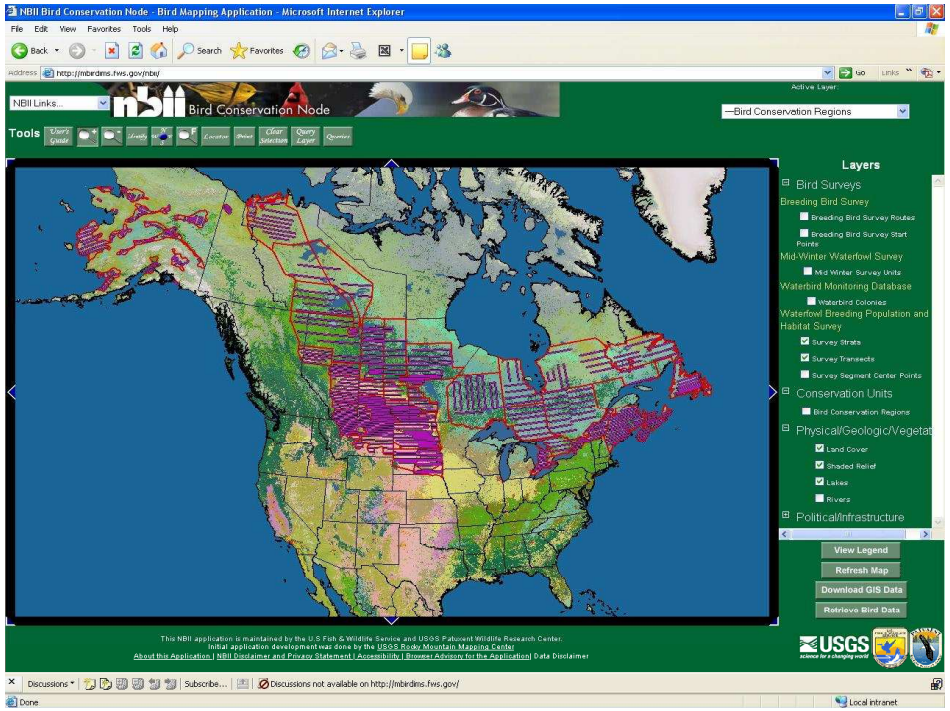
ArcIMS

- The next screen is the initial view presented when the web site is accessed. It shows most of North America. The shaded areas represent bird conservation regions. You can click on a "View Legend" button to view the key.



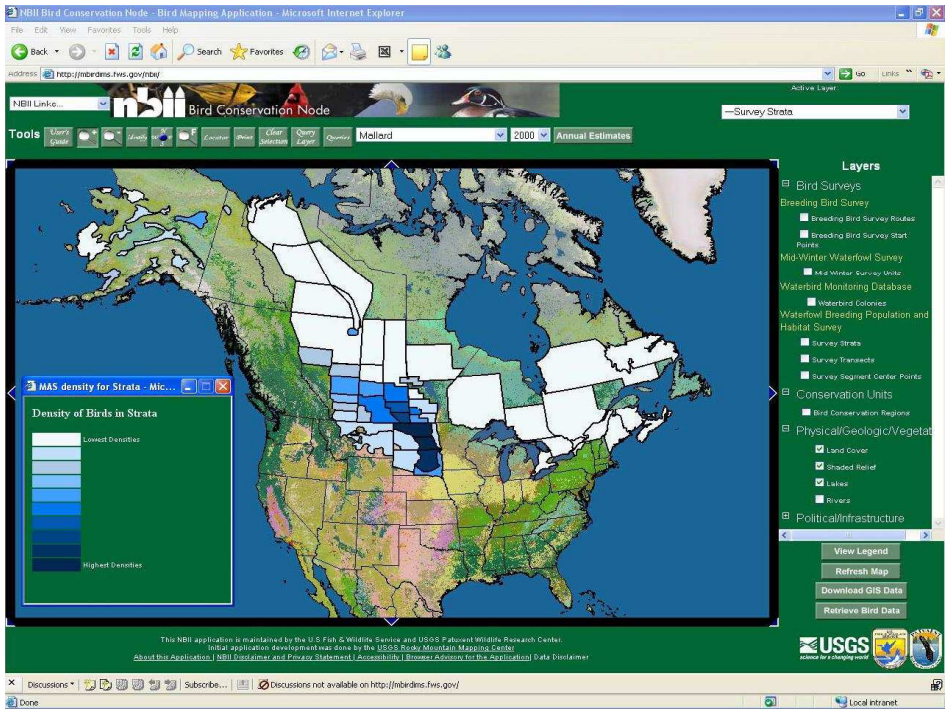
Aerial Surveys

- One of the major functions of our office is to sample the breeding grounds to estimate waterfowl populations annually.
- The next view shows the flight lines that are surveyed each year in May.



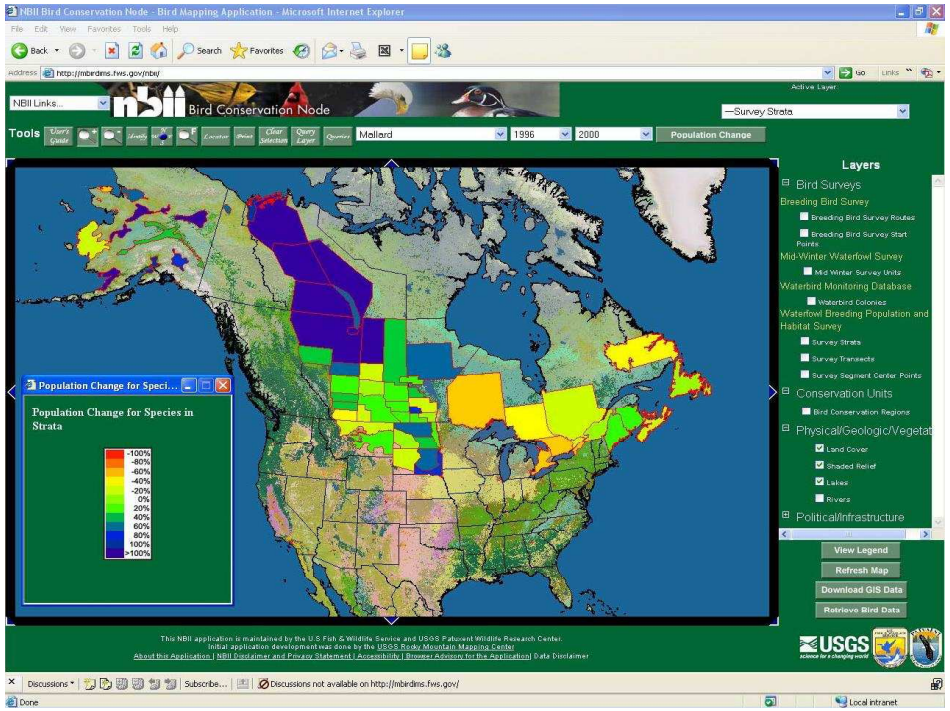
Aerial Surveys
(cont'd)

- An example of survey results is shown on the next slide.
- It shows the results of a query on Mallard Duck abundance by survey stratum.
- You may have noticed that the background has been replaced by a layer showing land cover types.



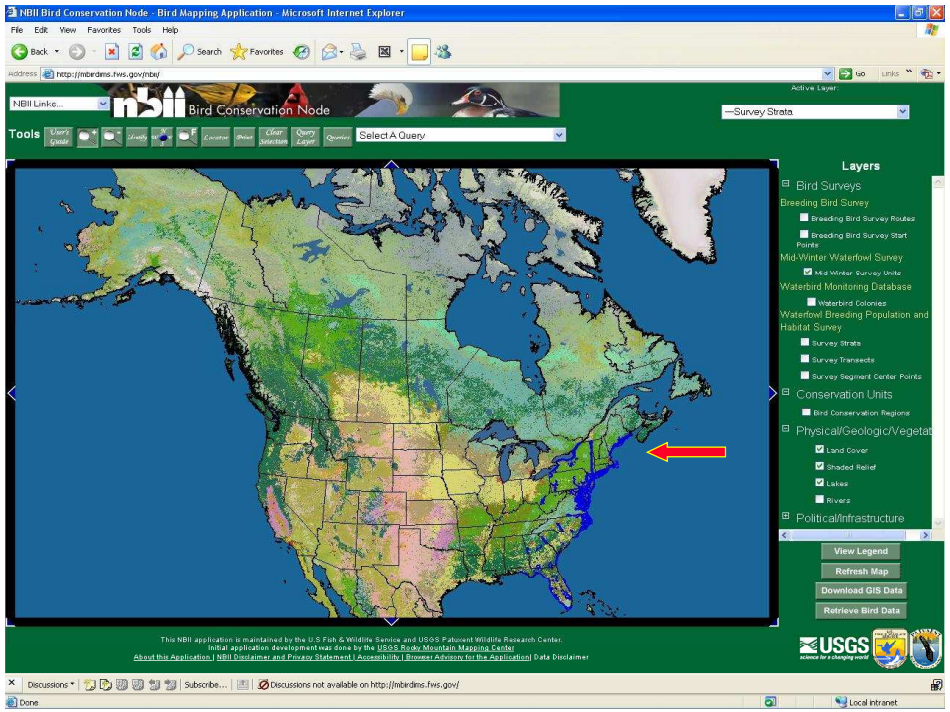
Aerial Surveys
(cont'd)

- The next example illustrates the changes in the population estimate for Mallards from 1995 to 2000.



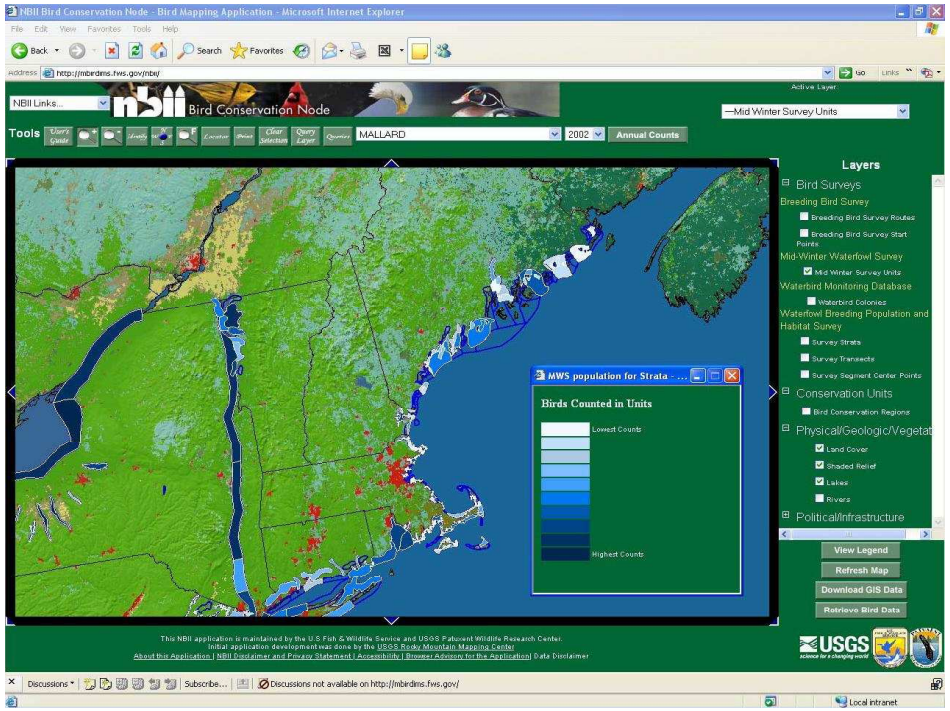
Aerial Surveys
(cont'd)

- In the next view the outlines for the Mid-Winter waterfowl survey zones have been displayed. This is a late December – January survey of wintering waterfowl, primarily along the Atlantic coast but covering inland waters of Atlantic coastal states.



Aerial Surveys
(cont'd)

- The next view shows counts of Mallards by survey unit for 2002.



Summary

This little demonstration hardly does justice to what can be done using a GIS to analyze data, but hopefully it has exposed the potentials.

But remember that you can't do anything unless you have the data necessary to answer the questions.

We'll finish up exploring a few methods for getting data into a GIS.

Polygons and Lines can be digitized from maps or other sources. Once initialized to known reference points, digitizing software automatically generates the correct geographic coordinates.

Points can be collected using GPS receivers.

Raster data is generated by photographs or satellite imagery.

Next up is a simple example of using satellite imagery as a background reference.

This is a satellite image, 978 pixels wide and 598 pixels high. The origin, pixel (0,0), is in the lower left corner.

In order to use this image effectively in a GIS, the coordinates of each pixel must be transformed to match the other layers being used.

Original Image from previous slide

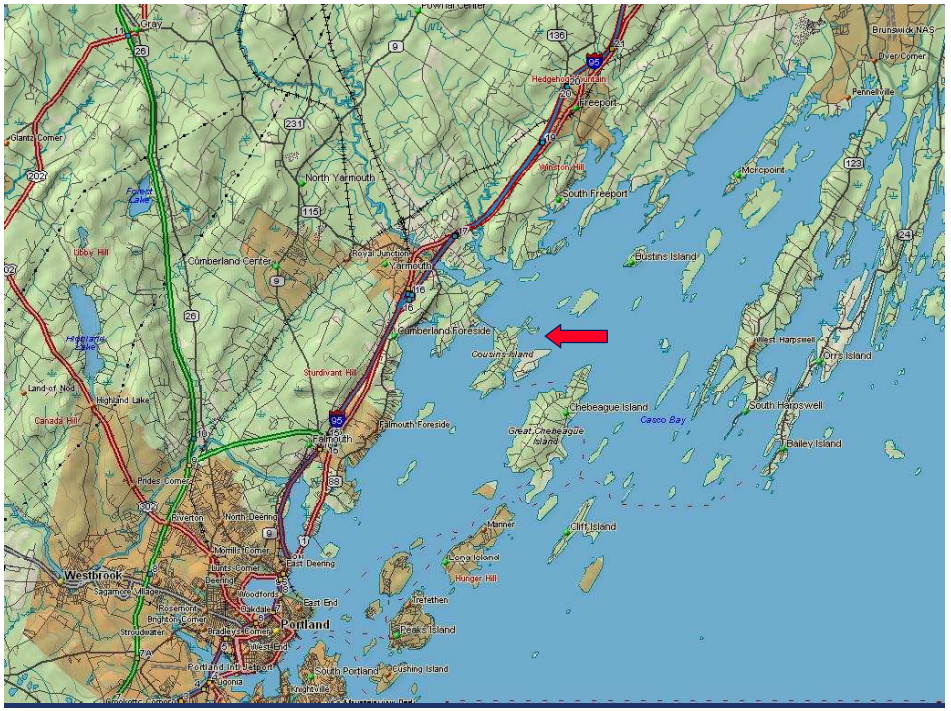
Transformed Image

Overlaid with Moosehorn cover type polygon outlines

Aerial Surveys

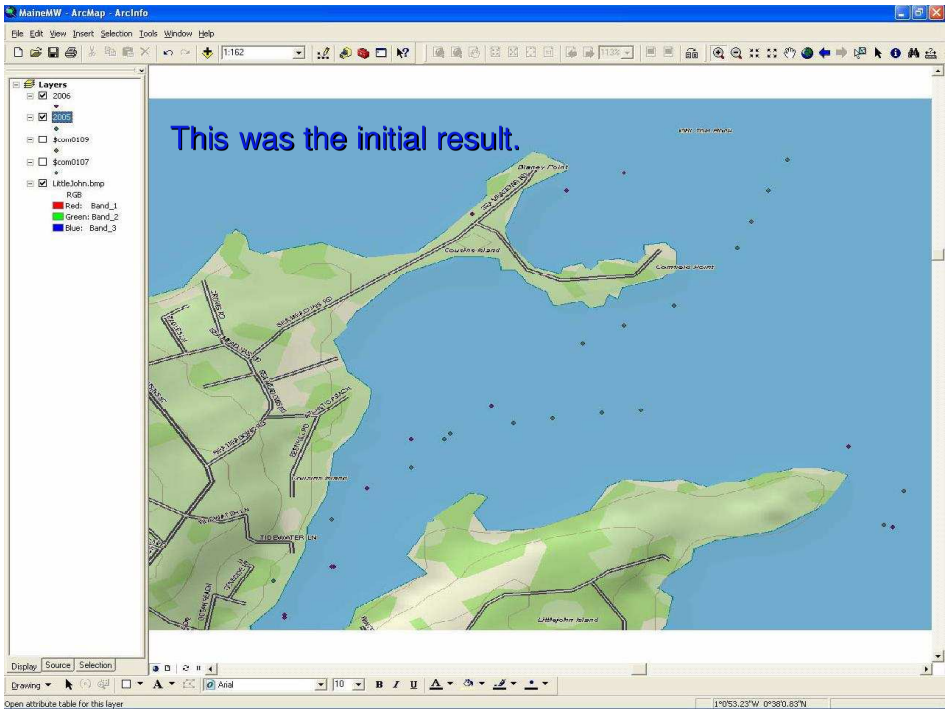
Aerial Surveys

- In January of this year I received a request through one of our pilot/biologists from a biologist in Maine who was interested in some specific data. There was concern about some development plans on Little John Island off the coast of Maine east of Yarmouth. He wondered if we could provide information about waterfowl species and counts in the area from the last two mid-winter waterfowl surveys.



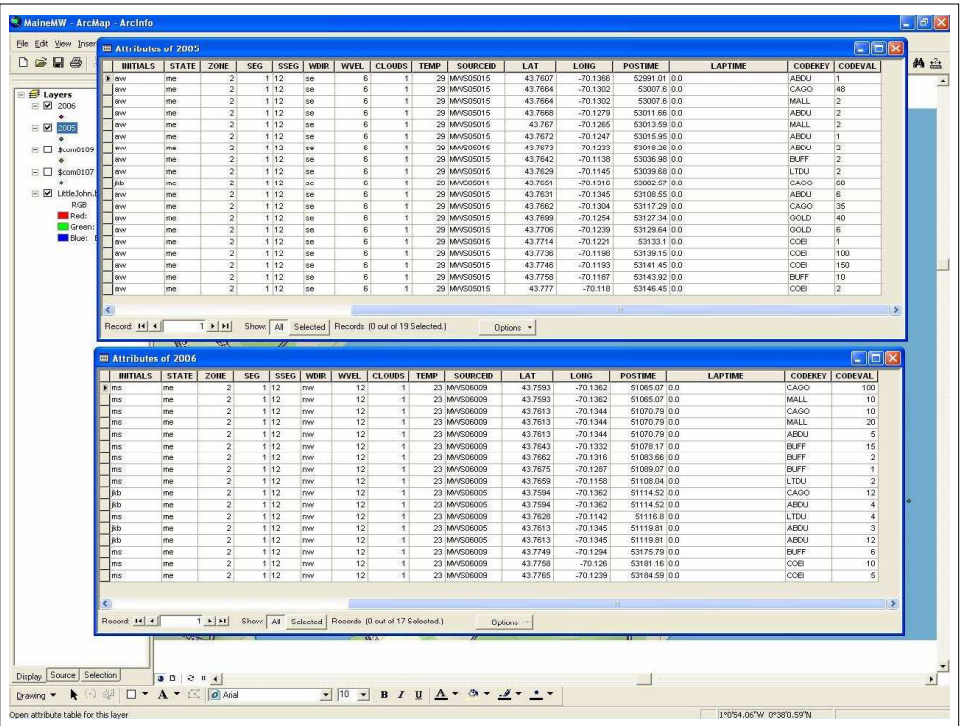
How to display "raw" survey data in a meaningful manner?

- The first thing was to create a base layer map to define the area of interest. Then the survey data points were plotted and those in the area of interest were selected using a "Select Features" tool which allowed me to "box" the area of interest and extract the attribute information for the points of interest.



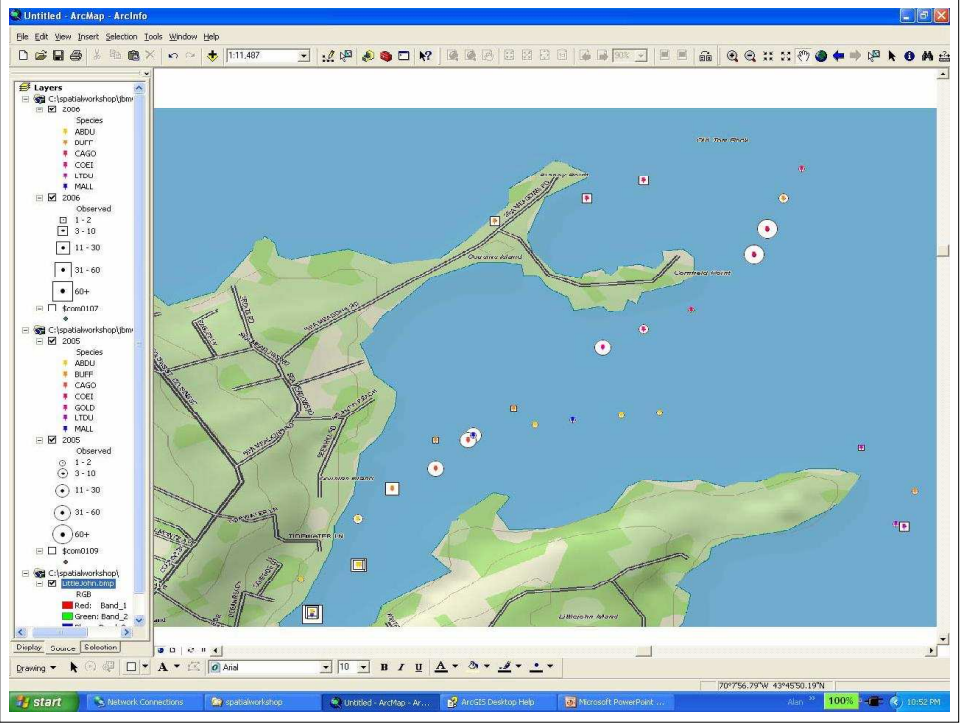
How to display "raw" survey data in a meaningful manner? (cont'd)

- There were not really many observations in the "target" area.
- Next we look at the attribute tables for the selected observations each year.



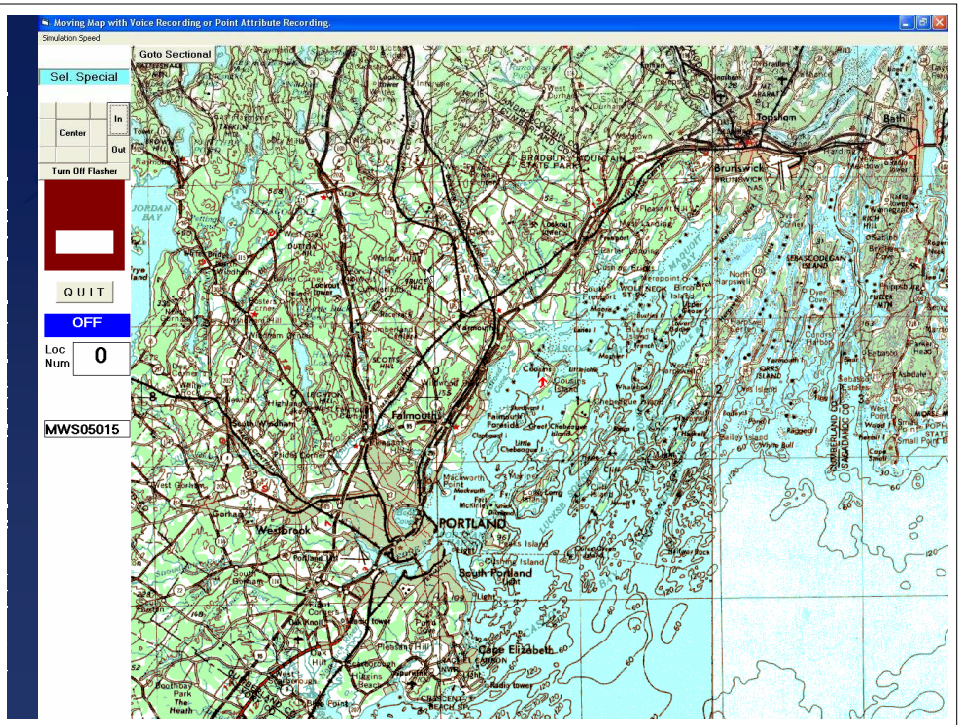
How to display "raw" survey data in a meaningful manner? (cont'd)

- Finally, symbology was selected to try and display the information in a more meaningful way than just as dots on a map.



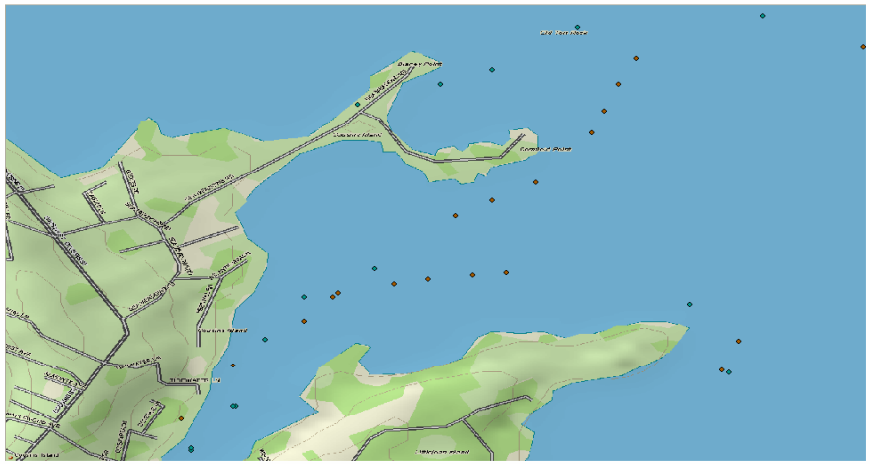
But how was this information actually collected?

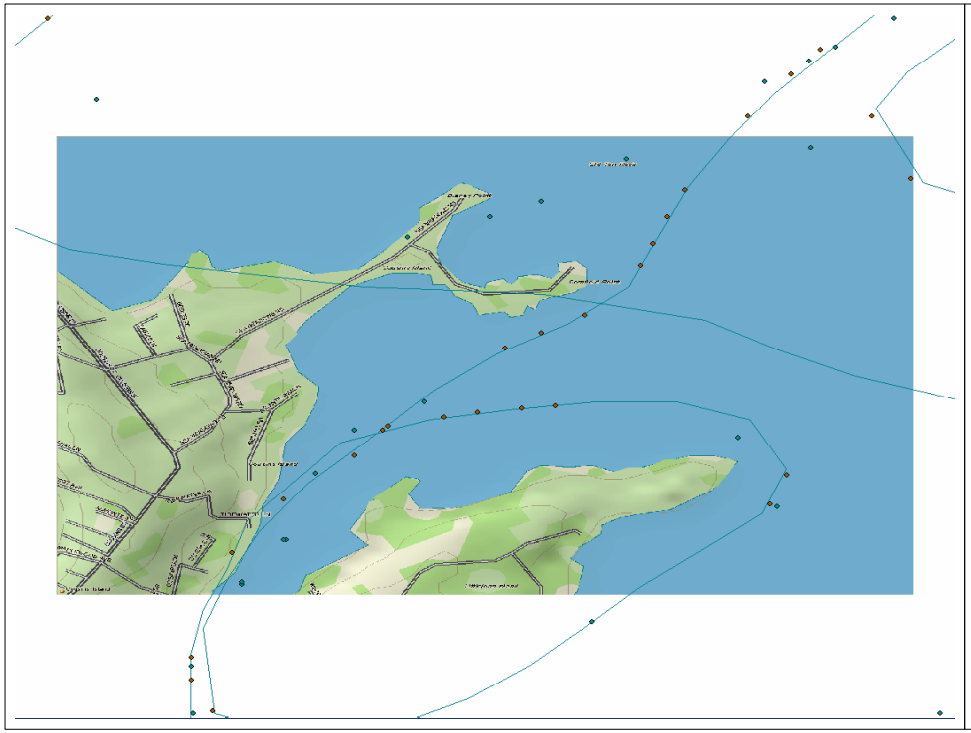
The next slide shows a simulation of part of the 2005 survey flight in the selected area.



Aerial Survey Data Collection

- Going back to the specific area of interest, we'll now see the data plots followed by an overlay of the flight path.






The End

Spatial Modeling Of Counts

J. Andrew Royle
USGS Patuxent Wildlife Research Center

Workshop on Spatial Statistics
Beltsville Agricultural Research Center
March 16, 2006

Outline

- Introduction: Count data in ecology and spatial dependence
- Generalized Linear Modeling (GLM) framework
- Spatial correlation models
- Examples: North American BBS data
- Detection bias in animal surveys

Introduction

Ecology: *The study of spatial and temporal variation in abundance*

A general theme of ecological studies: Collect spatially referenced counts, $y(s)$, with the goal of making inferences about "abundance"

For example,

- Characterize the spatial distribution of a population
- Map occurrence of a species – "range map"
- Evaluate landscape factors that influence variation in abundance

Introduction

Data: $y(s_i) \equiv y_i$ are spatially referenced *counts*, e.g., number of birds counted at site s_i (a point, quadrat, transect)

Genesis of Spatial Dependence –

- Omitted habitat covariates
- Demographic processes

  → Recruitment, dispersal, etc..
- Interactions between individuals/species

  → Predation, competition

Objectives

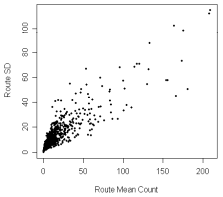
What do we do with spatial models of abundance?

- Mapping/prediction or simple description
- Small area estimation, inference
- Shrinkage estimation of model parameters
- "Honest" estimation of covariate effects

Considerations for Modeling Counts

Why not just use a kriging-type model?

- counts are positive valued
- counts are discrete
- mean related to variance (empirically)



← Route SD vs. mean, house finch (routes ≥ 10 years)

Kriging is a linear procedure, for normally distributed data that does not respect these features.

Generalized Linear Models (GLMs):

Classical statistics deals with normal distributions and linear models.

- $y_i \sim \text{Normal}(\mu_i, \sigma^2)$
- $\mu_i = \beta_0 + \beta_1 x_i$

Kriging is also a normal, linear procedure

GLMs (Generalized Linear Models) represent an analogous class of models for non-normal data

Elements of Generalized Linear Models (GLMs)

A probability model for the observations:

- $f(\mu_i, \theta)$
  - $\mu_i = E[y_i]$
  - $\theta =$ a variance parameter

Common choices of f for count data

- Poisson
- Binomial

Generalized Linear Models (GLMs)

Modeling covariates effects:

$$h(E[y_i]) = \sum_{j=1}^{J} \beta_j x_{ij}$$

instead of (for normal data)

$$E[y_i] = \sum_{j=1}^{J} \beta_j x_{ij}$$

- $h(\cdot)$ is called the *link* function (it *links* the mean of $f(\cdot)$ to the linear function of covariates)
  - Poisson: $log(\mu_i)$
  - Binomial: $log(\mu_i/(1-\mu_i))$

Poisson Regression

Probability model for the data:

$y_i \sim \text{Poisson}(\mu_i)$

μ_i is the mean of y_i at location s_i

$log(\mu_i) = \beta_0 + \beta_1 x_i$

x_i = a covariate, describing landscape or habitat structure

GLMs for Spatial Data

Introduce a spatially indexed random effect, z_i:

$$h(\mu_i) = \sum_{j=1}^{J} \beta_j x_{ij} + z_i$$

- z_i is a spatially correlated random effect
- Exploit conventional Gaussian spatial process models for z_i (kriging)
- Several possibilities are described shortly

Binomial counts

If y is the number of "successes" in T independent Bernoulli trials ("coin flips"), then y has a binomial distribution

- T = sample size
- parameter π = "success probability"

Binomial data examples

- Nest success/productivity data
- Capture-recapture or band recovery data
- Occupancy data (y_i units occupied out of T_i)
- Harvest success

Binomial counts

Goal: model variation in π_i

Logistic regression model:

$$log(\pi_i/(1-\pi_i)) = \sum_{j=1}^{J} \beta_j x_{ij} + z_i$$

Poisson Counts

Aggregate a Poisson point process (equal area units)

$$y_i \sim \text{Poisson}(\mu_i)$$

y_i results from counting (unique) individuals in space

Goal: model variation in μ_i

Log-linear model:

$$log(\mu_i) = \sum_{j=1}^{J} \beta_j x_{ij} + z_i$$

Spatial Models for z —

Assume that $z_i \equiv z(s_i)$ is a Gaussian spatial process:

- $z_i \sim$ Normal
- $E[z_i] = 0$
- $Var[z_i] = \sigma^2$
- $Corr(z_i, z_j) = k_\theta(||s_i - s_j||)$

Joint normality of $\mathbf{z} = (z_1, z_2, \ldots, z_n)$:

$$\mathbf{z}_{n\times 1} \sim \text{Normal}(0, \mathbf{\Sigma}(\theta))$$

There are a number of ways to specify $\mathbf{\Sigma}(\theta)$

1. Classical or Direct Construction

"Kriging for counts" – A direct specification of a joint distribution for the spatial process, $z(s)$

Specify a model for the correlation between $z(s)$ at any two locations:

$$Corr(z(s_i), z(s_j)) = k_\theta(||s_i - s_j||)$$

e.g., exponential decay –

$$k_\theta(s, s') = e^{-||s-s'||/\theta}$$

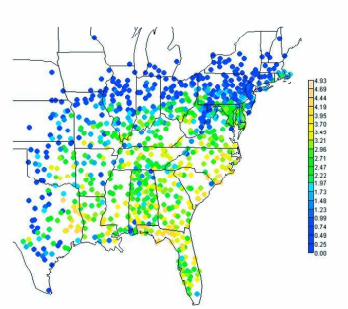
This function $k_\theta(s, s')$ "fills-in" the $n \times n$ elements of $\mathbf{\Sigma}(\theta)$:

$$\mathbf{z}_{n\times 1} \sim \text{Normal}(0, \mathbf{\Sigma}(\theta))$$

Estimation/prediction requires repeated mathematical operations on $\mathbf{\Sigma}(\theta)$

Example: Range Mapping

- Carolina Wren counts from the BBS
- abt. 1000 routes
- Goal is to make a relative abundance/range map



← log(mean) route count (across years)

$\mathbf{\Sigma}(\theta)$ is 1000×1000 and does not yield to kriging-like estimation and prediction.

Kriging for Counts

Diggle, P.J., J.A. Tawn and R.A. Moyeed. 1998. Model-based geostatistics. *Journal of the Royal Statistical Society, Ser. C.*

2. Kernel Smoothing/(Process Convolution) Construction

Express $z(s)$ as a linear combination of *iid* "random effects"

$$z(s) = \sum_{j=1}^{R} w_\theta(r, s)\alpha(r_j)$$

where

$$\alpha(r) \sim \text{Normal}(0, \sigma_\alpha^2)$$

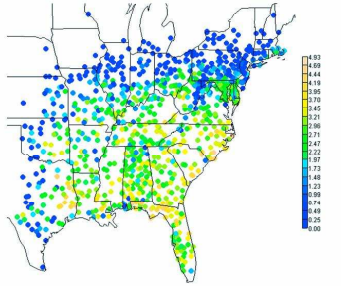
- $w_\theta(r, \cdot)$ is a kernel centered at r
  "kernel" = weighting function
- z an average of "noise" –
  $z(s)$ is a weighted average of *iid* noise $\alpha(r_j); j = 1, 2, \ldots, R$.
- A classical mixed model (Laird and Ware; PROC MIXED)
- $R << n$

Kernel Smoothing/Convolution Construction

- Equivalence between this method and "kriging", i.e., a precise relationship between the choice of $w_\theta(\cdot)$ and the correlation function.
- This is more computationally efficient in large problems. Do not have to operate on $\mathbf{\Sigma}(\theta)_{n\times n}$.
- Higon, D. 1998. A process-convolution approach to modeling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*

Example: Range Mapping

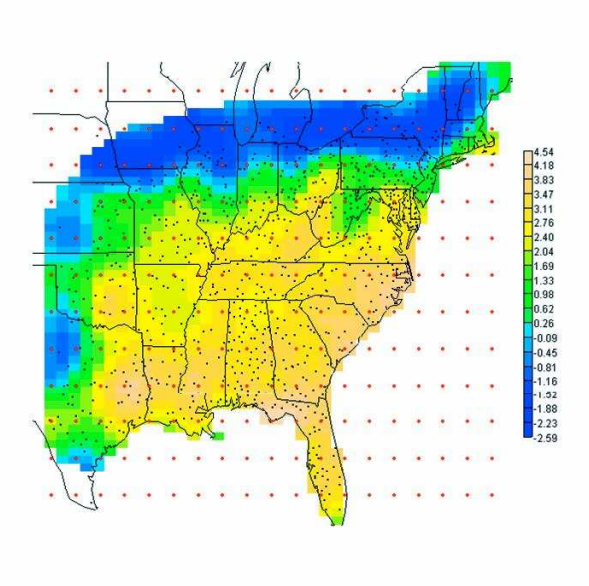
- Carolina Wren counts from the BBS
- abt. 1000 routes
- Goal is to make a relative abundance/range map
- Method: Gaussian kernel convolution model



← log(mean) route count (across years)

← Data locations and grid of "support points" – Gaussian kernel centers

Estimated spatial process:



3. Lattice models

Usually used when *data* have discrete or areal support. e.g., areal measurements: counties, geographic strata, etc..

Conditional autoregression (CAR):

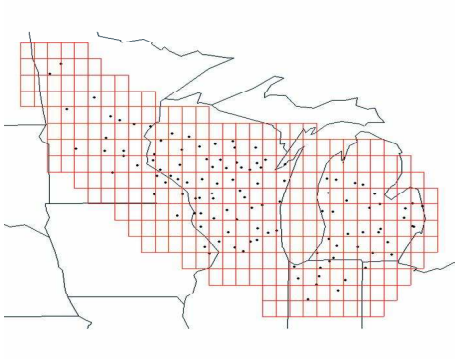
$$z_i = \rho \sum_{j \sim i} w_{ij} z_j + \epsilon_i$$

$\{w_{ij}\} \equiv \mathbf{W}$ is the *adjacency* matrix.

- 0s and 1s indicating neighbors
- length of boundary
- "average distance" between cells

Lattice models for non-lattice data

If data locations do not form a natural lattice, then make one up:



$$log(\boldsymbol{\mu}) = \mu\mathbf{1} + \mathbf{Hz}$$

- $\boldsymbol{\mu}$ is $n \times 1$
- $\mathbf{z}$ is $p \times 1$ CAR process
- $\mathbf{H}$ is $n \times p$

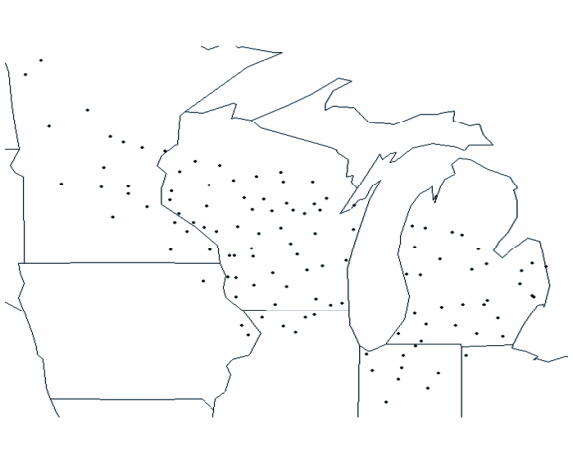
$\mathbf{H}$ associates each observation with one or more of the p random effects, which are arranged on a lattice

BBS Bobolink counts, arbitrary grid for embedded CAR model

Example: Spatial Variation in Bobolink Counts

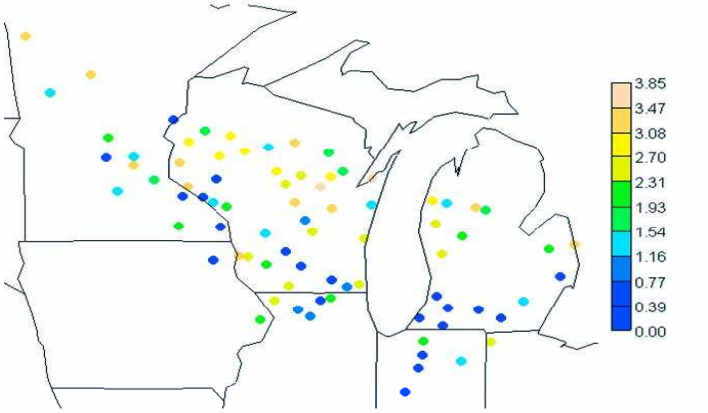
- Species: Bobolink
- BBS route counts in the upper-midwest (a physiographic stratum)
- Several habitat covariates thought to influence abundance
- CAR model with incidence adjacency matrix

Data Locations



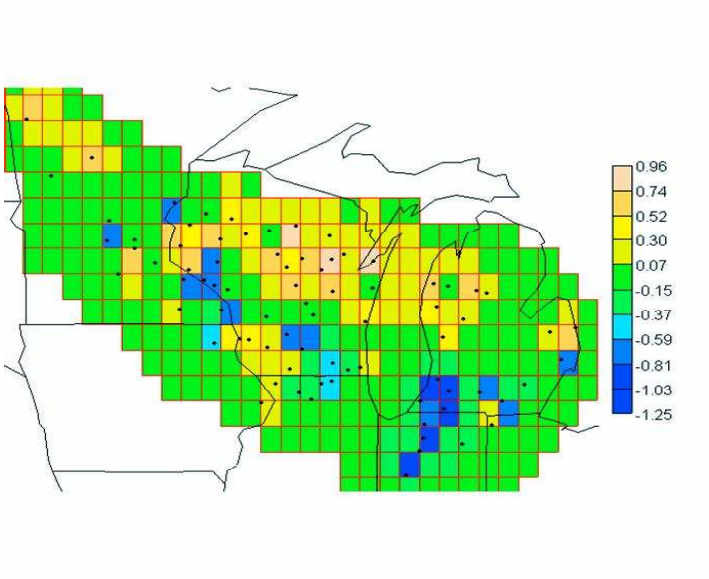
100 or so routes in upper midwest
y_i = count of bobolinks on BBS route i, located at s_i.

Data



$log(count)$

Predictions



Estimation and Implementation

- Markov chain Monte Carlo

geoR,geoRGLM add-on **R** libraries

PROC MIXED/GLIMMIX for some models

WinBUGS for all models described here

Abundance and Detectability

In Ecology, we have an acute inability to observe the state variable of interest in many problems: Abundance, or occurrence

$N(s) = \#$ of animals in population s (population size)

Observe a sample count, $y(s) \leq N(s)$

Abundance and Detectability

Binomial Observation Model:

$$y(s) \sim \text{Binomial}(N(s), p)$$

$$\begin{aligned} y(s) &= \text{observed count} \\ p &= \text{“detection probability”} \end{aligned}$$

- Detection is important because y is a “biased estimate” of N
- p can vary in response to many factors (e.g., intensity, env. conditions)
- Variation in y is not just due to variation in N.
- But (variation in) N is the object of inference

Simple Count Surveys (Binomial counts)

When detection is imperfect, $N(s)$ is not distinguishable from p (they are confounded). For example, the model consisting of:

(1) $y(s) \sim \text{Binomial}(N(s), p)$ and

(2) $N(s) \sim \text{Poisson}(\mu(s))$

is equivalent to the model

$$y(s) \sim \text{Poisson}(p\mu(s))$$

Thus, models for $y(s)$ describe variation in the product $p\mu(s)$. This is insufficient for some important inference problems.

Abundance and Detection

Therefore, much effort has been directed toward developing alternative sampling protocols/methods that allow variation due to the detection process to be decoupled from variation in abundance.

- capture-recapture
- double or multiple observer sampling
- distance sampling
- "removal" methods

Most methods yield a multivariate count statistic $\mathbf{y}$ that has a multinomial sampling distribution –

$$\mathbf{y}|N \sim \text{Multinomial}(N; \boldsymbol{\pi})$$

Differences among protocols are manifest in parameterization of $\boldsymbol{\pi}$

Example of Multinomial Observation Models

A double-observer protocol: Two observers independent record observations of individuals and, after the fact, "reconcile" their observation lists. This yields an *encounter history* for each individual of the form:

1	1	observed by both observers
1	0	observed by 1st
0	1	observed by 2nd
0	0	not observed

Data are encounter history *frequencies* – n_{11}, n_{10}, n_{01} and n_{00} (missing data), which have a multinomial distribution, with cell probabilities $\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}$. These are functions of detection probability p_1 (1st observer) and p_2 (2nd observer).

The General Hierarchical Model

1. Multinomial Likelihood –

$$\mathbf{y}|N \sim \text{Multinomial}(N; \boldsymbol{\pi})$$

2. Abundance model –

$$N_i \sim Poisson(\mu_i)$$

3. Model for the Poisson mean

$$log(\mu_i) = \mathbf{x}_i'\boldsymbol{b} + z(s_i)$$

4. The spatial process – Spatial dependence is induced through the correlated random effect, $z(s)$.

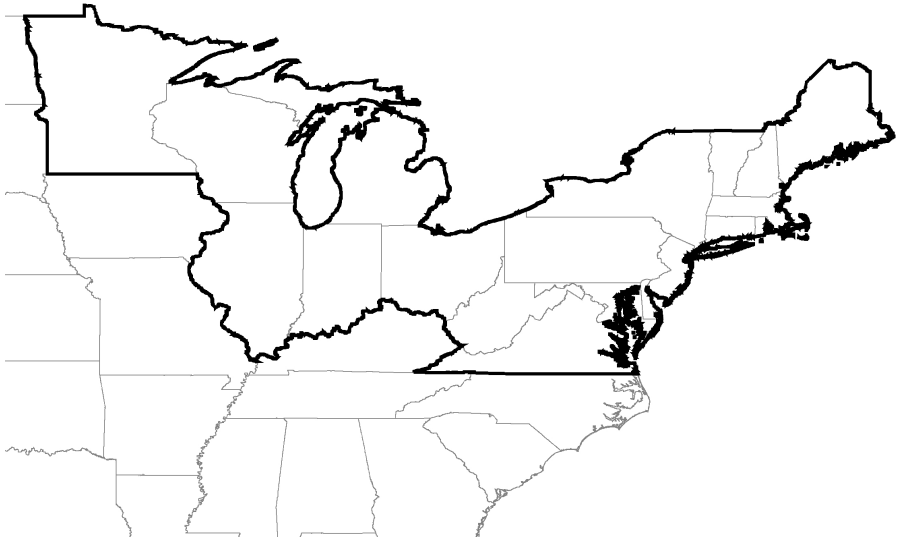
Summary

- Many ecological studies yield data that are counts: of animals, or Bernoulli trials
- Poisson/Binomial GLMs with spatially correlated random effects
  1. Kriging-type models
  2. Regression-on-noise ("convolution") formulation
  3. Lattice models (CAR)
- Abundance/occurrence processes, detection bias: yields a hierarchical model wherein the spatial model governs the latent (unobservable) abundance parameter, $N(s)$.

A hierarchical spatial count model with application to American Woodcock

Wayne Thogmartin, USGS Upper Midwest Environmental Sciences Center

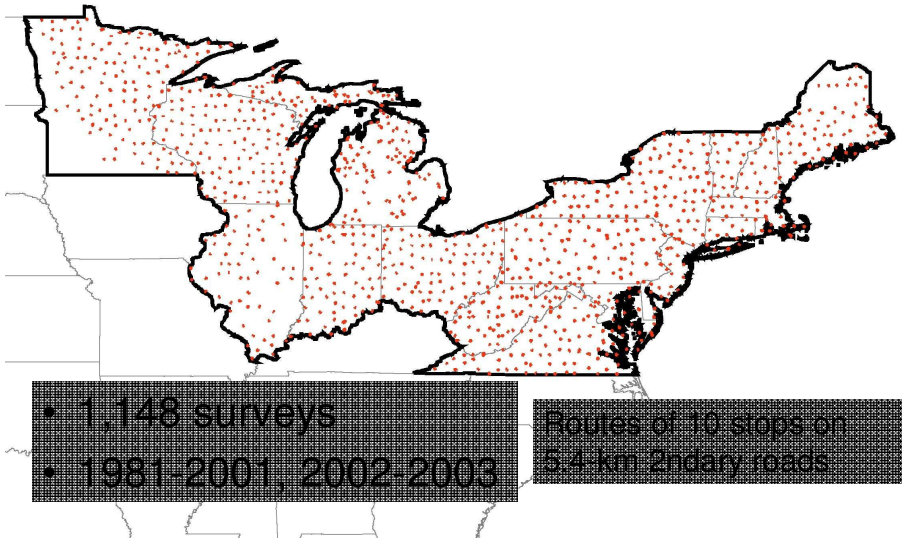
American Woodcock primary breeding range in the United States



Objective

- Our objective is to model and map predicted woodcock relative abundance across their primary breeding range in the United States.

American Woodcock Singing Ground Surveys



Survey Design

- Woodcock "peenting" surveys are annually conducted on secondary roads in the upper midwestern and northeastern United States.

Summary Statistics - Woodcock

- Mean count for 9,142 surveys (space x time) was 3.39 birds per survey (SD = 4.00)
- Zero counts comprised 27% of the surveys
- Median count was 2, and maximum count was 47
- 1,581 observers

Spatial Poisson Count Model

$$Z(s_i) = \mu(s_i) + \Sigma c_{ik}[Z(s_k) - \mu(s_k)] + \omega(s_i) + \gamma(s_i) + \varepsilon(s_i)$$

ω Observer effects: *observers count birds differently*

γ Year effects: *to accommodate observed decline in abundance*

μ Environmental effects

ε Extra-poisson variation

Spatial CAR (Conditional AutoRegression)

Spatial Poisson Count Model

- The expectation is treated as Poisson.
- Because observers count birds differently (e.g., older birders have a hard time hearing some species, novice birders have a hard time recognizing birds with unusual calls), we wish to adjust the counts to offset the effect of observer.
- We are using a time series of counts from a number of surveys. We can leverage this time series to inform our association of counts with habitat IF we control for annual variability and any sort of trends that may occur in the data (many birds are declining in abundance, and so it would be 'unfair' to compare counts from 1981 with those from 2001 if the species is in the midst of a decline).
- Environmental factors are included as a linear combination of variables derived from classified satellite imagery. These environmental factors will form the primary basis for mapping the predicted species abundance.
- Typically, the variance of counts exceeds the mean of those counts, so we have a term to soak up that extra-Poisson variation. This is generally not a serious issue as much of the extra-Poisson variation is 'structural' in nature, i.e., because of observers, routes, or years consistently leading to lower or higher counts than may be expected. This is adjusted for through hierarchical modeling (described shortly).
- We expect counts to be correlated over space, and so we model this correlation with a spatial 1st-order conditional autoregression.

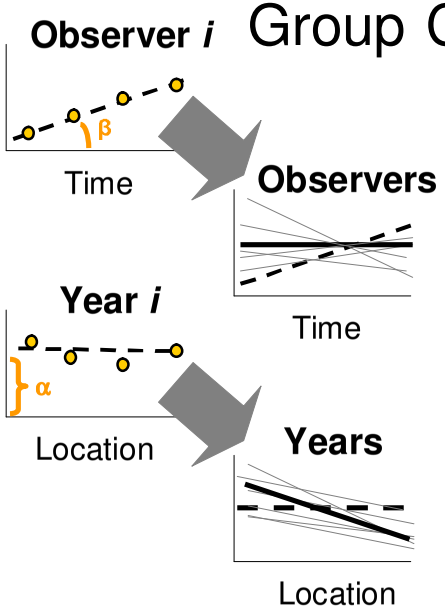
Hierarchical Modeling

- Bayesian: Data and prior specification used to identify a posterior distribution for parameter estimates (β)
  - Standardized Likelihood x Data = Posterior Probability
- Hierarchical: clustering of β for observer, year, and route effects because of group correlation
- Correlation may occur because of design, over time, and/or across **space**

Model Fitting

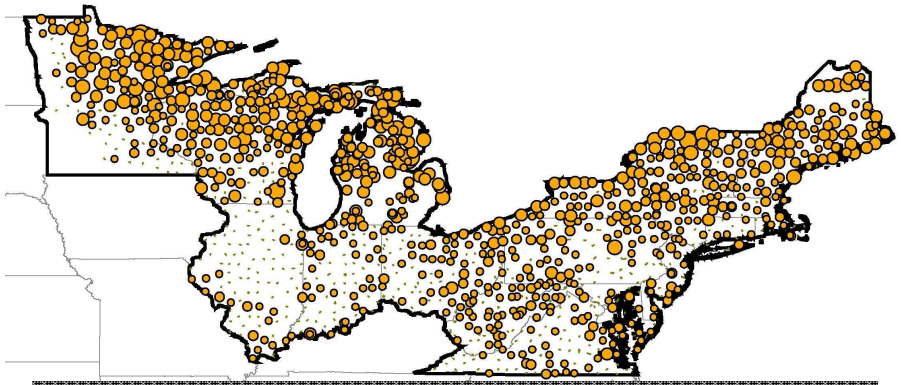
- To fit this model, the approach we employ is described as a hierarchical model. In this workshop, we're most interested in effects over space, but to identify those spatial effects free of the clutter of the sample design and the temporal correlation between survey's, we need to accommodate nuisance effects that would otherwise obfuscate the spatial effects.

Group Correlation



- Observer variability: Some observers will count woodcock similarly
- Temporal variability: Some years may be above the grand mean, others below

Mean Counts: Spatial Considerations

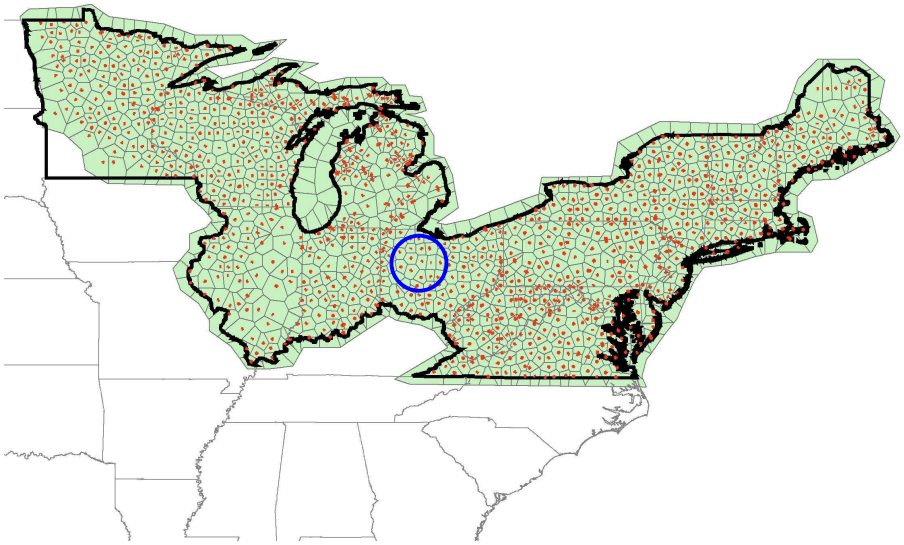


- Tobler's Law: We expect that which is closer to be more similar than that farther apart

Observed Counts

- An average of the point-specific time series shows that there is a general north-south gradient in woodcock abundance. Woodcock are more abundant in the north and less abundant in the south.
- This gradient results in sites near to one another being more similar than those farther from one another.
- We may wish to accommodate this spatial correlation to reduce the bias imposed on estimation of the slopes associated with the environmental factors. This correlation ostensibly describes environmental factors for which we have insufficient ability to map (i.e., understory plant composition, earthworm abundance, etc.).

Spatial Correlation: Lattice-based Solution



Irregular Lattice

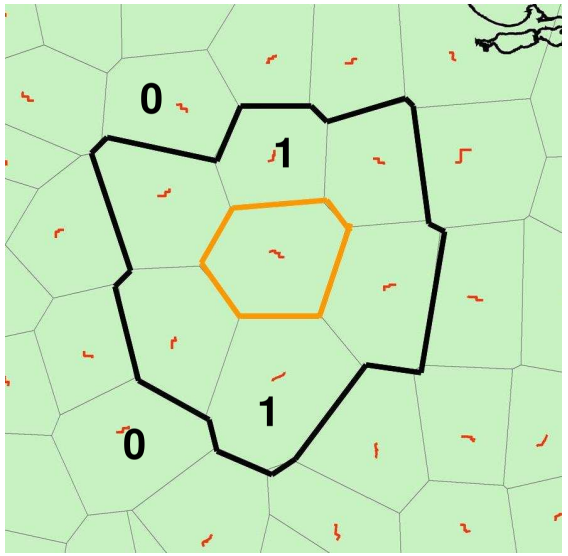
- We identify the domain of interest or influence around each route by tesselating the routes, creating an irregular lattice. This irregular lattice will be used to identify the neighborhood structure.

Neighborhood

- 1st order Conditional Autoregression

Value of *i* is akin to a weighted 'average' of surrounding cells

Surrounding cells weighted 1, distant cells weighted 0



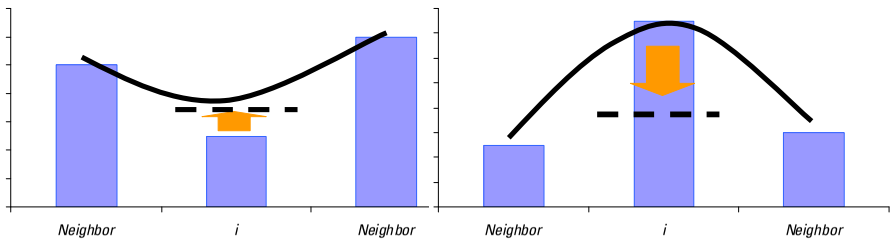
Shrinkage

- Observed counts will be more variable than the mean expectation. Shrinkage or smoothing gives a stable estimate of the pattern of the underlying expected counts, whereas the raw counts lead to a noisy or blurred picture of the true, unobserved count process.

Conditional Autoregression

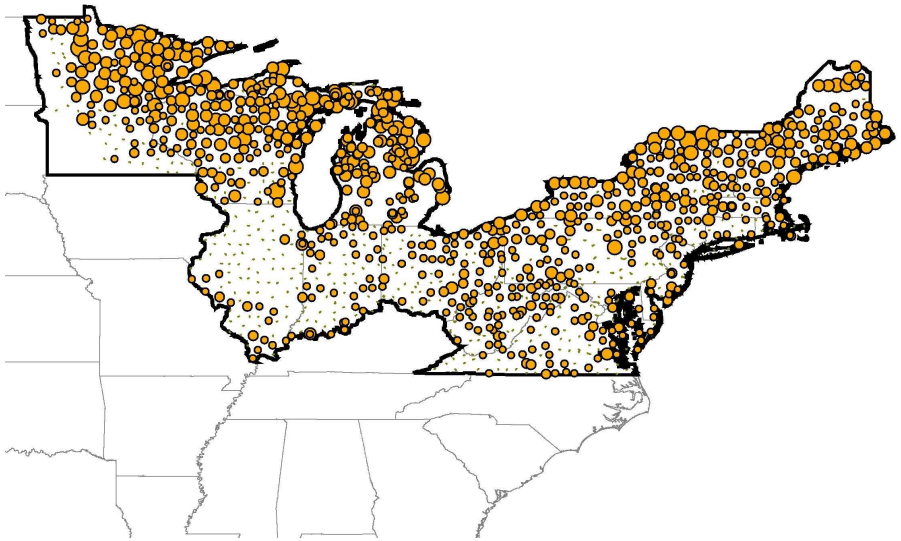
- Probability of observing a particular value at a given site is a conditional probability, i.e., it depends upon the values in the surrounding neighborhood
- Advantages:
  - Conservative
  - High Specificity (correctly classifying occurrences) even in sparse data situations

Smoothing

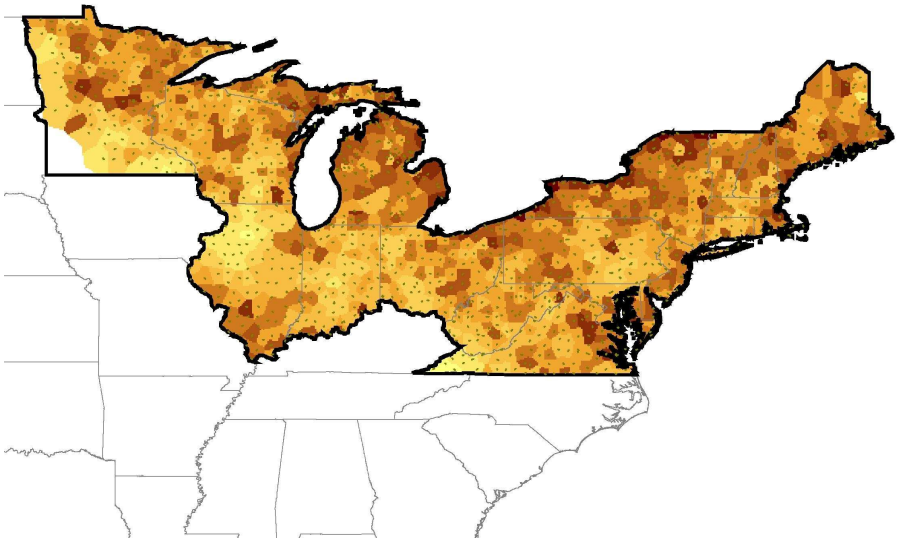


Shrinkage provides a stable estimate of the pattern of the underlying expected counts

Mean Counts



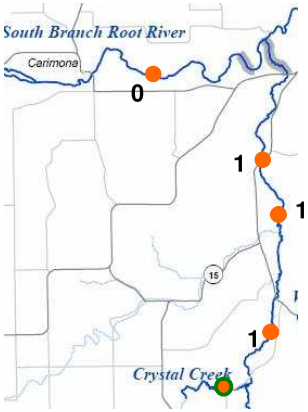
Smoothed Expectation



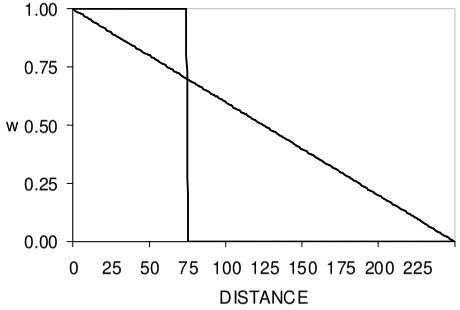
Alternatives to 1° CAR

- Ad hoc weighting

$\omega = 1$ (1° & 2 °), 0 (>2 °)



Euclidean Distance-based weights



Weights determined by degree of interaction or similarity

Alternatives to 1° CAR

- We may wish to include not just our nearest neighbors, but also those neighbors in the surrounding ring immediately beyond the nearest neighbors; this would be a 2nd-order CAR.
- We also might want to use proximity as defined by a metric other than Euclidean distance. For instance, maybe only those points along a stream or road are considered part of the neighborhood and given a weight of 1, whereas all others are given a weight of 0.
- Others have used distance-based weightings, after having done semivariogram analyses to identify the degree of spatial correlation. These distance weightings can be 1 for all points within a certain distance (the range in geostatistical parlance) and 0 otherwise, or the 0-1 gradient can be continuous and reflect the distance from the point in some linear fashion.
- Regardless, symmetry needs to be observed. That is, if you are my neighbor, I am your neighbor.

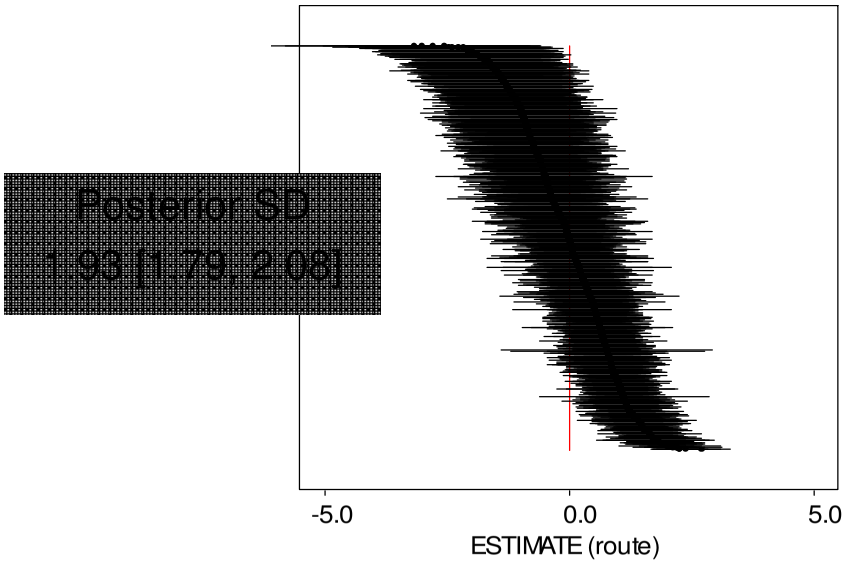
Parameter Estimates for μ for Models at 3 Spatial Scales

Variable	Finest Scale (350 ha)	Medium Scale (4,000 ha)	Coarsest Scale (106,000 ha)
INTERCEPT	**0.02 (0.10)**	**0.07 (0.11)**	**0.06** (0.15)
START OF SEASON	-0.37 **(0.17)**	**-0.33 (0.12)**	**-0.33** (0.16)
AGGREGATION INDEX	**-0.29 (0.04)**	-0.36 **(0.05)**	**-0.26** (0.07)
HUMAN (%)	**-0.22 (0.04)**	-0.26 **(0.04)**	**-0.15** (0.05)
GRASS (%)	**-0.01 (0.05)**	-0.21 **(0.05)**	**-0.14** (0.07)
ASPEN (%)	**0.09 (0.04)**	**0.12 (0.05)**	0.20 **(0.08)**
TOPOCONVERGENCE	0.10 **(0.04)**	**0.00 (0.05)**	**NA**
SHRUB (%)	0.17 **(0.11)**	**0.17 (0.11)**	**0.12** (0.14)
FOREST (%)	0.18 **(0.05)**	**0.15 (0.05)**	**0.09** (0.05)
FOREST×FOREST(%)	**NA**	-0.01 **(0.05)**	**NA**

Environmental Factors

- We modeled woodcock at three spatial scales, and used an information-theoretic approach to averaging models within scale. We found little variability in woodcock response to the environment across scales. Woodcock were generally negatively related to the day of the year in which the growing season began, which may reflect the importance of earthworms to the woodcock diet. Woodock were also negatively related to landscapes in which forest, shrub, and field were aggregated into clumps as opposed to fine distributions among each other. The relation of forest and aspen were positive, but the importance of forest declined with the coarsening of scale, whereas the importance of aspen increased as the scale coarsened.

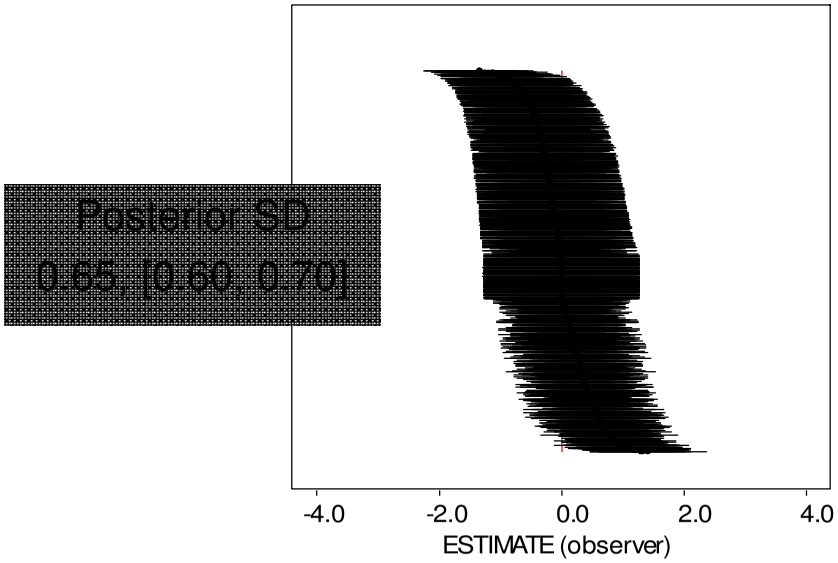
Route Random Effect



Route Random Effect

- A caterpillar plot of the individual route effects, ordered by route estimate, indicates a small number of routes reduce the expected counts relative to the predictions of the environmental variables, whereas a number of routes increase the expected counts relative to the predictions. These route-level reductions and increases are variability that we can not explain with the environmental variables we have identified in the course of our model.

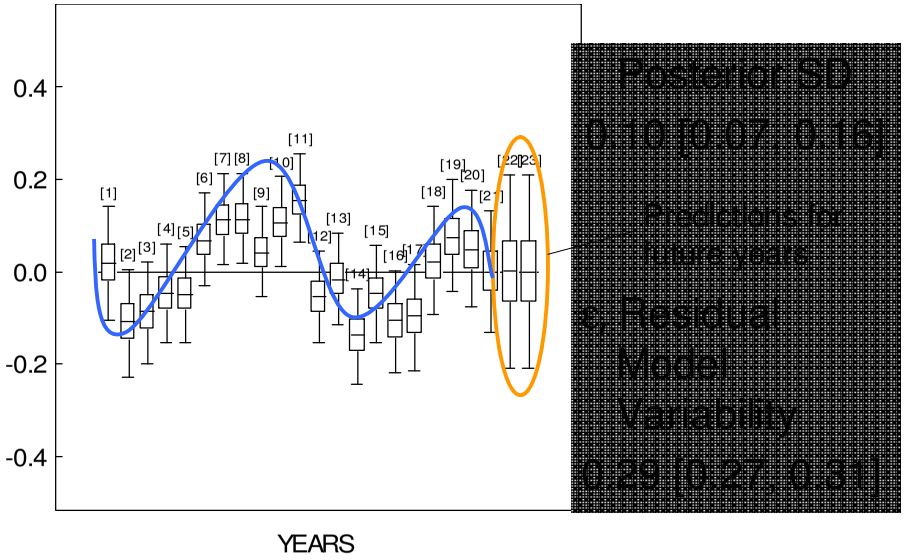
Observer Random Effect



Observer Random Effect

- Aside from a small number of observers who under- or over-counted relative to the other observers, most observers had little effect on the overall count expectation, indicating that we should have little concern in general for the effect of observers on surveys of woodcock.

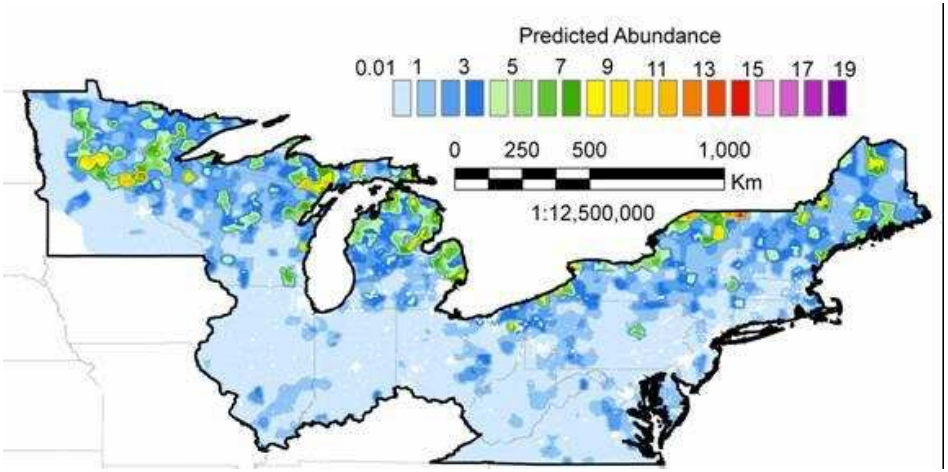
Year Random Effect



Year Random Effect

- May want to address the potential cyclicity with an AR(1) (i.e., an autoregressive term of lag 1); this may reduce the error variance around the out-years (2002 and 2003).

Predicted Woodcock Relative Abundance

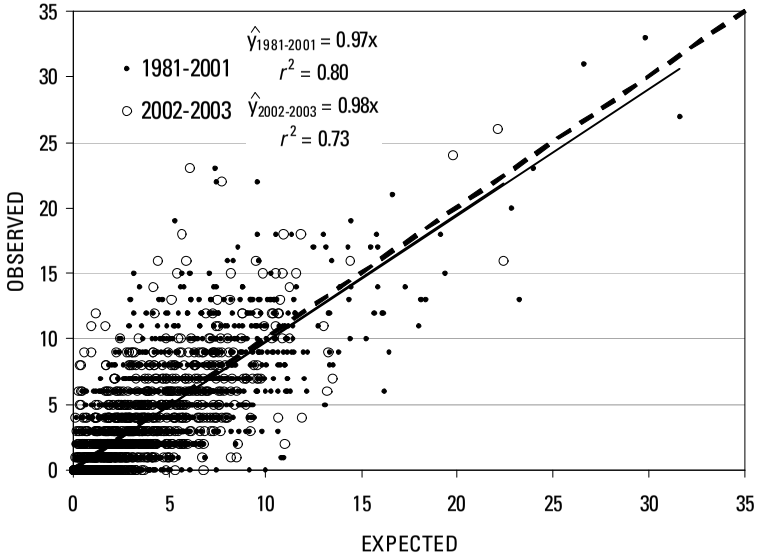


circa 1991

Mapped Predictions

- The result of mapping the environmental variables and the route effect together yields a map of predicted woodcock relative abundance.
- Because we treat the counts as Poisson, we must first exponentiate the linear combination of variables and slope estimates to map the count expectation. $e^{(\beta 1 * X1 + \ldots + \beta k * Xk + \text{route effect } Z)}$

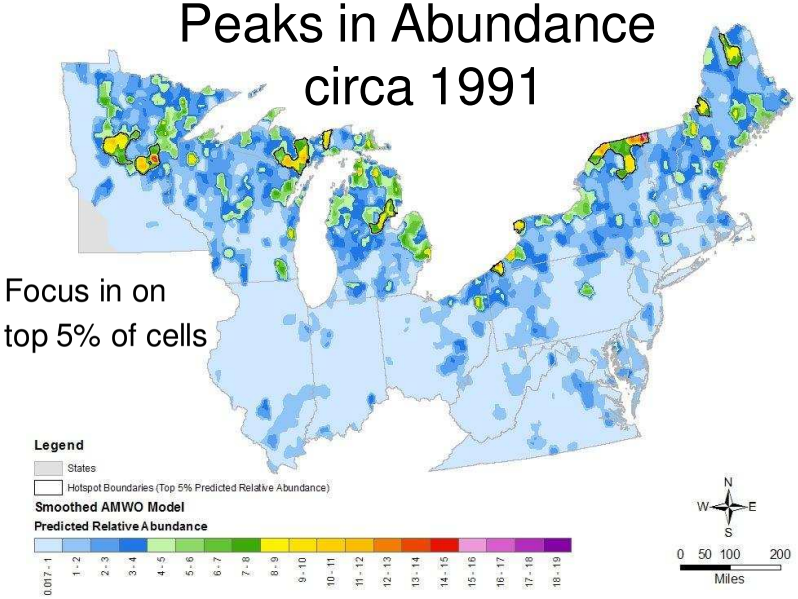
Model Evaluation



Model Evaluation

- Evaluation of the model by imputing values as determined by the final model structure (i.e., based upon the estimated model parameters [slopes]) indicated near one-to-one correspondence between the model predictions and the observed data for both those data withheld from model construction and data for the two years subsequent to the modeling effort.

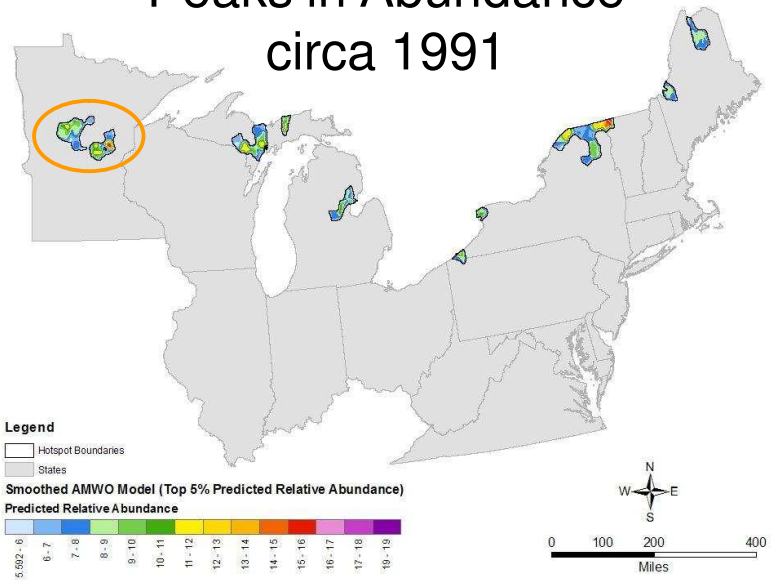
Predicted Woodcock Peaks in Abundance circa 1991



Management Application

- To increase the efficiency of conservation delivery, it would be best to manage the species where our efforts would do the most good for the most individuals of the population. Unless we are led to believe otherwise, that efficiency comes by conserving the species where it is most abundant. Thus, we use our map of predicted abundance to focus on specific areas of high or peak abundance.

Predicted Woodcock Peaks in Abundance circa 1991



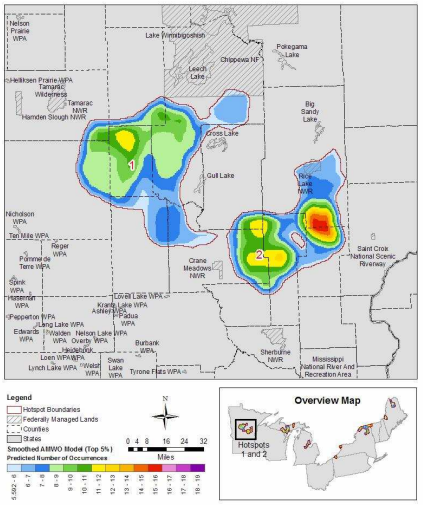
Management Application

- There are 10 such areas. These areas are the top 5% of the distribution in the expected counts.

Regional Conservation Planning

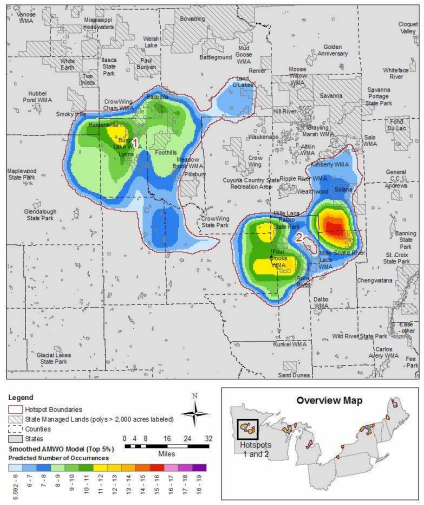
Smoothed American Woodcock (AMWO) Model Overlayed with Federal Lands

Hotspots 1 and 2



Smoothed American Woodcock (AMWO) Model Overlayed with State Lands

Hotspots 1 and 2



Management Application

- We found, from our analysis of the mapped predictions relative to the state and federal land management agencies (i.e., "the conservation estate"), that the proportion of the population occurring on private lands varied between 70.5% in Minnesota to 94.1% in Maine, with a grand mean of 79.9%. The proportion of the predicted population was 7.2% on federal land and 12.9% on state land, which was marginally higher than the proportion of the area under federal and state management (6.4% and 11.4%, respectively).
- We plotted our predictions against data layers describing the land management context with the idea that land managers and private lands biologists can effectively direct species-specific conservation efforts to those specific areas where the species is high in abundance. We can also use these sorts of maps to direct research activities, to better learn why species in these areas are highly abundant. We may also be able to use constituent aspects of the model to identify areas where the species can be most effectively increased by simple modification of the landscape (i.e., if we affect certain management practices in areas where the species occurs, might we see better bang for our buck in some areas rather than other areas; are there limiting factors that we can not overcome regardless of our management efforts [e.g., climate (start of the growing season) can not be managed, but only accommodated]).

Questions?

- For more information: http://www.umesc.usgs.gov/terrestrial/migratory_birds/bird_conservation/amwo_american_woodcock.html
- wthogmartin@usgs.gov

Predicting the Spread of Invasive Species

Mevin B. Hooten

Department of Statistics
University of Missouri

March 15-16, 2006

Title Page [notes]

- The basic idea here involves a method for incorporating a scientifically meaningful deterministic model in a more general probabilistic framework for estimation and prediction while accounting for uncertainty at multiple levels.

Acknowledgements

- Christopher Wikle
- Robert Dorazio
- J. Andrew Royle

Acknowledgements [notes]

Various components of this work can be found in:

- Wikle, C.K., (2003). Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology , 84, 1382-1394
- Royle, J.A., and C.K. Wikle, (2005). Efficient Statistical Mapping of Avian Count Data. Ecological and Environmental Statistics , 12, 225-243.
- Royle, J.A. and R. Dorazio, (2006). Hierarchical models of animal abundance and occurrence. In Review.
- Hooten, M.B., C.K. Wikle, R.M. Dorazio, and J.A. Royle (2006). Hierarchical matrix models for characterizing invasions. In Review.

Characteristics of Invasive Species

- Invasive: quickly spreads and becomes abundant.
- Can be naturally introduced or imported.
- Successful Invasions:

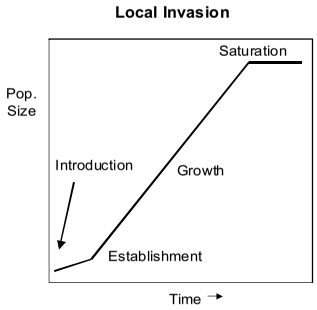
Invasive Species: phases of a successful invasion

1.) Introduction

2.) Establishment

3.) Dispersal / Growth

4.) Saturation

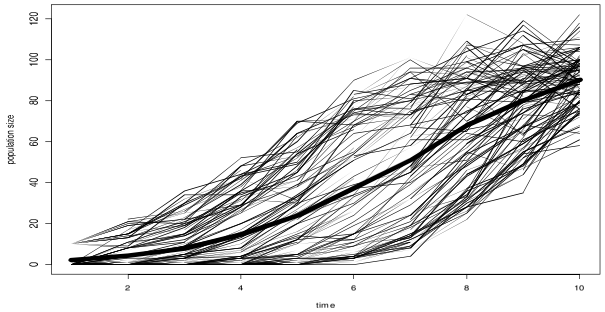


Characteristics of Invasive Species [notes]

- The idea with this figure is that it represents growth in the population size for an organism over time.
- The growth curve shown is very generalized, of course there are all manner of more complex forms of population growth. The basic idea is that after introduction, population size grows rapidly until resources become limiting.
- As the population size approaches the carrying capacity (i.e., saturation) other forms of dynamical behavior could ensue (e.g., stability, periodicity, chaos).

Characteristics of Invasive Species (cont'd)

- Multiple growth curves for various locations:



Characteristics of Invasive Species (cont'd) [notes]

- Studying total population size is useful, but we want to make inference about the population size at numerous locations over time.
- These plots with multiple growth curves representing the growth in population size at each location of interest are informative, but it's difficult to see the interaction between locations (that is, the movement of organisms between locations).
- A sequence of maps is helpful here, such as those in the results section of this presentation.

Characteristics of Invasive Species (cont'd)

Impacts of exotic species:

- Pests can attack humans and livestock (e.g., Killer Bee).
- Cause or transmit disease (e.g., West Nile Virus and Avian Flu).
- Disrupt native food webs (e.g., Peacock Bass and the exotic zooplankton, *Daphnia lumholtzi*.

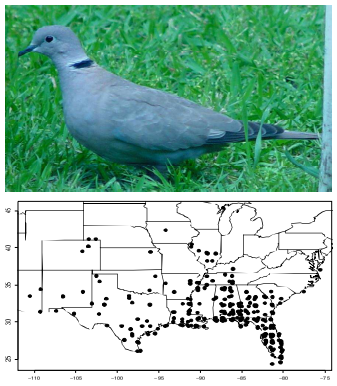
Characteristics of Invasive Species (cont'd) [notes]

- Obviously these are some of the more prominent examples in the media.
- The point of this slide is to provide some justification for wanting to study these processes in more detail in order to better understand them and thus make better management decisions.

History

Eurasian Collared-Dove (ECD):

- Invaded Europe in 1930's.
- Introduced to Florida mid-1980's.
- Count data collected through N. Amer. Breeding Bird Survey, documenting invasion.
- Imperfect detection.



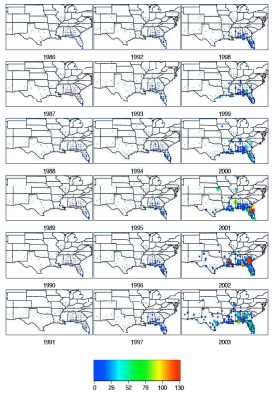
History [notes]

There are several good references for this species:

- Hengeveld, R. (1993) What to do about the North American invasion by the Collared-Dove? Journal of Field Ornithology 64:477-489.
- Romagosa, C., and R. Labisky. (2000) Establishment and dispersal of the Eurasian Collared-Dove in Florida. Journal of Field Ornithology 71:159-166.

Characteristics

- No replicate data through BBS.
- Separate dataset used to estimate detection probability.
- Data for years: 1986-2003



Characteristics [notes]

- We desire to estimate the “true” population size, that is, the real number of birds in a given location at a given time.
- Our data represents only the “observed” number of birds. In this type of data collection we could miss a few birds even though they were there.
- Treating the probability of missing a bird as a parameter in our model, we would need more than one observation to estimate that parameter as well as the “true” population size.
- In the case where we only have the one space-time observation (as with the BBS data) we must estimate the probability of detection separately.

Impacts

ECD biological threats (Romagosa and Labisky 2000):

- Competition for resources with native avifauna.
- Transmission of disease.

“ECD will probably colonize all of North America within a few decades.”

- Just how probable is it?

Impacts [notes]

- One of the most important game animals in this country is the mourning dove.
- It would not be good if the ECD causes problems for the mourning dove.
- The goal here is to associate (determine) some level of probability with the ongoing invasion at various locations and times.

Overview

Environmental/Ecological Sciences

- Common classes of scientifically meaningful behavior.
- Often with non-linear and spatially varying dynamics.
- A hierarchical modeling framework can be employed to accommodate such behavior.

Overview [notes]

- The question is: How do we make use of all of this scientific knowledge while characterizing complex dynamics in a rigorous statistical model?
- A hierarchical framework allows us to characterize very complex systems by breaking the problem down into simpler and more intuitive components.
- It also allows us to incorporate scientific knowledge (e.g., functional model forms and parameter spaces) into the model.

Spatio-temporal Processes

Examples:

- Diffusion: Spreading process; similar to “dispersal” in ecology.
- Growth: Process increasing in intensity; a simple form of population growth in ecology.
- Density Dependent Growth: Process increasing in intensity non-linearly; a more realistic form of population growth.

Spatio-temporal Processes [notes]

- These processes, when formulated mathematically, can be written as Partial Differential Equations (in continous space and time) or difference equations (in discrete space and time).
- Difference equations can be derived as approximations to partial differential equations.
- There are many other deterministic models capable of exhibiting dynamical behavior (as discussed in the methods).

Examples

- Diffusion: spreading

diffusion movie...

Examples [notes]

- The movies that are shown on the current and following slides are only example simulations of these types of processes to give you some idea of what they might look like.
- Alone they appear quite simple, but in combination they can represent more realistic invasive behavior.
- You will see from the movie in the results section, that the behavior is a combination of both diffusion and non-linear growth.

Examples

- Growth: increasing in intensity

growth movie...

Examples [notes]

- This is a linear form of growth where the population size increases at a constant rate in a given area.
- The differences are subtle between this movie and the next but very important for exhibiting realistic behavior.
- In these two growth movies, the process is not spreading out (diffusing), but rather growing independently at each location.

Examples

- Density Dependent Growth: non-linear increase in intensity

non-linear growth movie...

Examples [notes]

- Again, the differences in types of growth are subtle here.
- In this movie, the growth rate slows down as a function of intensity in the process (population size) and after reaching a carrying capacity it ceases to grow further.

Definitions and Notation

- "Population": loosely, the true number (N) of organisms at a place and time (also let $\lambda =$ mean population $= E(N)$).
- "Count": the observed number (n) of organisms at a place and time where $n \leq N$.
- Bolded variables denote vectors and matrices (e.g., $\mathbf{x} = [x_1, \ldots, x_m]'$).
- "|" = given; as in conditional probability.
- Square bracket notation refers to a probability distribution (e.g., $[x|\beta] = \text{Prob}(x|\beta) = f_x(\beta)$).
- "$\sim$" = is distributed as ... (e.g., $x|\beta \sim [x|\beta]$).
- "$\propto$" = is proportional to ... (e.g., $[\beta|x] \propto [x|\beta][\beta]$).

Definitions and Notation [notes]

- Another way of saying: $x|\beta \sim f_x(\beta)$ is that x is a sample from the probability distribution f given the parameter β.
- These kind of expressions: $[\beta|x] \propto [x|\beta][\beta]$ will be used later to illustrate the hierarchical nature of the models.
- In this general case we may be interested in estimating the parameter β given the data (x). To do so, we need only know the distribution of the data given the parameter $[x|\beta]$ (often called the likelihood) and any prior knowledge about the distribution of the parameter $[\beta]$.

Hierarchical Specification

- We want to characterize real environmental processes in the presence of data.
- We have *a priori* scientific knowledge about the process evolution.
- If we assume the data are a realization from such a process, which is latent and evolves dynamically, then a hierarchical probability model is useful:

  [data|process][process]

- Our knowledge of the process contains uncertainty, so we must learn about the process parameters as well:

  [data|process][process|parameters][parameters]

Hierarchical Specification [notes]

- This is just a very general representation of a hierarchical model.
- In the specific application of modeling invasive species, each of these components will have specific probability distributions associated with them.

Hierarchical Components

- [data|process]: Specified in the usual statistical sense. Accounts for possible observational uncertainty and/or measurement error.
- [process|parameters]: Specified with discretized scientific model (for computation).
- [parameters]: Specified according to *a priori* scientific knowledge or lack thereof.
- [process, parameters|data]: We want to learn about the true process given the data (via Bayes).

Hierarchical Components [notes]

- These models are very data and process specific.
- Each different scientific problem (and dataset) will require a different model specification. That is, different probability distributions and process models.
- The specification given in the following slides is relevant to the spatio-temporal ECD model only, though the general framework holds for many similar problems.

Data Model

$$n_{i,t}|N_{i,t},\theta \sim \text{Beta-Binomial}(N_{i,t},\theta),$$

where, $i = 1,\ldots,m$, $t = 1,\ldots,T$, and

- $n_{i,t}$: sample count at location i and time t.
- $N_{i,t}$: Population size at location i and time t.
- θ: probability of detection parameters (assumed to be known).

Data Model [notes]

- Here $n_{i,t}$ for all locations i and times t, are the data.
- $N_{i,t}$ is the "true" population size and the thing we want to estimate.
- $\theta = \{\alpha, \beta\}$ are the parameters corresponding to the probability of detection. In this case we have estimated them using a separate model (see Royle and Dorazio 2006).
- The Beta-Binomial model is an "over-dispersed" binomial model where the $n_{i,t}$ is a random integer from zero to $N_{i,t}$. This data model allows us to account for the uncertainty in the probability of detection through the parameters α and β.

Process Model

$$N_{i,t}|\lambda_{i,t} \sim \text{Poisson}(\lambda_{i,t}) \quad \text{,for} \quad t = 1,\ldots,T,$$

where, $\boldsymbol{\lambda}_t$ is the mean and variance of the population size at time t and is modeled via a latent dynamic process:

$$\begin{aligned}\boldsymbol{\lambda}_t &= \mathbf{H}\boldsymbol{\lambda}_{t-1} \quad , \\ &= \mathbf{MG}\boldsymbol{\lambda}_{t-1} \quad \text{,for} \quad t = 1,\ldots,T,\end{aligned}$$

- $\mathbf{G} = \mathbf{G}(a, b, \boldsymbol{\lambda})$ is the growth matrix.
- $\mathbf{M} = \mathbf{M}(\delta)$ is the movement (dispersal) matrix.

Process Model [notes]

- The Poisson model is a common model for "relative abundance" and allows for a substantial amount of variability in the "true" population size.
- The dynamical process model is a version of a "matrix model", as it is known in ecology.
- Conventionally, matrix models are used to study demographics in population growth. Here we modify it to study dispersal in population growth.
- Matrix Models are thoroughly discussed in:

Caswell, H. (2001) Matrix Population Models. Sinauer Associates, Inc., Sunderland, MA.

Process Model (growth and dispersal)

Growth Model:

$$\mathbf{G}(a, b, \lambda_{i,t}) = \exp\{b\left(1 - \frac{\lambda_{i,t}}{a}\right)\},$$

Dispersal Model:

$$\mathbf{M}(\delta) = [(M_{i,j})]_{m \times m},$$

$$M_{i,j} \propto \exp\{-\frac{d_{i,j}^2}{\delta_j}\}.$$

Process Model (growth and dispersal) [notes]

- This slide illustrates how these matrices are parameterized.
- The growth model is a Ricker growth equation; one form of a non-linear or density dependent growth equation. For further information see:

Kot, M. (2003) Elements of Mathematical Ecology. Cambridge University Press, Cambridge, UK.

- The dispersal model is a Gaussian (i.e., Normal) dispersal kernel. That is, it is a function that weights the dispersal of organisms from one location to another based on the distance between them ($d_{i,j}$). This is the component of the model that allows for **spatial** effects.
- The parameters controlling the rate of dispersal (δ_j) vary by location, allowing for organisms to move more easily in different areas. This also allows for a heterogeneous environment.

Parameter Model

Probability distributions for the parameters can be specified based on prior scientific knowledge (or lack thereof):

$$a \sim Gamma(\alpha_a, \beta_a)$$
$$b \sim Normal(\mu_b, \sigma_b^2)$$
$$log(\boldsymbol{\delta}) \sim Normal(\boldsymbol{\mu}_\delta, \boldsymbol{\Sigma}_\delta)$$
$$log(\boldsymbol{\lambda}_1) \sim Normal(\boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda)$$

Parameter Model [notes]

- These prior probability distributions represent *a priori* scientific knowledge about the parameters.
- They would be different for different problems of interest and should be specified vaguely (i.e., with large varibility) if little is known about the parameters.

Implementation

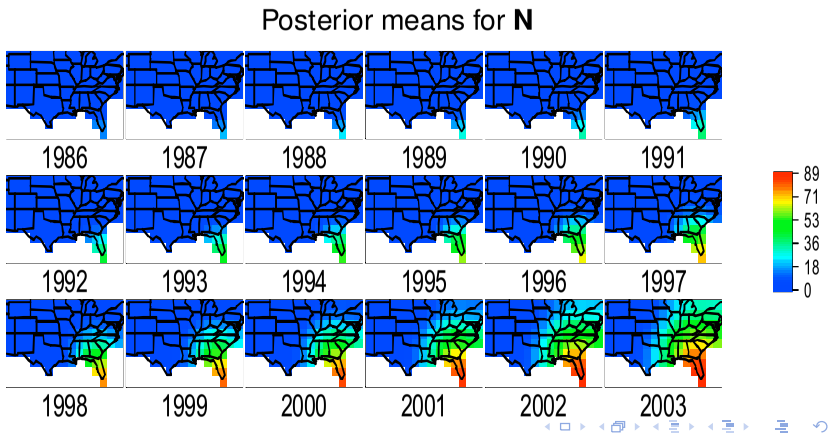
- We want to estimate the model parameters as well as characterize and predict the population size over time given the data.
- Bayes theorem and various sampling methods can be utilized to do this:

$$[\mathbf{N}, \lambda, a, b, \delta | \mathbf{n}] \propto [\mathbf{n} | \mathbf{N}, \theta][\mathbf{N} | \lambda, a, b, \delta][\lambda, a, b, \delta]$$

Implementation [notes]

- So we're interested in learning about $\mathbf{N}$, λ, a, b, and δ given the data ($\mathbf{n}$). That is, the we want to estimate the stuff on the left hand side of the equation (what's known as the posterior distribution).
- What allows us to do so is by writing the complicated joint model as the series of simpler probability models on the right hand side of the equation.
- We can't find these exactly, but we can get pretty close by taking numerous samples from these distributions and calculating our statistics based on the samples. One such method for sampling is called Markov Chain Monte Carlo (MCMC).

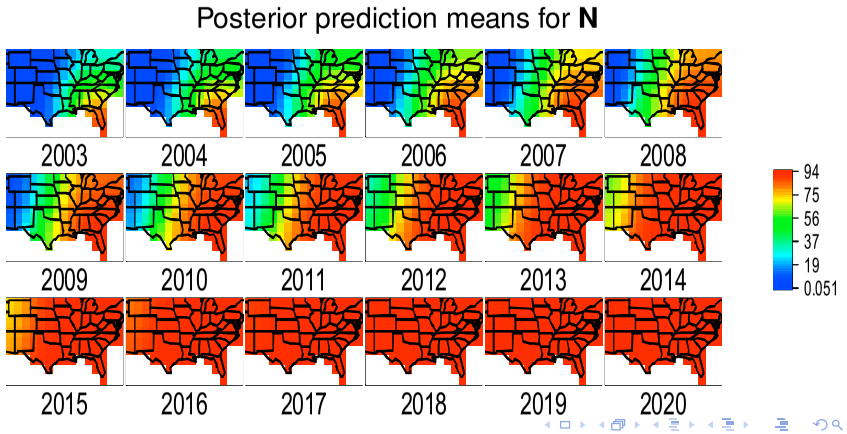
Estimation (maps)



Estimation (maps) [notes]

- These maps show the mean of the posterior distribution for the "true" population size.
- Notice how the invasion starts in South Florida and spreads out as well as grows in intensity over time.
- Also notice how the population in South Florida ceases to grow after about 2001. It has reached its carrying capacity there.

Prediction (maps)



Prediction (maps) [notes]

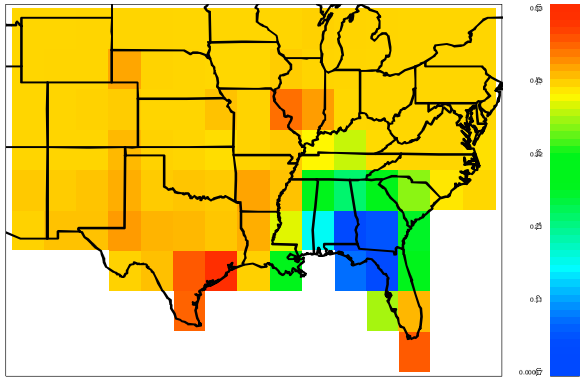
- Recall that speculation in the year 2000 suggested that the ECD would invade most of North America within a few decades.
- These maps provide a forecast by displaying the mean posterior predictions for future years.
- This model provides some statistical justification for such speculations.

Estimation and Prediction (movie)

Estimation and Prediction (movie) [notes]

- This movie just combines the previous two slides into a sequence of images.

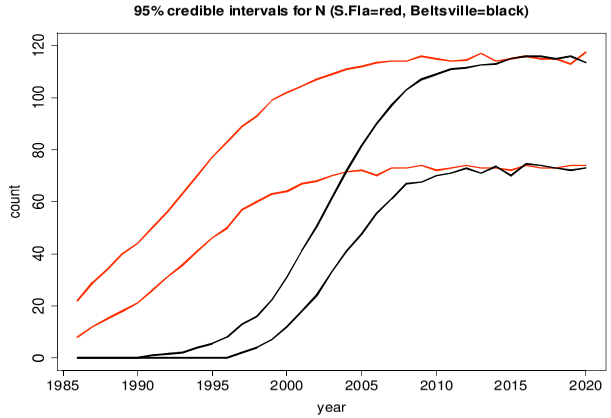
Dispersal (δ)



Dispersal (δ) [notes]

- This image represents the mean of the posterior distribution for the dispersal parameters (δ).
- Notice how there is a pocket of low dispersal in Northern Florida.
- This area of low dispersal is only marginally significant (based on the variability of the posterior distribution; not shown), but the effect of which can be seen in the previous movie.
- Essentially, ECDs are dispersing slower in that area than in some others.

Estimation and Prediction (growth curves)

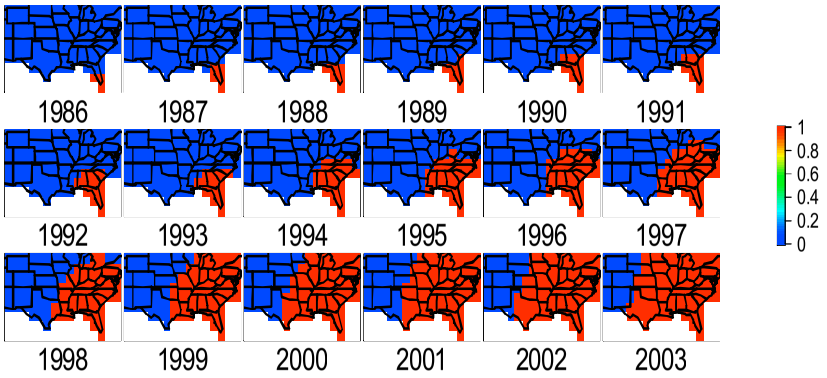


Estimation and Prediction (growth curves) [notes]

- These curve envelopes allow us to compare the population growth for two locations simultaneously.
- For each location, the lower line represents the 2.5^{th} percentile of the posterior distribution for population size and the upper represents the 97.5^{th} percentile.

Presence/Absence

$P(\mathbf{N} > 0) \geq 0.95$



Presence/Absence [notes]

- An advantage of obtaining the output from Bayesian models is that we can easily calculate probabilities.
- These maps show in red all areas where the probability of presence is at least 0.95.
- These can be viewed as probabilistic range maps.
- They are not to be confused with political election maps.

Summary

In this setting, the matrix model specification accommodates:

- *a priori* scientific knowledge.
- Flexible dynamical behavior.
- Multiple sources of uncertainty.
- Long-range predictions (assuming no population collapse).

In addition to:

- More intuitive parameterization than other models (e.g., partial differential equation based models).
- More accessible to ecologists and managers.

Summary [notes]

- In addition, there are many extensions to this model that can be (and were) implemented.
- These include things like: letting other parameters vary spatially and the comparison of models with different specifications.

Diagnostics for Spatial Models

Mark Otto and David Meek

U.S. Fish and Wildlife Service, Laurel, MD
U.S.Department of Agriculture—Agricultural Research Service, Ames, IA

16 March 2006

Box-Jenkins Iterative Modeling Procedure

1. Identify the model: transformation, regression, trend, correlation structure
2. Estimate model parameters
3. Check that the model fits the assumptions
4. Repeat 1–3 until diagnostics check out

General Concepts

See that the estimated model fits the assumptions. The usual assumption is that the residuals have a zero mean and constant variance.

- Residuals do not show any consistent patterns
  - Fitted values
  - Regression variables
  - Important spatial coordinates
  - Time
- Residuals are white noise: check with the variogram of the residuals

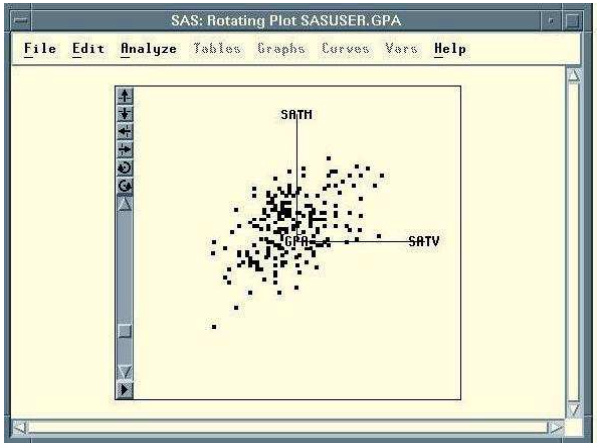
Stem Plots

```
stem(rbyc)
 The decimal point is at the |
 -2 | 33333333333333333333333333333333333333333333333333333333333333333333
 -1 |
 -1 |
 -0 |
 -0 |
  0 | 000000000000000
  0 | 77777777777777
  1 | 11111111111144444
  1 | 666666666668889999
  2 | 1222222222334
  2 | 55667
  3 |
  3 | 8
  4 | 0
```

5

Spinning

- Spatial data is ≥ 3 dimensional
- Spin map to change point of view

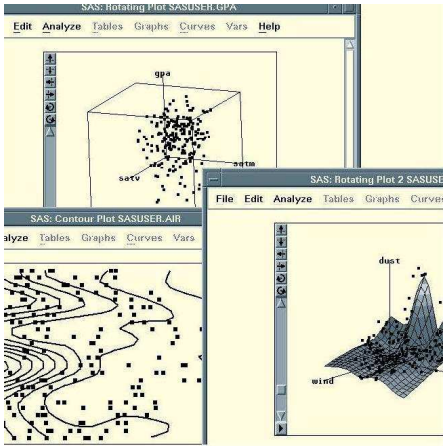


These were done in SAS Insight.

6

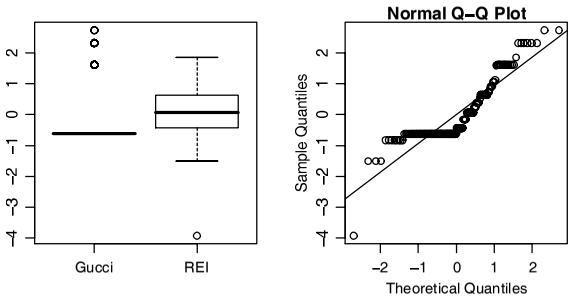
Brushing

- Identify interesting points
- Link to identify in other graphs and tables



7

Box and Normal Probability Plots



8

When the independent variables are factors the residuals can be shown in box-plots. Here the levels are different and the variation is much different between the factors. This is on the original data. The lines in the normal probability plot show the strong correlations in the data.

Regression Diagnostics

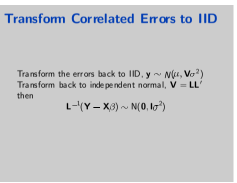
- Studentized residuals: mean removed, standard variance
- Influence statistics: change due to missing observation
- Plot the above by direction, fitted value, fixed, time

These diagnostics differ from regression diagnostic in that they are transformed by the same linear transformation used on the correlated residuals to make them IID, with $\mathbf{L}^{-1}$.

Transform Correlated Errors to IID

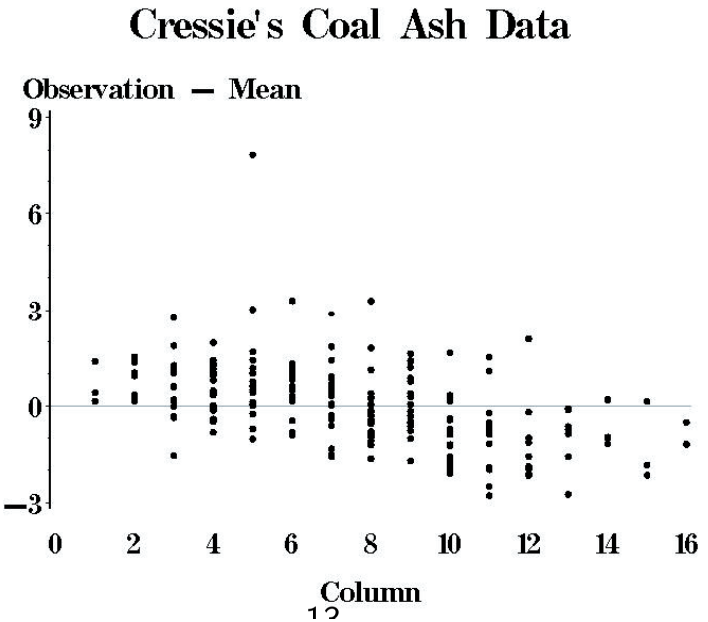
Transform the errors back to IID, $\mathbf{y} \sim N(\mu, \mathbf{V}\sigma^2)$
Transform back to independent normal, $\mathbf{V} = \mathbf{L}\mathbf{L}'$
then

$$\mathbf{L}^{-1}(\mathbf{Y} - \mathbf{X}\beta) \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2)$$



Have talked about how to define correlations in other talks

Standardized Residuals

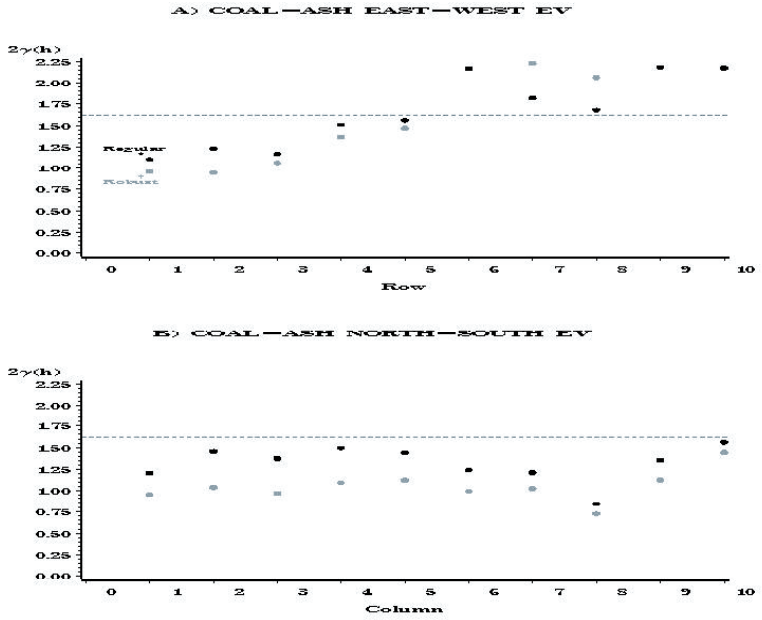


12

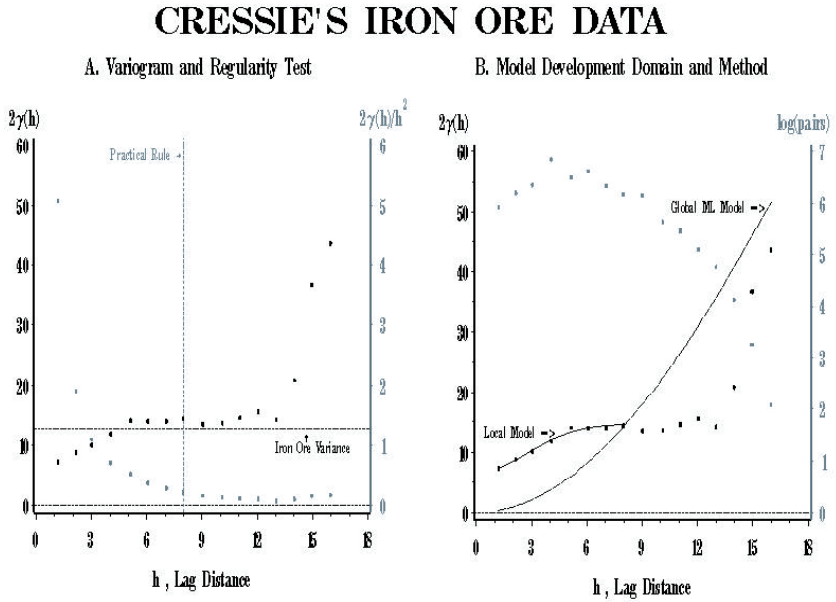
Variogram Plots of Residuals

- Omnidirectional classical, $\gamma(h)$, and robust, $\gamma^{\dagger}(h)$ variograms
- Try directional $\gamma_{\alpha}(h)$ in two (0 and 90 deg) to six (0, 30, ..., 150 deg) depending on the limits of the data
- Pair count vs. h Include a reference line for white noise variogram, $\hat{\gamma} = \sigma^2$
- Regularity test, $\gamma(h)/h^2$ vs. h
- Correlation structure, $\rho(h)$ or $C(h)$ vs. h

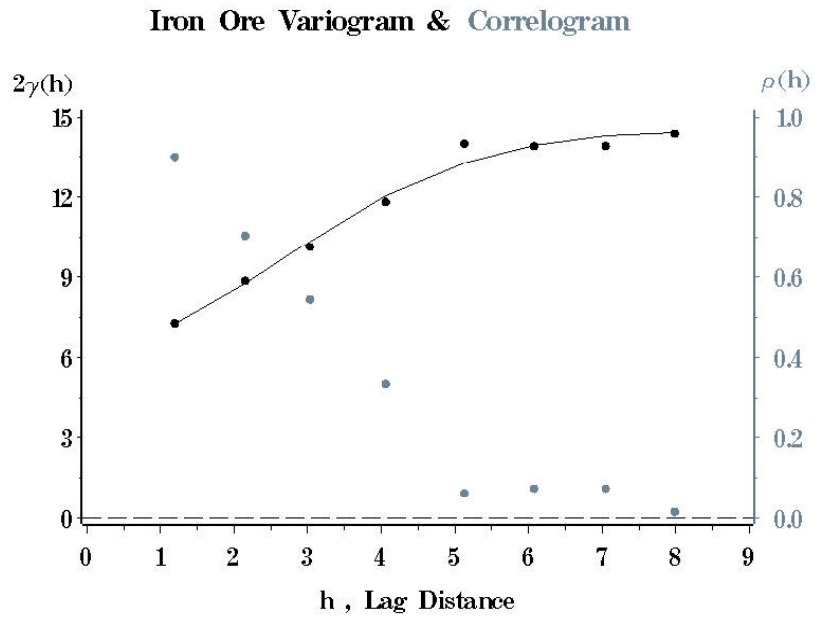
14



15



16



17

Model Mispecification Grid Check

- Ribeiro and Diggle (2004) suggest eye-ball variogram parameters to fit empirical
- Suggest fitting model so residual variogram is white-noise

18

Outliers

- Two types: point (a single outlying value) and patch (a region at a different level)
- Difficult to identify with correlated data because they don't stand out on their own. It is how the differ from nearby values
- Use a priori knowledge (at least use to confirm)
- Can confound spatial correlation structure

19

Point Outlier

- Set outlier detection critical value, $t = 3.5$ (p=0.01 controlling for an experiment-wise error rate)
- Identify the fixed effects and initial correlation structure
- Estimate and fix correlation structure
- Add outlier dummy for each observation and estimate
- Add most significant outlier to fixed effects
- Repeat 2–4 until no more significant outliers
- Reassess the fixed effects and correlation structure

20

Iterative Outlier Detection

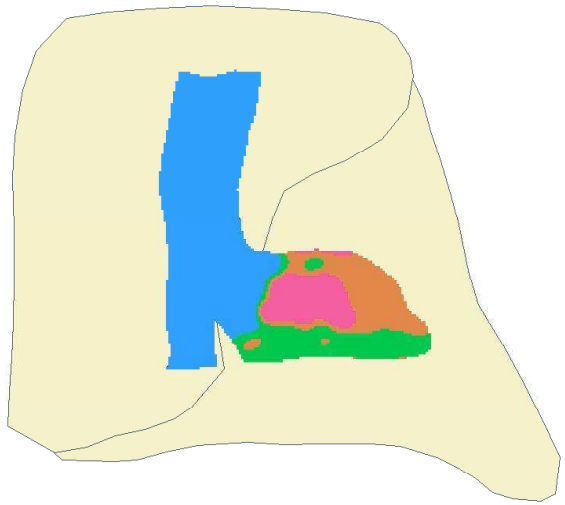
```
Iteration 1.
  Outlier        Est       SE       t
+ Obs47          2.31      0.1322   4.2 *
  Obs141         1.22      0.1411   3.1

Iteration 2.
  Obs47          2.37      0.1072   3.9 *
+ Obs141         2.07      0.1100   3.6 *
```

Outlier at 47 was picked up on the first iteration but 141 was below the critical level. The model was re-estimated with new correlations and variance. An outlier at 141 was then picked up. After the next round no more observation were over the critical value

Region Breaks

- Region a different level
- Too difficult to test every possible patch
- Use ArcGIS patch identification tool (classification algorithm)



Green patch identifies sole fishing area with higher fishing and bycatch

What Diagnostics do not Cover

- Measurement error:
  - Errors caught by repeated measurement at the sample points
  - Errors independent regression variables
  - Inaccuracies in the locations and thus the measures of distance between points
- Micro-scale variation: errors at scales below the smallest distance increment
- Aliasing, variation at periodicities covered by spacings in the data
- Emphasizes the importance of good design to address the important sources of error

Conclusions

- Use the Box-Jenkins iterative modeling approach
- Use Exploratory Data Analysis stem-plots, qq-plots, and brushing for multi-dimensional views of the data
- Look for patterns in residual and influence diagnostics
- Check that classical and robust variograms look like white noise and there is no anisotropy in the residuals
- Iteratively identify point outliers. Use them to look harder at outlying values. Have a priori reasons for including outliers and patches in the model.

25

Introduction to Spatial Point Pattern Analysis

by

Stephen L. Rathbun
Department of Health Administration, Biostatistics, and Epidemiology
College of Public Health
University of Georgia
Athens, GA 30605
rathbun@uga.edu

References:

Ripley, B.D. (1981). *Spatial Statistics.* Wiley, New York.

Diggle, P.J. (2003). *Statistical Analysis of Spatial Point Patterns,* 2nd Ed. Oxford University Press, London.

Upton G.J.G., and Fingleton, B. (1985). *Spatial Data Analysis by Example.* Wiley, New York.

Waller, L.A., and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data.* Wiley, New York.

Møller, J., and Waagepetersen, R.P. (2004). *Statistical Inference and Simulation for Spatial Point Processes.* Chapman and Hall/CRC, Boca Raton, FL.

Definition: A *spatial point pattern* is comprised of the locations of events.

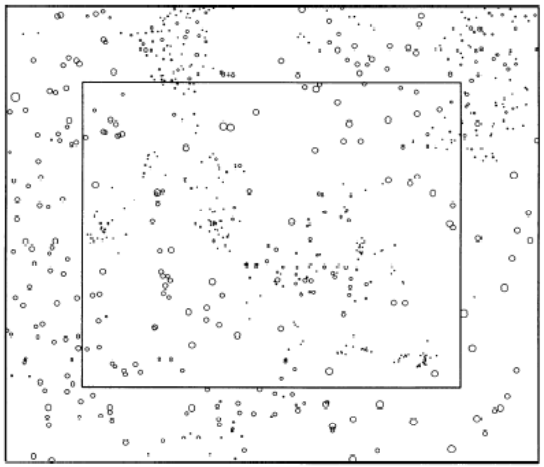
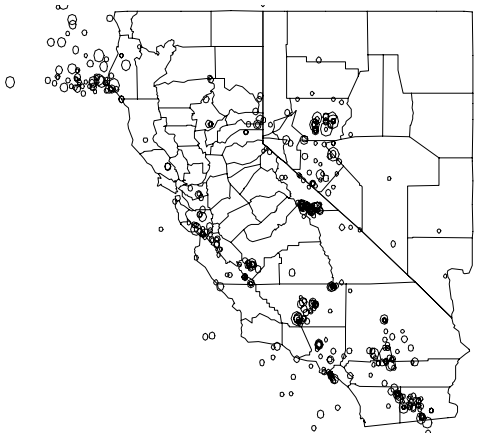


Figure 2. Map of All Longleaf Pines in the 150 × 120 m Study Region B (Inner Rectangle) and the 30 Meter Wide Guard Region $B_+ - B$. The direction north is toward the right side of the page.

California Earthquakes



Point pattern analysis is primarily concerned with modeling the locations of events, for example the locations of:

- Trees
- Birds' nests
- Ants' nests
- Earthquake epicenters
- Cancer cases
- Galaxies

Objectives: Point Pattern Analysis

1. To determine if the point pattern is completely random;
2. If the pattern is not completely random, fit an explanatory point process model to the data.

Complete Spatial Randomness

Definition: A point pattern is *completely random* if it is realized from a homogeneous Poisson process.

Definition: For a homgeneous Poisson process with intensity λ

1. The number of events (trees) $N(A)$, in a study region A is Poisson distributed with mean $\lambda|A|$

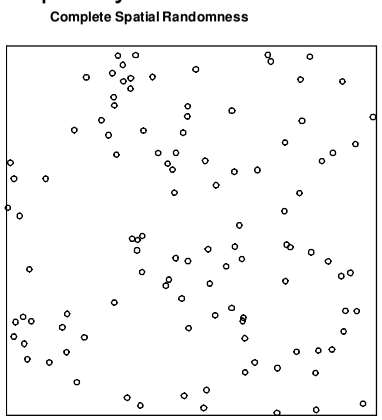
$$\Pr\{N(A) = n\} = \frac{1}{n!} e^{-\lambda|A|} (\lambda|A|)^n$$

2. Conditional on the number of events (trees), the event locations are independently sampled from a uniform distribution on A.

Definition: The *intensity* λ is equal to the mean number of events per unit area.

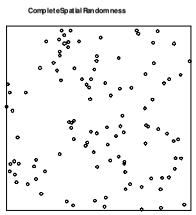
Note: In ecology, the intensity is called the density. In statistics, we use the term intensity to distinguish it from a probability density function.

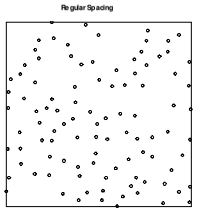
Completely Random Pattern

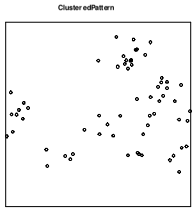


Complete spatial randomness is the null model against which spatial point patterns are often compared.

Completely Random | Regular | Clustered







In Ecology:

- Regular spacing may result from intraspecific competition for limited resources;
- Clustered patterns may result from:
  - Clustering of offspring around their parents;
  - Response to a heterogeneous environment.

Ripley's K-Function

Ripley's K-function is the most effective tool for assessing departure from complete spatial randomness.

Definition:

$$K(r) = \frac{\text{Mean number of trees within distance } r \text{ of an arbitrary tree}}{\lambda}$$

Estimation:

$$\widehat{K}(r) = \frac{1}{\widehat{\lambda} N} \sum_{i \neq j} w_{ij} I(d_{ij} \leq r)$$

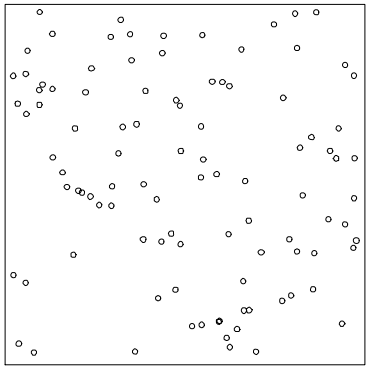
where

$$\widehat{\lambda} = \frac{N}{|A|}$$

is the number of trees in the study region divided by the area of the study region.

What is this?

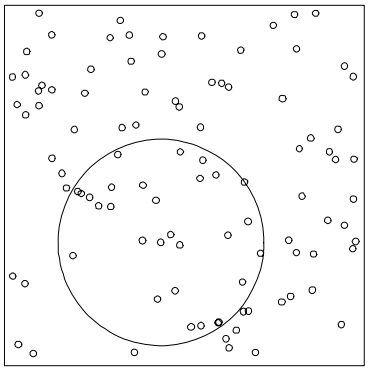
Consider the point pattern of trees:



Here there are 100 trees in a 10×10 region. So

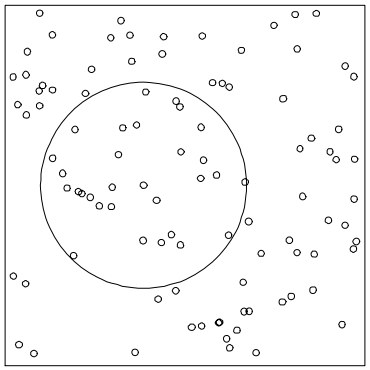
$$\widehat{\lambda} = \frac{100}{10 \times 10} = 1$$

Place a circle of radius r around an arbitrary tree:



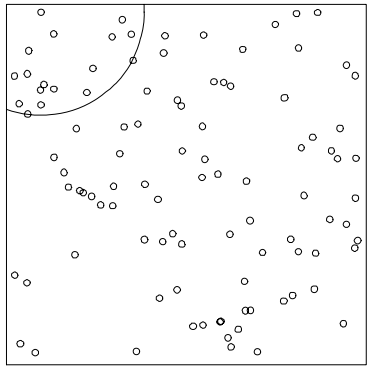
Count the number additional of trees within the circle.

Repeat for each of the remaining trees:



Counting the number of additional trees within each circle.

Edge Correction



For trees close to the edge of the study region, we cannot observe the number of trees within radius r.

Here, we give the neighboring trees a weight w_{ij} equal to one divided by the portion of the circle of radius d_{ij} inside the study region.

13

The results are averaged over all base trees

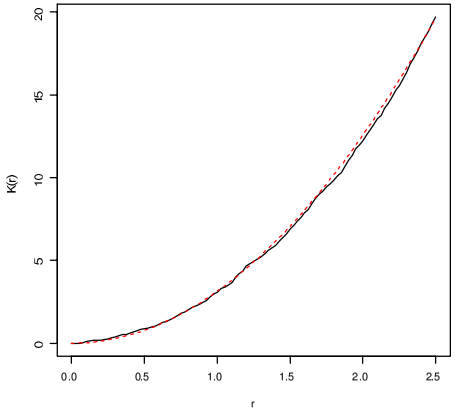
$$\frac{1}{N}\sum_{i\neq j} w_{ij} I(d_{ij} \leq r)$$

and then divided by the estimated intensity $\hat{\lambda}$ to obtain the estimate

$$\widehat{K}(r) = \frac{1}{\hat{\lambda} N}\sum_{i\neq j} w_{ij} I(d_{ij} \leq r)$$

14

Plot $\widehat{K}(r)$ against r



Note: Under complete spatial randomness,

$$K(r) = \pi r^2$$

15

Note: Even for strong departures from complete spatial randomness, the difference between the empirical K-function and its expectation under complete spatial randomness is small.

Therefore, a plot of the K-function may not be very informative.

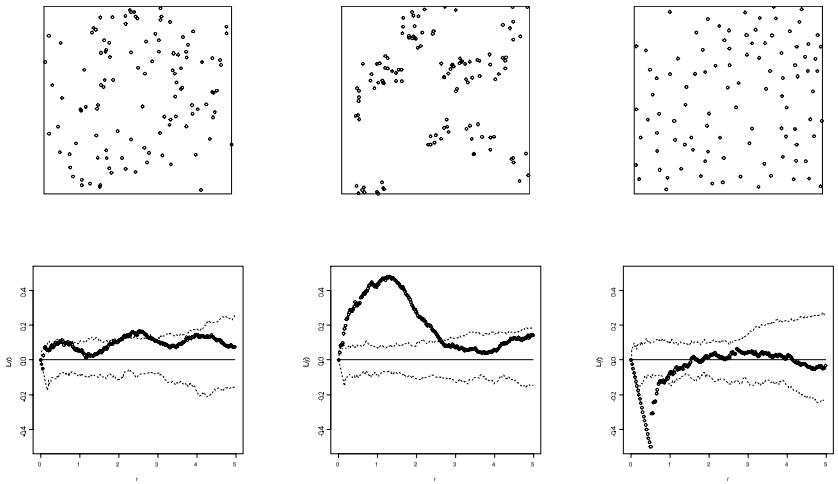
Solution: Linearizing Transformation:

$$L(r) = \sqrt{K(r)/\pi} - r$$

- Under complete spatial randomness
$$L(r) = 0$$
- For clustered patterns
$$L(r) > 0$$
- For regular spacing
$$L(r) < 0$$

16

Completely Random | Clustered | Regular



Note: By plotting the L-function against distance, all scales of pattern can be examined.

Point Process Models
Inhomogeneous Poisson Process

Definition: The *intensity* of a point process is

$$\lambda(\mathbf{s}) = \lim_{|d\mathbf{s}|\to 0} \frac{E\{N(d\mathbf{s})\}}{|d\mathbf{s}|}$$

The intensity can be viewed as a local density. Regions with high intensities will tend to contain large numbers of trees, while regions with low intensities will tend to contain few trees.

- $N(d\mathbf{s})$ is the number of trees in a small region $d\mathbf{s}$ surrounding the location $\mathbf{s}$
- $E\{N(d\mathbf{s})\}$ is the mean number of trees in $d\mathbf{s}$
- $|d\mathbf{s}|$ is the area of the region $d\mathbf{s}$
- Thus, the intensity $\lambda(\mathbf{s})$ is the mean number trees per unit area, as a function of location $\mathbf{s}$.

Inhomogeneous Poisson Process

The inhomogeneous Poisson process may be used to model the impact of spatial variation in environmental characteristics (e.g., elevation, light intensity, nutrient concentrations) on a point pattern.

Definition: For an inhomgeneous Poisson process with intensity λ

1. The number of events (trees) $N(A)$, in a study region A is Poisson distributed with mean

$$\Lambda(A) = \int_A \lambda(\mathbf{s}) d\mathbf{s}$$

That is, the probability that the number of events $N(A)$ equal to n is

$$\Pr\{N(A) = n\} = \frac{1}{n!} e^{-\Lambda(A)} (\Lambda(A))^n$$

2. Conditional on the number of events, the event locations are independently sampled from a probability density function proportional to $\lambda(\mathbf{s})$.

Space-Varying Covariates

Let

$$x_1(\mathbf{s}), x_2(\mathbf{s}), \cdots, x_p(\mathbf{s})$$

denote the values of p space-varying covariates at the location $\mathbf{s}$ in the study region A (e.g., elevation, light intensity, nutrient concentrations, etc.).

The impact of these space-varying covariates on a spatial point pattern may be modeled through the intensity function:

$$\lambda(\mathbf{s};\boldsymbol{\beta}) = \exp\{\beta_0 + \beta_1 x_1(\mathbf{s}) + \beta_2 x_2(\mathbf{s}) + \cdots + \beta_p x_p(\mathbf{s})\}.$$

An inhomogeneous Poisson process with the above intensity is called a *modulated Poisson process.*

Reference

Cox, D.R. (1972). The statistical analysis of dependencies in point processes. In P.A.W. Lewis (ed.), *Stochastic Point Processes,* pp. 55-66. New York: Wiley.

Parameter Estimation

The *maximum likelihood estimator* is obtained by finding $\hat{\boldsymbol{\beta}}$ that maximizes the log likelihood:

$$L(\boldsymbol{\beta}) = \boldsymbol{\beta}' \sum_{i=1}^{n} \mathbf{x}(\mathbf{s}_i) - \int_A \exp\{\boldsymbol{\beta}'\mathbf{x}(\mathbf{s})\}$$

where

- $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n$ denote the locations of n trees in the study region A.
- $\mathbf{x}(\mathbf{s})$ = vector of covariates at the location $\mathbf{s}$ in A.

Problem: This requires that the values of the covariates be observed for:

- All of the trees in the study region.
- All locations in the study region.

The former may be impractical, and the later impossible to obtain.

Two Approaches:

1. Rathbun (1996) *Biometrics* **52**, 226-242.
2. Rathbun, Shiffman, and Gwaltney (2006) In *Models for Intensive Longitudinal Data.* T.A. Walls and J.L. Schafer (eds.). Oxford.

Approach 1

- Sample the covariates at a collection of sites

$$\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m$$

- Use kriging to predict the values of the covariates at the locations of the trees, and at the unsampled sites.
- Substitute predicted values into the log likelihood:

$$\hat{L}(\boldsymbol{\beta}) = n\beta_0 + \beta_1 \sum_{i=1}^{n} \hat{x}(\mathbf{s}_i)$$

$$- \int_A \exp\{\beta_0 + \beta_1 \hat{x}(\mathbf{s}) \underbrace{+\frac{1}{2}\beta_1^2(\sigma^2 - \mathrm{var}(\hat{x}(\mathbf{s})))}_{\text{Bias Correction}}\} d\mathbf{s}$$

- Find $\tilde{\boldsymbol{\beta}}$ that maximizes the approximate log likelihood $\hat{L}(\boldsymbol{\beta})$.

Example: Titi Hammock Data

Beech-Magnolia Forest in South Georgia

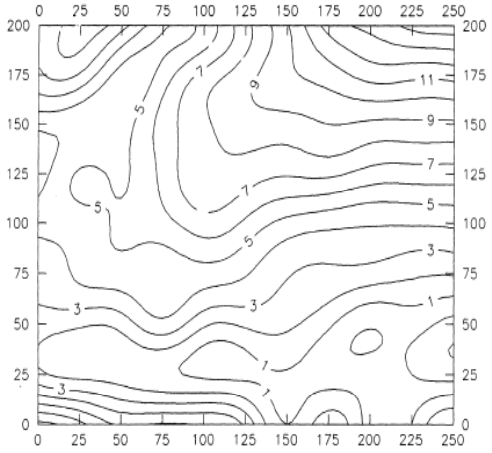


Figure 3. Contour map of predicted elevations in Titi Hammock.

Results:

Parameter estimates for a modulated Poisson process with intensity (5.2). Standard errors are given in parentheses.

	No bias correction		Bias corrected	
Species	$\hat{\beta}_0$	$\hat{\beta}_1$	$\tilde{\beta}_0$	$\tilde{\beta}_1$
Bay	−4.3729	−0.9204	−4.4300	−0.8819
	(0.1355)	(0.0936)	(0.1330)	(0.0883)
Beech	−5.4476	−0.0613	−5.4495	−0.0609
	(0.1384)	(0.0262)	(0.1381)	(0.0261)
Blue beech	−5.0711	0.1362	−5.0667	0.1355
	(0.0821)	(0.0113)	(0.0819)	(0.0113)
Holly	−5.2169	0.0161	−5.2164	0.0160
	(0.1099)	(0.0183)	(0.1097)	(0.0182)
Ironwood	−3.2833	−0.7621	−3.3264	−0.7384
	(0.0760)	(0.0432)	(0.0749)	(0.0415)
Magnolia	−5.1454	−0.0277	−5.1462	−0.0276
	(0.1135)	(0.0203)	(0.1132)	(0.0202)
Tulip poplar	−4.8605	−1.0234	−4.9270	−0.9725
	(0.1750)	(0.1356)	(0.1717)	(0.1265)

Approach 2

Data Requirements: Covariates are observed

- Locations of the trees

$$\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n$$

- Random locations from the study region

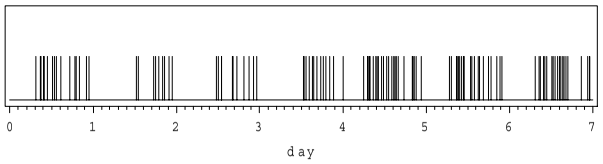
$$\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m$$

Find $\tilde{\boldsymbol{\beta}}$ that maximizes the approximate log likelihood

$$\widehat{L}(\boldsymbol{\beta}) = \boldsymbol{\beta}' \sum_{i=1}^{n} \mathbf{x}(\mathbf{s}_i) - \frac{|A|}{m} \sum_{j=1}^{m} \exp\{\boldsymbol{\beta}' \mathbf{x}(\mathbf{u}_j)\}$$

Example: Ecological Momentary Assessment of Smoking

Times at which cigarettes were lit by a smoker



Time-Varying Covariates

- Negative Affect
- Arousal
- Attention
- Restlessness

Results

	Modulate Poisson	
Parameter	Estimate	SE
Intercept	-0.05924	0.00839
Negative Affect	0.01950	0.01077
Arousal	-0.01594	0.01078
Attention	-0.01787	0.01198
Restlessness	0.21017	0.01577

Extensions:

- Obtain covariates on a thinned sample of trees. Visit each tree and sample the covariates with known probability p. More generally, p may depend on location.
- Use alternative designs for covariate sample sites:
  - Stratified Random Sample
  - Transect Samples - Random parallel transects, and random sites along each transect.

29

Example: Spatial and Spatio-temporal Patterns of Yellow Crinkle Disease in Papaya

Dixon, P.M. and Esker, P.
Department of Statistics,
Iowa State University

Spatial point patterns

- Locations of events in space or space x time
- Goals / questions include:
  - Visualizing probability of an event
  - Are events clustered?
  - Are events clustered with other types of events?
  - In space or space x time
- Illustrate using Papaya Phytoplasma data

Phytoplasmas in Papaya

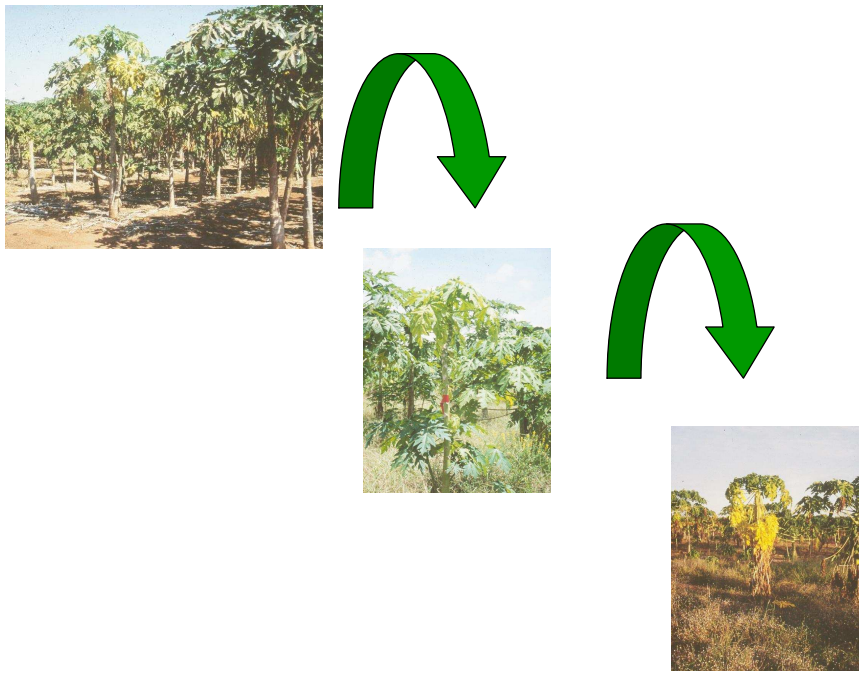
- In Australia, three predominant types of economic importance:
  - Papaya dieback
  - Papaya mosaic
  - **Papaya yellow crinkle**
    - **Tomato big bud (TBB)**
    - **Sweet potato little leaf V4 (SPLL-V4)**
  - Visual scouting for symptoms
    - confirm / identify with PCR and RFLP

Notes

- Phytoplasmas: specialized bacteria lacking cell walls
  - Transmitted by sap feeding insects, e.g. leafhoppers, psyllids
  - Responsible for a variety of yellowing or wilting diseases in large number of plant species
- Identification usually by molecular methods, e.g. sequencing

Notes

- Data from papaya plantation in Northern Territory, Australia, outside Katherine Australia, 14 S, 132 E
- Location of Katherine within Australia



Notes

- Photos show uninfected and two stages of phytoplasma infection.

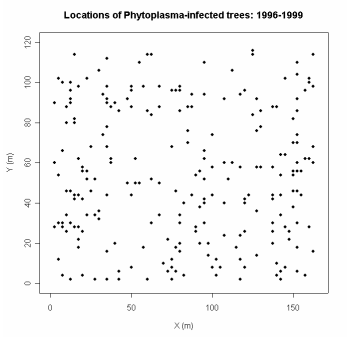
- Upper left photo: uninfected plants and the general layout of the plantation.
- Middle photo: papaya with early symptoms of yellow crinkle disease.
- Lower right photo: plant that is near death due to a phytoplasma infection.
- Infected plants remained in the plantation for this study.

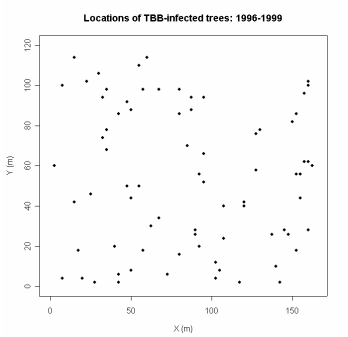
Field Study: May 1996 – April 1999

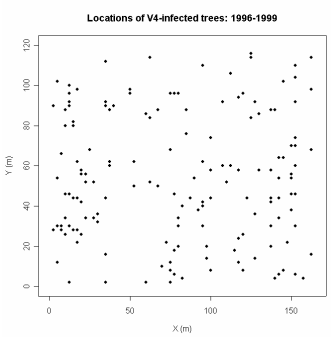
- Padovan and Gibb (2001)
- Time, incubation to death: Esker et al. (2006)
- Plantation measured: 65 x 58 = ~ 3,800 plants
- Planted: January 1996
- Monthly census of all plants (began May 1996 through April 1999)
- Plants left in plantation until plant death (month of death then noted)
- 154 cases of V4 infection; 76 cases of TBB

Notes

- Padovan, A.C. and Gibb, K. S. 2001. Epidemiology of phytoplasma diseases in papaya in Northern Australia. J. Phytopathology 149:649-658
- Esker, P.D., Gibb, K.S., Padovan, A., Dixon, P.M, and Nutter, F.W. Jr. 2006. Use of survival analysis to determine the postincubation time-to-death of papaya due to yellow crinkle disease in Australia. Plant Disease 90:102-107.

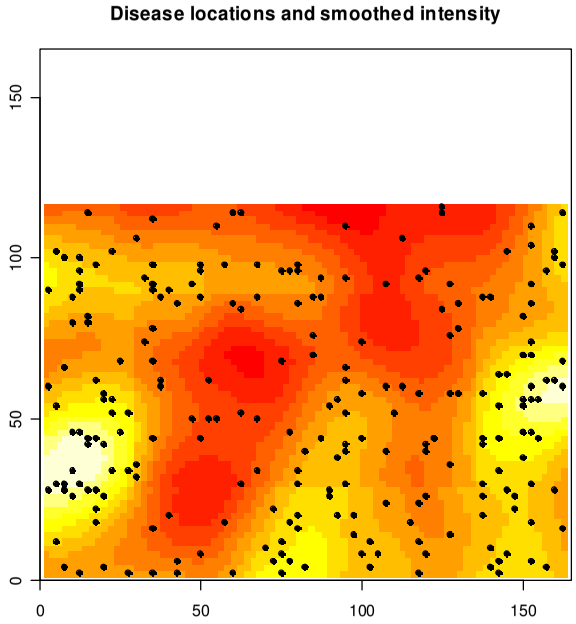






Notes

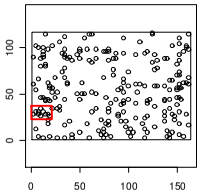
- These plots show locations of diseased plants. Time when symptoms first noticed is ignored.
- Is there any general trend in disease incidence? E.g.:
- Is disease more frequent in one part?, or on the edge?
- Answers not obvious
- Are locations of diseased plants clustered? I.e. are diseased trees surrounded by other diseased trees?
- Hard to tell from the plots of locations. Appear to be places with many and places with few, but the eye is easily fooled.
- Even harder to tell if there is any relationship between the two types

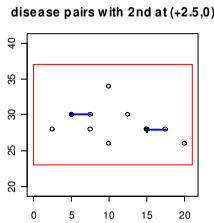


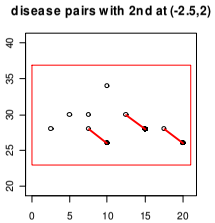
Notes

- Estimate intensity (average number per m^2) using a non-parametric kernel smoother
- On color plot, white = highest intensity, yellow = intermediate intensity, red = lowest intensity
- Black and white plot has contour lines to show the same thing.
- These plots show average number; don't look at relationships between points
- Clustering is a property of sets of points.
  - Clustering: neighborhood of a diseased tree is unusually likely to have another diseased tree
- When set of possible diseases locations is a grid (e.g. trees in a plantation), can count distance-direction pairs: Look at each diseased location. Count number of diseased tree with another diseased tree 'next door' to the east, 'next door' to the north, …, two trees away to the east, …, one north and one east, … Use all combinations of lag distance and direction pairs up to some maximum distance and you get the next plot.

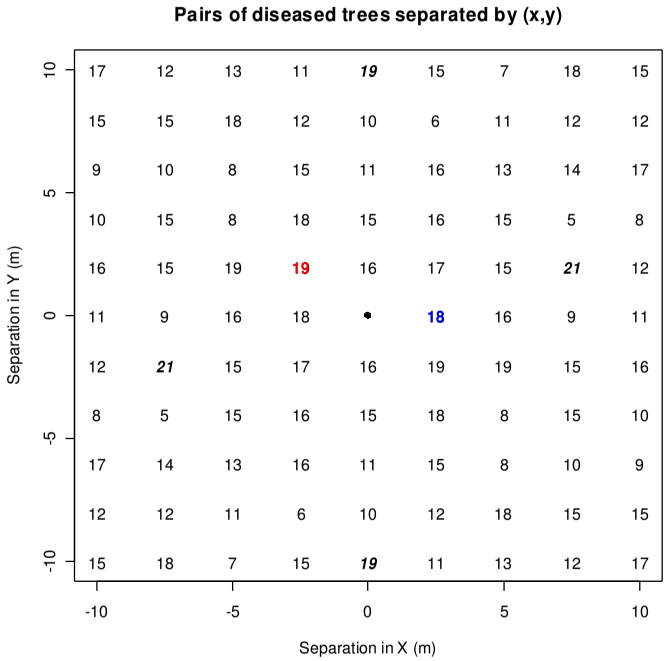
Distance-Direction Pairs







Consider each diseased tree
How many diseased trees are 2.5 m to right?
  2 in this small part of the data
How many are 2.5m to left and 2m up?
  3 in this small part of the data.

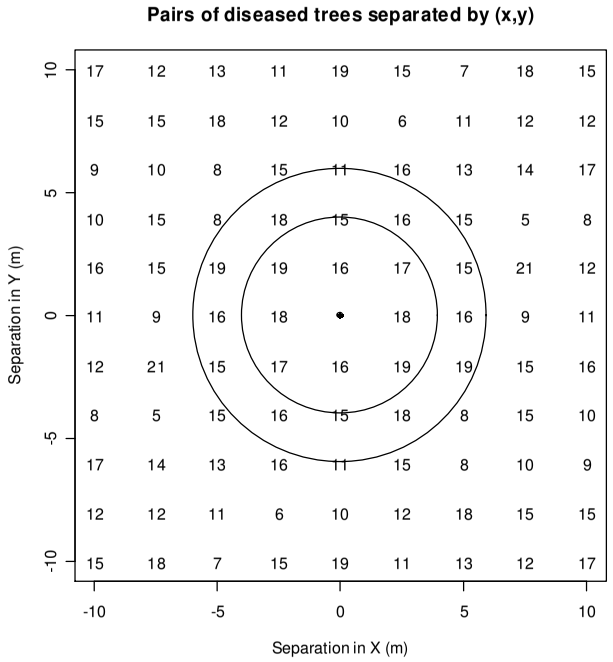


Notes

- Counts in red and blue are the full data set equivalents of the counts illustrated on previous slide
- Visually, no obvious pattern to the counts.
- Disease not more frequent among adjacent individuals or in one specific direction
- A very simple model: Complete Spatial Randomness
  - Locations of diseased trees are independent of each other (no clustering)
  - Probability that a tree is diseased is constant across the study area (no trend)
- Number of diseased individuals in any distance/direction count has Poisson (m) distribution
  - Ignoring edges: m = n(n-1)/(t-1) n = # diseased trees, t = # trees
  - m = 230(229)/3769 = 13.97
  - P[X > 20 | m = 13.86] = 0.047
  - Should account for edge effects: fewer pairs of points separated by 5 trees than by 1 tree.
  - Values in bold italic are significantly larger than expected for that distance and direction

Distance / direction pairs

- Consider each pair of diseased tree
- Tabulate distance/direction between them
- Counts larger than expected (ca. 13)
  - But not unusually so
- Direction doesn't seem to matter
- Consider only distance to increase power

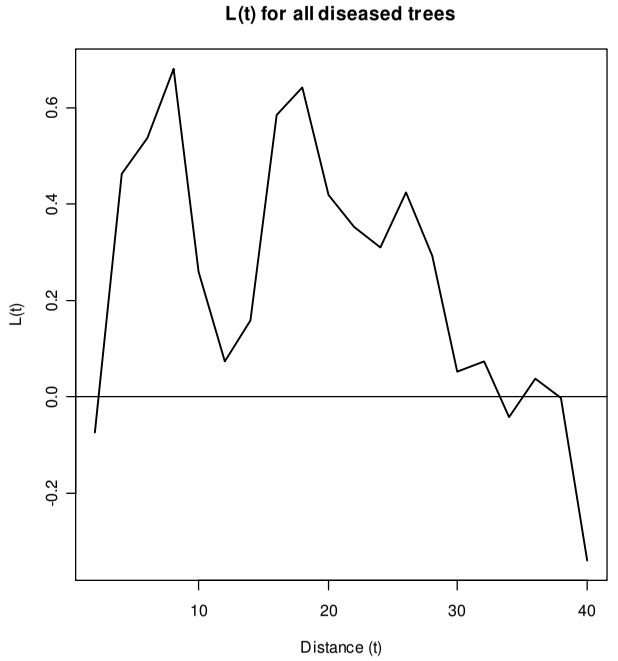


Ripley's K(t)

- Ripley's K(t) combines information across distances, ignores direction
- K(t) = E # add'n points w/i dist. t / intensity
- Often easier to work with $L(t) = \sqrt{(K(t)/\pi)} - t$
- Compare K(t) to πt^2 or L(t) to 0
- Clustering: L(t) > 0
- Segregation: L(t) < 0

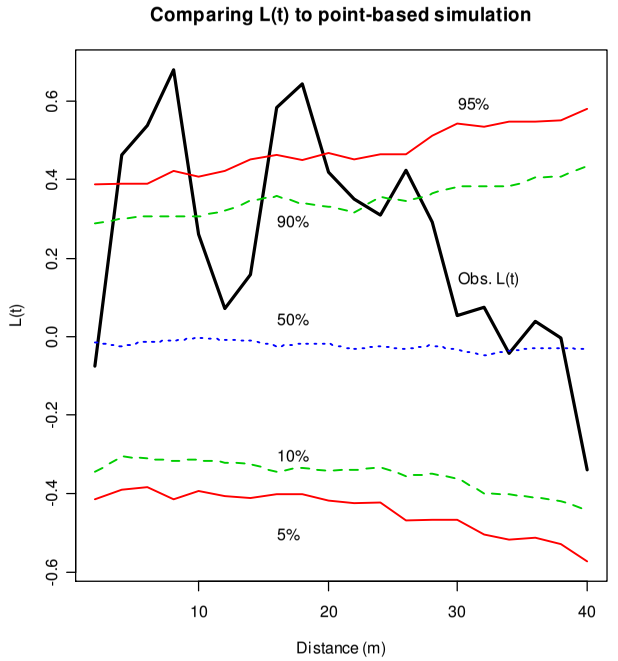
Notes

- E is shorthand for Expected value of. This is the theoretical average.
- Estimate by counting number of points w/i distance t of each point, divide by intensity
- Estimate intensity as # points / total area
- Again, need to account for edge effects, details in many books and papers
- Under CSR: Complete Spatial Randomness, defines a few slides ago:
- $K(t) = \pi t^2$
- $L(t) = 0$



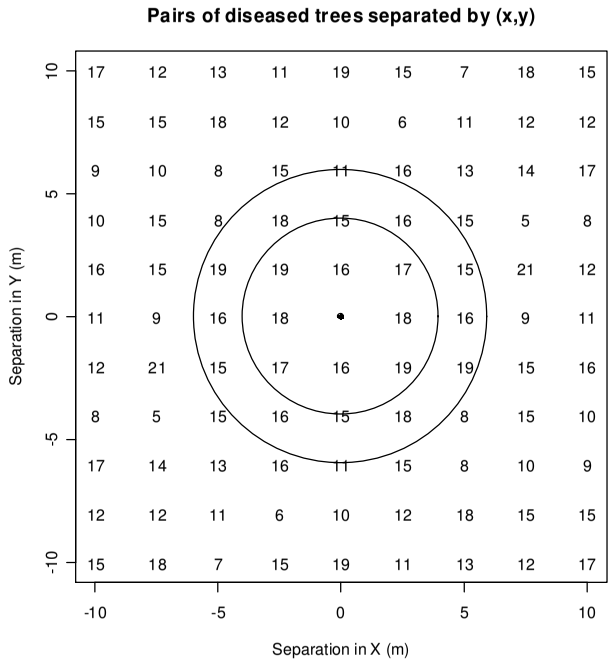
Are diseased trees clustered?

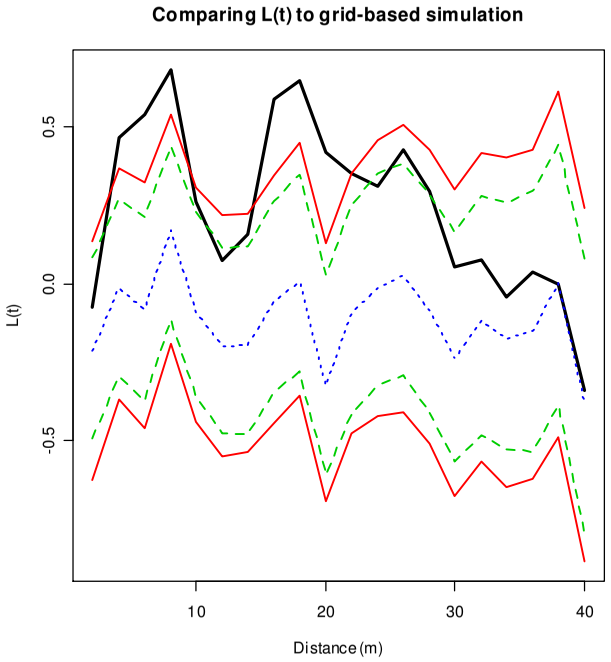
- Data: estimated L(t) > 0 from 4 – 30 m
- But: L(t) estimated from 230 points. How large is the sampling variation?
- Easy to construct test of H0: K(t) = 0 at that specific distance t.
- Simulate complete spatial random process, estimate K(t) and L(t), repeat many times. Calculate quantiles.



Interpretation of L(t) plots

- Focus on clustering, so one-sided test(s)
- More diseased trees than expected within 4m, 6m, 8m, 16m, and 18m of other diseased trees.
- But, trees aren't anywhere: on a grid
- Randomly choose 230 'diseased' locations from the 3770 possible grid positions





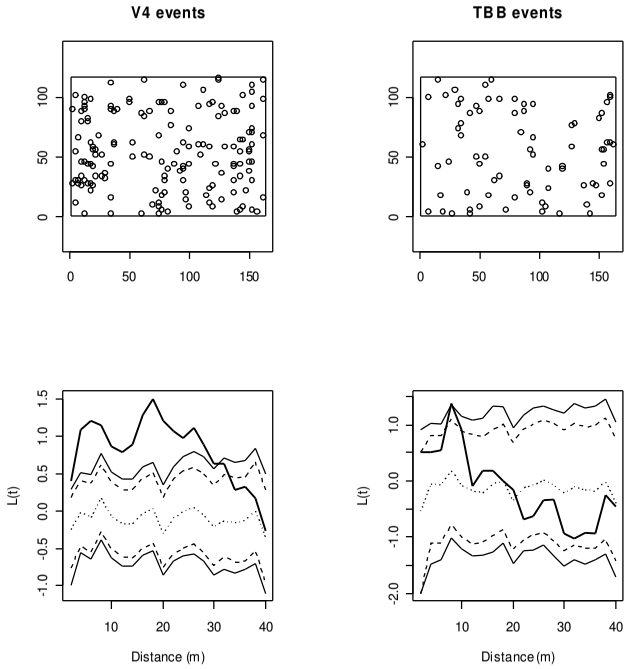
Notes

- Same legend as slide 23:
- Black line = observed L(t)
- Red lines = 0.05 and 0.95 quantiles
- Green lines = 0.10 and 0.90 quantiles
- Blue line = median (0.5 quantile)

- Lines are jagged because there are lots of trees separated by 2m, lots separated by 2.5m, lots separated by 4m
- But none separated by 1.5m, or 3.5m, because of the planting grid

Grid-based simulation

- Similar conclusions:
- More diseased trees than expected within 4m, 6m, 8m, 16m, 18m and 20m of other diseased trees.
- What about each type separately?

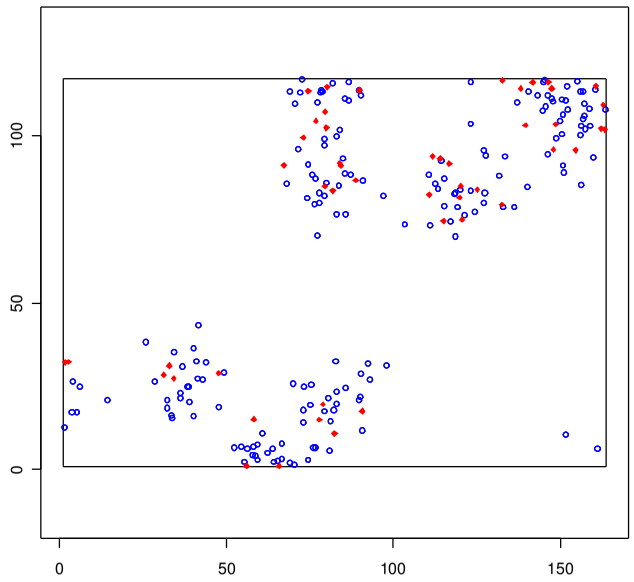


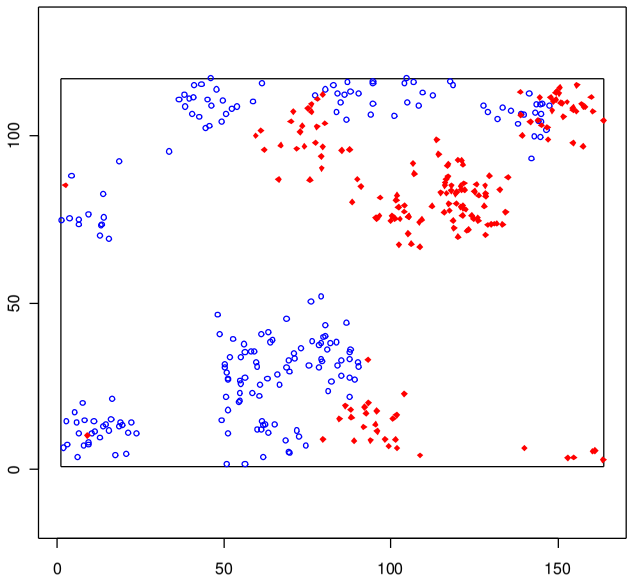
Notes

- Top pair of figures are locations of each type of phytoplasma, plotted separately
- Bottom pair of figures are L(t) for each type plotted separately
- Notice spread between 5% and 95% quantiles.
- Much larger for TBB events (only n=76) compared to V4 events (n=154)
- Can do the analysis for very small sample sizes (e.g. 30 locations), but power is very low.

Association of types

- Are V4 events surrounded by (more, fewer) TBB events?
- Are there clusters of diseased trees? Or, separate clusters of V4 and TBB?
- Concerns relationships between two (or more) processes, not the characteristics of each process.





Bivariate K functions

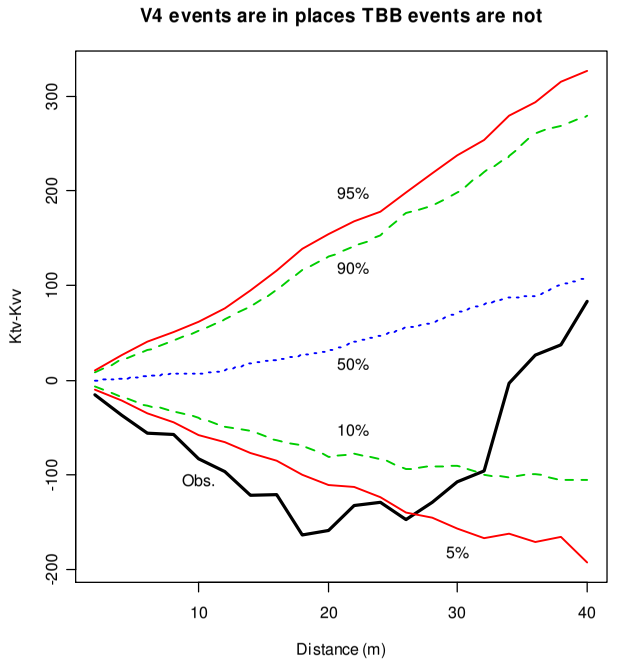
- Univariate K function: center circle on a location, count # additional events
  - 3 of these, All events. K(t)
  - V4 events to themselves. $K_{VV}(t)$
  - TBB events to themselves. $K_{TT}(t)$
- Bivariate K function: center circle on each V4 location, count number of TBB points in that circle. $K_{VT}(t) = K_{TV}(t)$

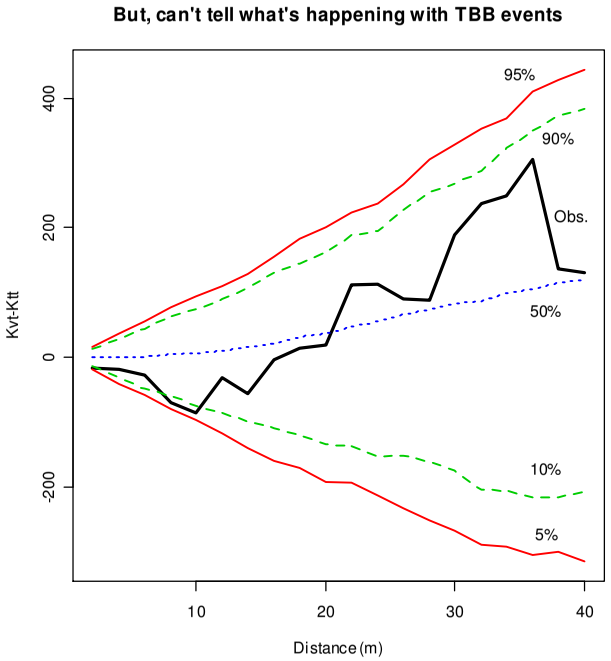
Notes

- Estimate each univariate K(t) by considering all locations, only V4 locations or only TBB locations
- Estimate bivariate K function by generalizing the estimate of K(t)
- Because of edge effects, estimated $K_{VT}(t)$ is not the same as the estimated $K_{TV}(t)$. Usually averaged.
- $K_{VT}(t)$ is sometimes called the cross-K function
- Two commonly used null hypotheses:
- Random labeling: Labels (TBB or V4) are randomly assigned to disease locations
- Under random labeling: $K_{VT}(t) = K_{TT}(t) = K_{VV}(t) = K(t)$,
- Independence: process generating TBB locations is independent of that generating V4 locations
- Under independence: $K_{VT}(t) = \pi t^2$
- Simulate random labeling by randomly assigning labels to observed disease locations
- Simulating independence is more difficult. Toroidal rotation is one possibility
- I used random labeling

Association between types

- Examine using differences of K functions
- Q: Are V4 events surrounded by (more, fewer) TBB events? $K_{VT}(t) - K_{VV}(t)$
- Q: Are TBB events surrounded by (more, fewer) V4 events? $K_{VT}(t) - K_{TT}(t)$:
- Null hypothesis:
  - points are randomly labeled,
  - both differences = 0

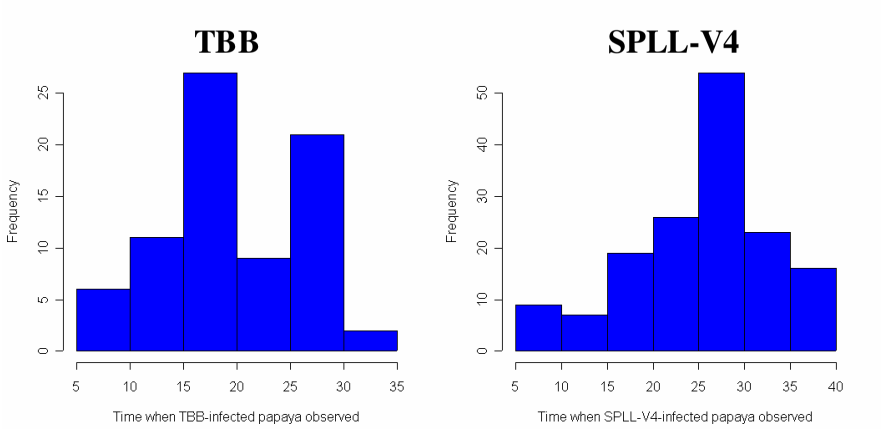




What about aggregation in time?

- Data includes the time an infection first noticed.
- So far, analyses have ignored time

- Is the number of newly infected trees constant over time?
- No.

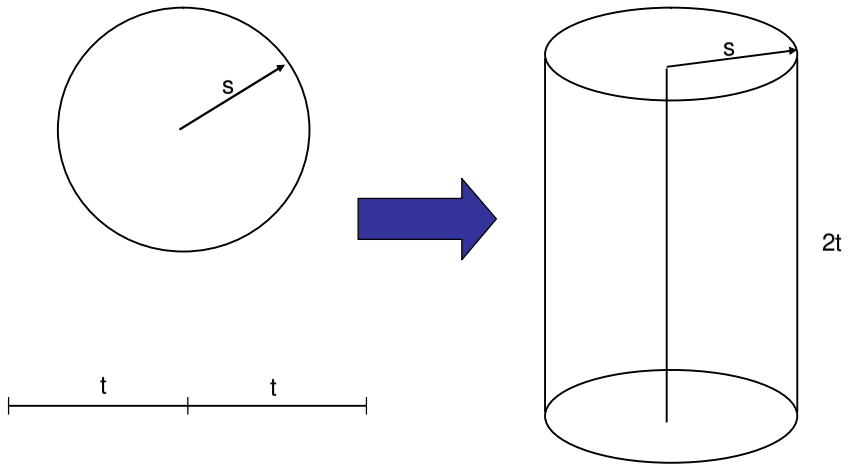


Evidence that there is a non-uniform frequency of the time when a papaya was found to be infected with either phytoplasma strain

Are space and time independent?

- We have shown Spatial Aggregation
  K(s) ≠ CSR
- And Temporal Aggregation
  K(t) ≠ CTR
- But are events close in space also close in time?
- Expect this if disease spreads by local transfer by insects.
- Define K(s,t) in terms of # events within s in space and t in time

Space-Time "Windows"



Courtesy: Dale Tessin, Depts. EEOB and Statistics, ISU

Space-time independence

- If space and time are independent,
  K(s,t) = K(s) K(t)
- Measure departure from independence by

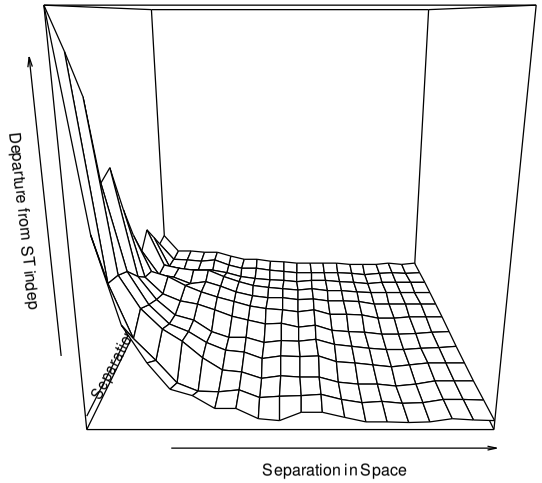
$$D_0 = \frac{K(s,t) - K(s)K(t)}{K(s)K(t)} = 0?$$

- $D_0 > 0$ when space-time contagion

Notes

- Division by K(s,t) is to equalize (at least approximately) the variance of K(s,t) – K(s) K(t)

D(s,t) to assess space-time independence



Notes

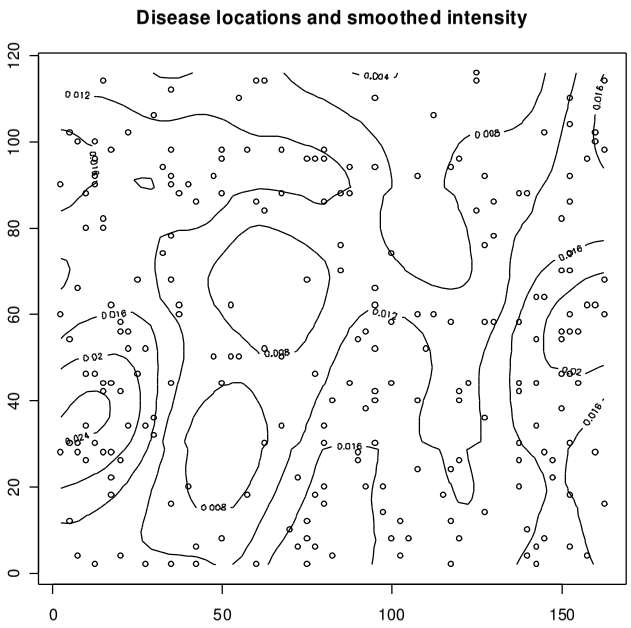
- (0,0) is the forward left corner
- D0 is >> 0 in the forward lefthand corner, that is for pairs of points close in space and close in time.
- Randomization test not shown, but results are highly significant. Space and time are not independent

What have we learned?

- Locations of diseased trees are clustered
  - Don't know whether contagion (infection) or consequence of environmental variation across plot
- Seem to be two processes,
  - one for each type (V4 and TBB)
- Rate of new infection varies over time
- Space-time assoc. suggests contagion
  - Events close in time tend to be close in space

Additional Materials

- Day-long short course on spatial point pattern analysis for ecologists on the web at:
- http://www.ci.uri.edu/projects/geostats/Theory.pdf concepts and examples
  - Large bibliography, emphasizing ecological applications at the end. Not updated since 2003.
- http://www.ci.uri.edu/projects/geostats/Hands-on.pdf computing using Splus (R is very similar)
- My favorite text is:
- Diggle, P.J. 2003. Statistical Analysis of Spatial Point Patterns, 2nd ed. Arnold / Oxford Univ. Press.



Spatial Statistical Software

by

Stephen L. Rathbun
Department of Health Administration, Biostatistics, and Epidemiology
College of Public Health
University of Georgia
Athens, GA 30605
rathbun@uga.edu

General Observations:

- Perhaps the single most limiting factor for dissemination a modern spatial statistical procedures is the limited availability of statistical software.
- Writing of statistical software involves the following trade-off:
  - Ease of use
  - Flexibility

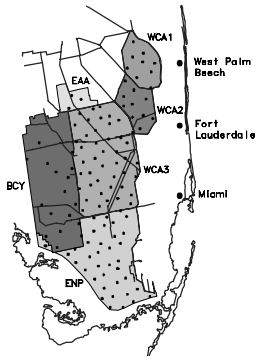
Outline: Review software for three areas of spatial statistics.

1. Geostatistics.
2. Spatial Point Patterns.
3. Lattice Data.

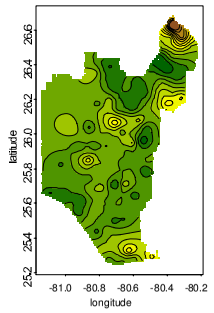
Geostatistics

South Florida Ecosystem Assessment.

Sample Sites



Predicted Total Mercury

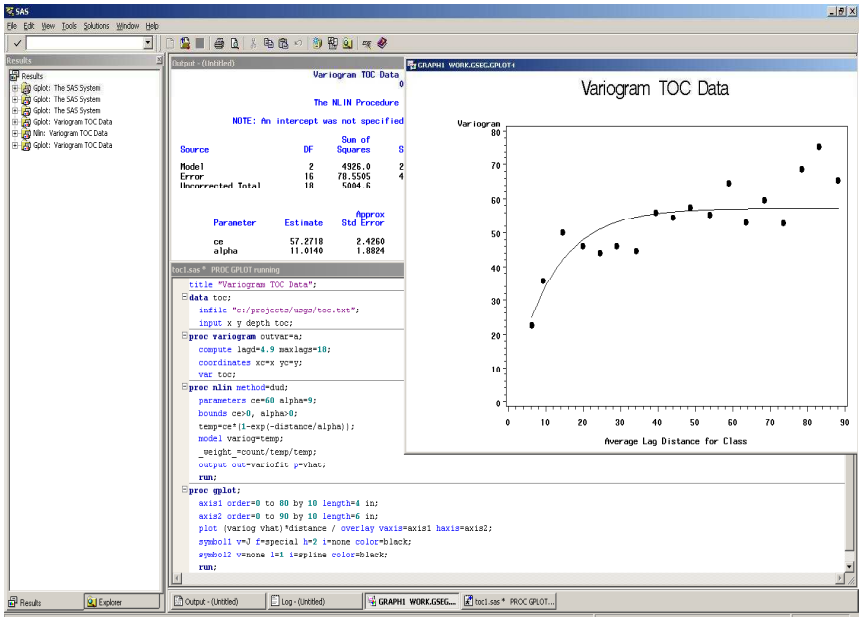


Geostatistical Software:

- SAS
- Surfer
- ArcGIS Geostatistical Analyst
- S+SpatialStats
- R

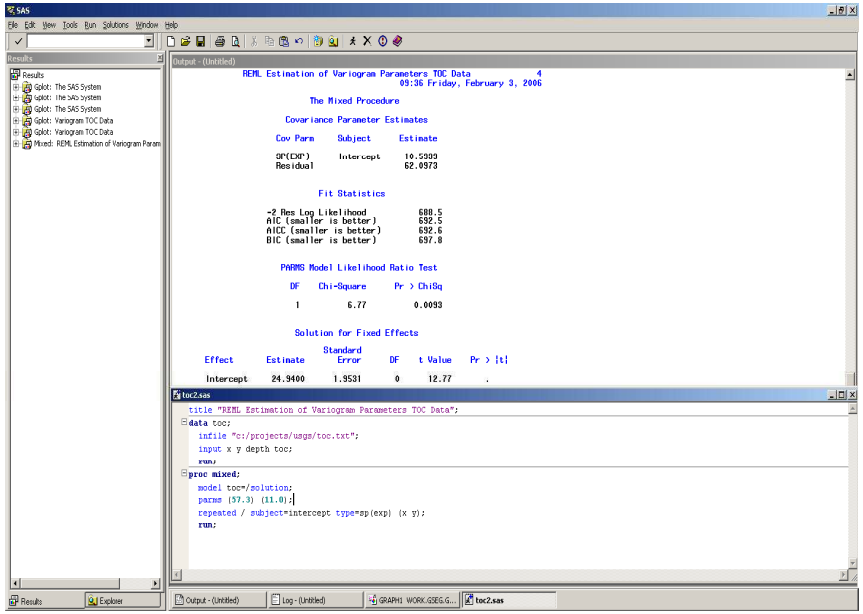
5

SAS: Variogram Model Fitting



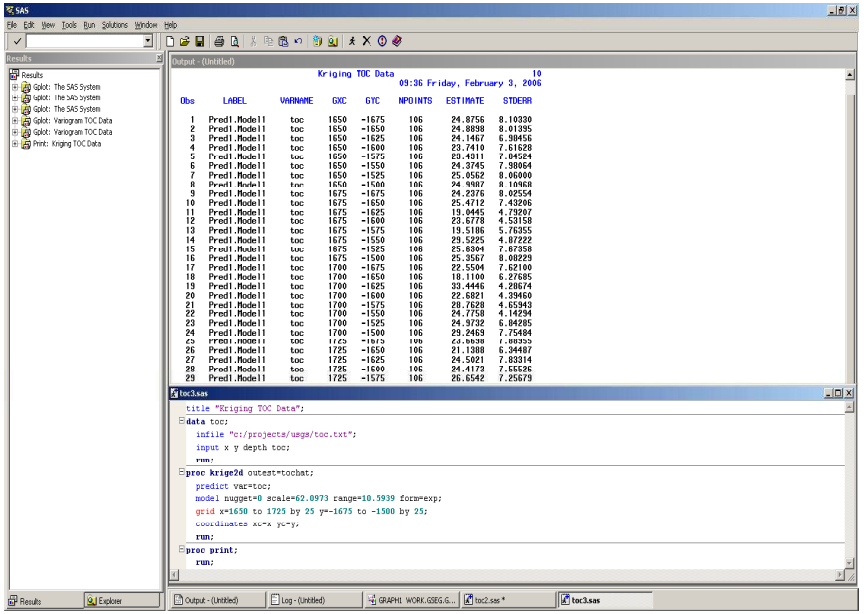
6

SAS: REML Estimation of Variogram Model Parameters



7

SAS: Kriging



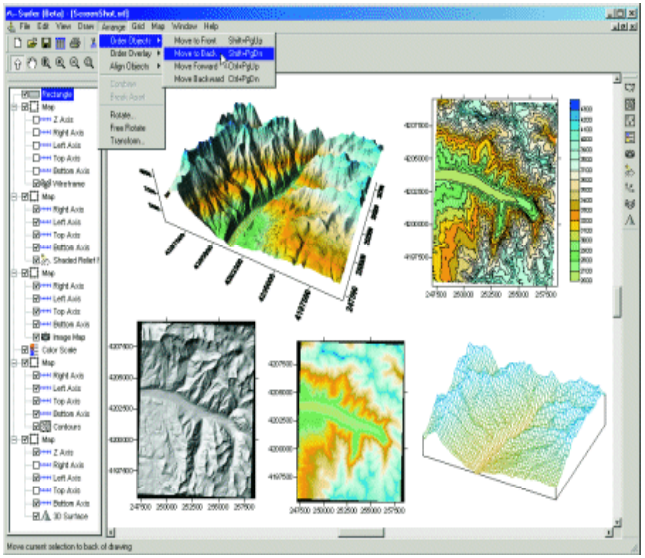
8

Comments: SAS Geostatistics

- SAS is not menu driven. Analysis is carried out by writing SAS programs in the SAS editor.
  - For those who have experience with SAS, the geostatistical procedures are easy to apply.
  - Harder to use than menu-driven software.
- SAS has procedures for:
  - Isotropic and anisotropic variogram estimation (proc variog);
  - Variogram model fitting:
    - Weighted Least Squares (proc nlin);
    - Maximum Likelihood and REML (proc mixed).
  - Ordinary Kriging (proc krige2d).
  - Universal Kriging (proc mixed).
  - Generalized Mixed Models (proc glimmix).
- Limitations:
  - Limited choice of variogram models.
  - Cannot draw good contour maps.

9

Surfer
http://www.goldensoftware.com/



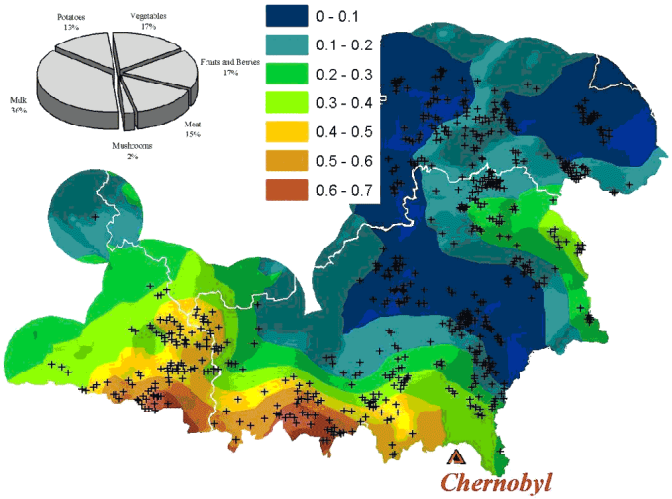
10

Comments: Surfer

- Menu Driven
- Surfer has procedures for:
  - Variogram Estimation;
  - Least Squares Estimation of Variogram Model Parameters
  - Wide Variety of Variogram Models: exponential, Gaussian, linear, log, power, quadratic, rational quadratic, spherical, wave, pentaspherical, cubic.
  - Ordinary Kriging
  - Excellent mapping capabilities: contour maps; 3D surface maps; wireframe maps; vector maps; shaded relief maps.
- Limitations:
  - Cannot fit Matern variogram;
  - Universal kriging not available.

11

ArcGIS Geostatistical Analyst
http://www.esri.com/software/arcgis/extensions/geostatistical/index



12

Comments: ArcGIS Geostatistical Analyst

- Menu Driven
- Geostatistical Analyst has procedures for:
  - Isotropic and anisotropic variogram estimation;
  - Least squares estimation of variogram parameters;
  - Wide variety of variogram models: circular, spherical, tetraspherical, pentaspherical, exponential, Gaussian, rational quadratic, hole effect, k-bessel, stable.
  - Variety of kriging methods:
    - Ordinary kriging
    - Universal kriging
    - Indicator kriging (Binary Variables)
    - Disjunctive kriging (Nonliner Geostatistics)
    - Cokriging (Multivariate Geostatistics)
  - Crossvalidation for model diagnostics.
- Limitation: Expensive ($2,500 for Geostatistical Analyst, $1,500 for ArcView 9.1)

13

Definition: Crossvalidation.

- Remove the data at site $\mathbf{s}_i$ from the data set;
  - Use the remaining data to obtain the kriging predictor $\hat{Z}_{-i}(\mathbf{s}_i)$ of the data at site $\mathbf{s}_i$
  - Compute the corresponding kriging variance $\sigma^2_{-i}(\mathbf{s}_i)$
- Repeat the above procedure for all sites.
- Compare observed values $Z(\mathbf{s}_i)$ with predicted values $\hat{Z}_{-i}(\mathbf{s}_i)$
  - Bias Measure

$$\mathrm{CV}_1 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Z(\mathbf{s}_i) - \hat{Z}_{-i}(\mathbf{s}_i)}{\sigma_{-i}(\mathbf{s}_i)}\right\}$$

  - Uncertainty Assessment

$$\mathrm{CV}_2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Z(\mathbf{s}_i) - \hat{Z}_{-i}(\mathbf{s}_i)}{\sigma_{-i}(\mathbf{s}_i)}\right\}^2$$

For a valid model, we should have

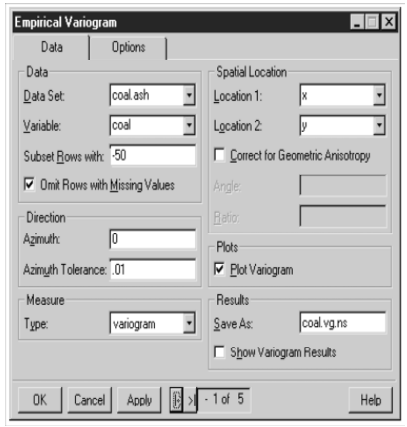
$$\mathrm{CV}_1 \cong 0 \text{ and } \mathrm{CV}_2 \cong 1$$

14

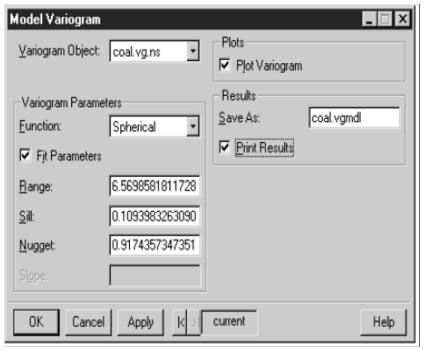
S+SPATIALSTATS

http://www.insightful.com/products/spatial/default.asp

Variogram Estimation



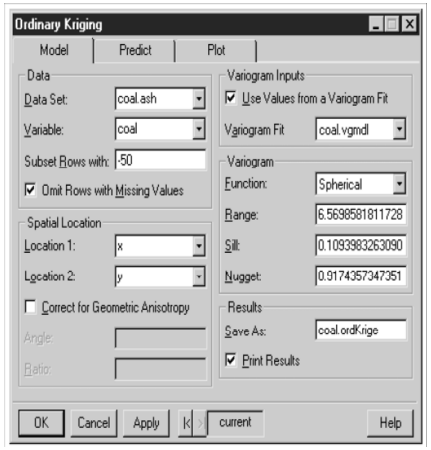
Least Squares Estimation



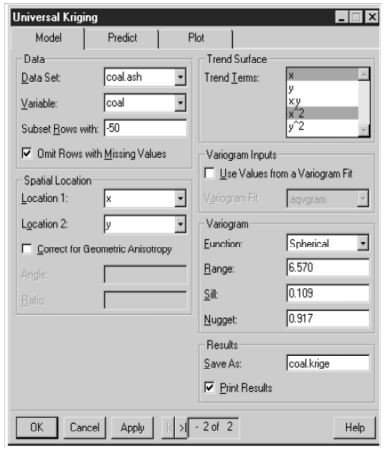
15

S+SPATIALSTATS

Ordinary Kriging



Universal Kriging



16

Comments: S+SPATIALSTATS

- Menu Driven
- S+SPATIALSTATS has procedures for:
  - Isotropic and Anisotropic Variogram Estimation
  - Least Squares Estimation of Variogram Parameters (Weighted least squares with some work)
  - Limited variogram models: Spherical, exponential, Gaussian
  - Ordinary and Universal Kriging
  - Good quality contour maps
- Software has not been kept up to date.
  - Effort has been made to improve user interface.
  - No effort has been made to include modern methods.

17

R

http://www.r-project.org/

Geostatistical Packages

- geoR http://www.est.ufpr.br/geoR/

  Frequentist and Bayesian geostatistics.
- geoRglm http://www.daimi.au.dk/~olefc/geoRglm/

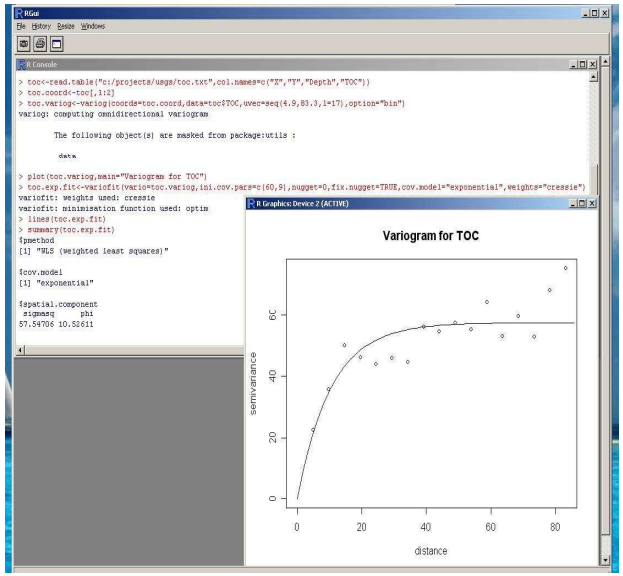
  Geostatistics for counts data. Poisson and binomial models.
- fields http://www.image.ucar.edu/GSP/Software/Fields/

  Best for global data. Includes great circle distance.
- gstat http://www.gstat.org/
- RandomFields http://www2.hsu-hh.de/schlath/R/RandomFields/RandomFields_doc.

  Spatial simulation.

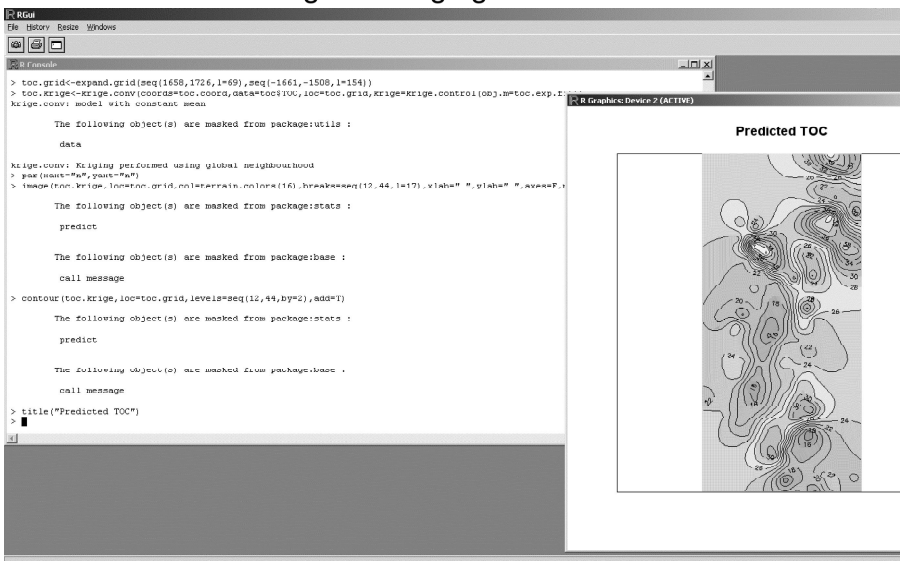
18

geoR: Variogram Modeling



19

geoR: Kriging



20

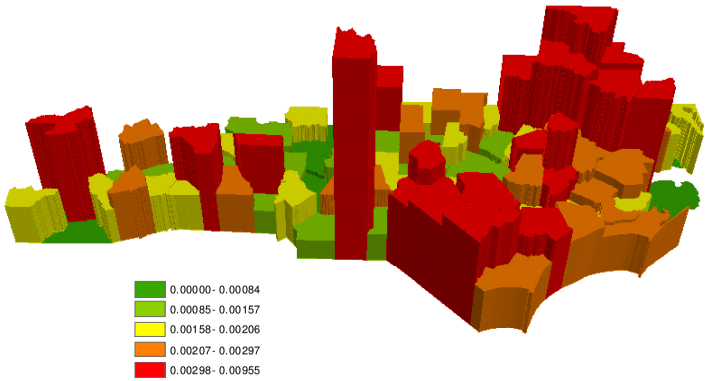
Comments: R

- Public domain software;
- Packages contributed by statistical researchers keep the software up to date;
- Command driven and interactive;
- GeoR has procedures for:
  - Variogram estimation;
  - Least squares, weighted least squares, REML estimation of variogram parameters;
  - Bayesian inference for model parameters;
  - Diverse variety of variogram models including the Matérn class;
  - Ordinary, universal and Bayesian kriging.
- GeoRglm has procedures for binomial and Poisson models for counts data;
- Fields includes great circle distance for investigating global data;
- Limitation: Not well documented.

21

Lattice Data

Sudden Infant Death Rates in North Carolina



22

Lattice Model Software

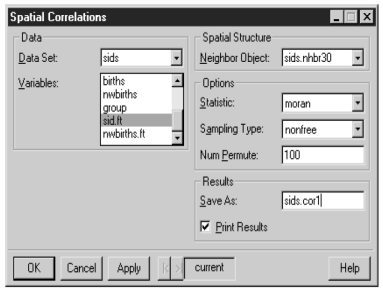
- S+SPATIALSTATS
- BUGS
- R package: spdep

23

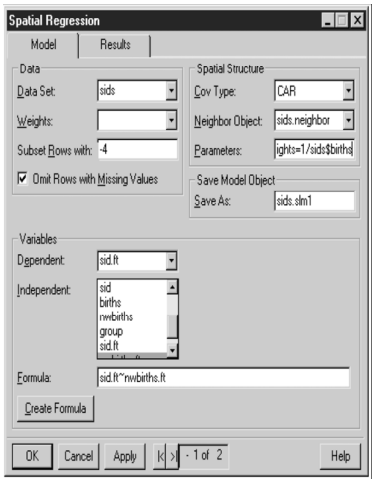
S+SPATIALSTATS

Lattice Models

Moran's Index



CAR Model



24

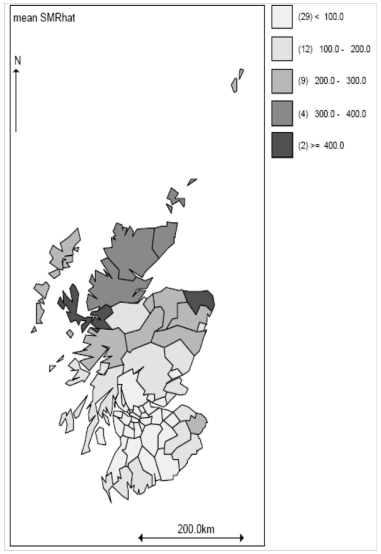
Comments: S+SPATIALSTATS

- Menu Driven
- S+SPATIALSTATS has procedures for:
  - Defining neighborhood matrices
  - Defining spatial weights matrices
  - Computing Moran's I
  - Fitting spatial regression models:
    - Conditional AutoRegressive
    - Simultaneous AutoRegressive
    - Moving Average

25

GeoBUGS

http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/geobugs.shtml



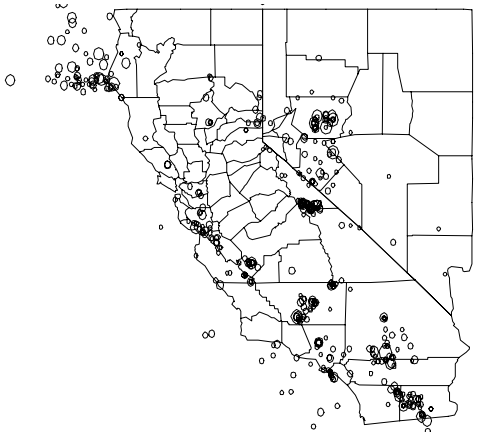
26

Comments: GeoBUGS

- Public domain software;
- Bayesian inference for lattice models:
  - CAR models
  - Poisson and binomial models with spatially dependent random effects.
- Data interface can use some work.

27

Spatial Point Pattern
California Earthquakes



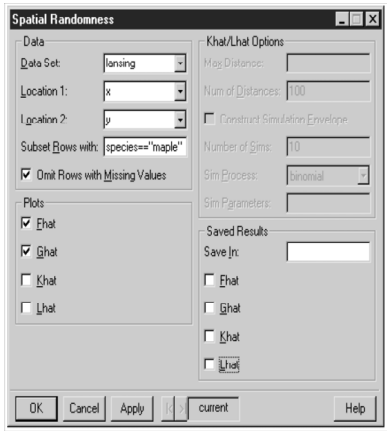
28

Point Pattern Software:

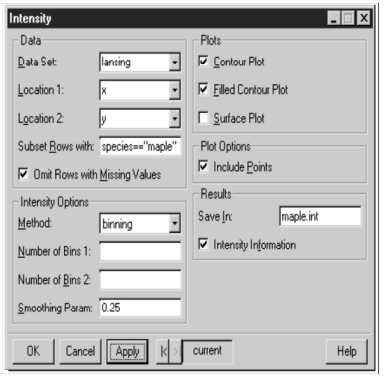
- S+SPATIALSTATS
- R

S+SPATIALSTATS

K-Function



Nonparametric Intensity



Comments: S+SPATIALSTATS

- Menu Driven;
- S+SPATIALSTATS has procedures for:
  - Computing F-, G- and K-functions;
  - Testing complete spatial randomness;
  - Nonparametric estimation of the intensity function;
  - Fitting the point cluster process model.

R: **Point Pattern Packages**:

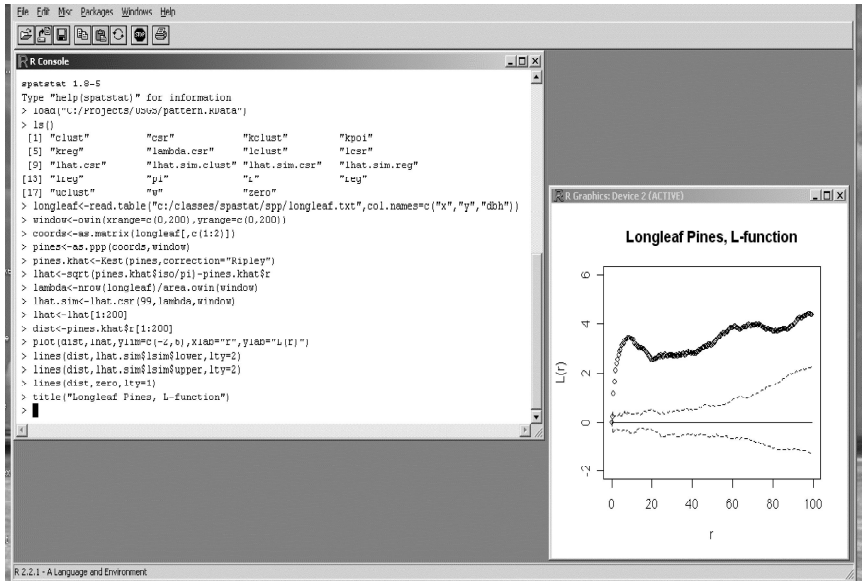
- spastat http://www.spatstat.org/

  Analysis of spatial point patterns.
- splancs http://www.maths.lancs.ac.uk/~rowlings/Splancs/

  Analysis of spatial and spatiotemporal point patternsn.
- MarkedPointProcess
  http://www2.hsu-hh.de/schlath/schlather.html#Software

  Analysis of marked point patterns.

R: spastat



Comments: R

- Public domain software;
- Packages contributed by statistical researchers keep the software up to date;
- Command driven and interactive;
- Spastat has procedures for:
  - Computing F-, G- and K-functions;
  - Testing complete spatial randomness;
  - Fitting the point cluster process model;
  - Simulating a variety of point process models;
  - Estimating parameters of modulated Poisson process model (covariates must be observed at all locations).

General Summary

- ArcGIS Geostatistical Analyst:
  - Menu driven;
  - A comprehensive collection of geostatistical methods;
  - Expensive.
- R:
  - Up-to-date methods for geostatistical and point pattern analyses;
  - Public domain;
  - Command driven and interactive.
- S+SPATIALSTATS:
  - Best for analysis of lattice data;
  - Menu driven.

Combining multi-scale spatial data

Prepared for the

Workshop on Spatial Statistics
For Agricultural and Environmental Applications

by

Mark West
USDA/ARS/NPA

Combining spatial data from different sources is a common problem and poses a real challenge. (Gotway C.A. 2002) provides an overview of the most recent approaches and progress made towards combining incompatible spatial data. What is covered in this presentation is narrowly focused on available geostatistical approaches which are challenging enough. By illustrating a few examples, my aim is to introduce basic concepts and terminology used in geostatisitical analysis and expose the limitations of these methods.

Gotway C.A., Y. L. J. (2002). "Combining Incompatible Spatial Data." Journal of the American Statistical Association **97**(458): 632-648(17).

What are multi-scale data?

- multiple sources
- collected from the same region using different formats and scales
- each source (layer) may have one or more attributes (weed infestation, percent bare ground)
- sources may have different levels of accuracy and precision

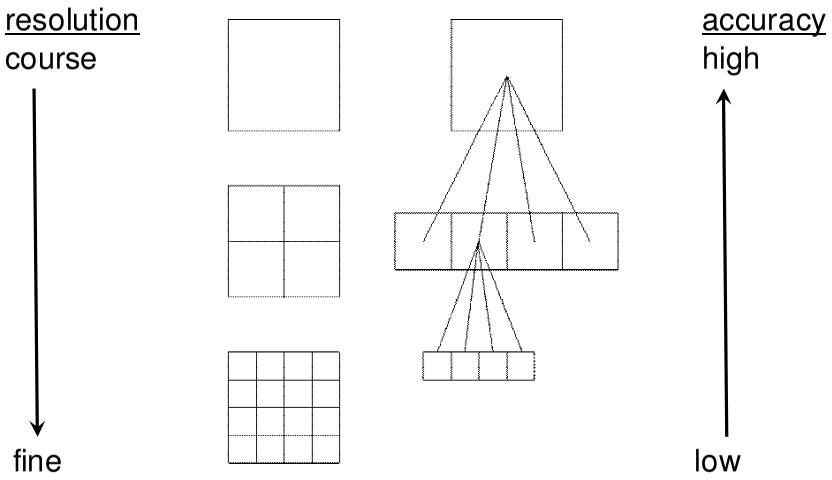
Multi-scale data may include different formats such as points, lines, polygons and grids. Edzer Pabesma (Pebesma 2004) has developed the **sp** and the **gstat** packages with R code for analyzing different types of spatial data.

(Zhu, Morgan et al. 2004) combine soil coring, penetrometer, and other topographic data to produce a fine map of depth-to-till for a Wisconson field. The data collected using these methods have different resolutions and accuracies. Soil coring provides accurate information on depth-to-till but because of its expense requires this information to be collected sparingly and hence results in a low resolution map of a field. Soil electroconductivity (EC) can provide information on depth-to-till and is easy to collect hence its resolution will be finer than the information collected from soil coring but has the problem of being less accurate with more error. The soil core and soil EC are multi-scale data.

Pebesma, E. J. (2004). "Multivariable geostatistics in S: the gstat package." Computers & Geosciences **30**(7): 683.

Zhu, J., C. L. S. Morgan, et al. (2004). "Combined mapping of soil properties using a multi-scale tree-structured spatial model." Geoderma **118**(3-4): 321.

Layers, Resolution & Accuracy



This slide was used to visually support the concept that resolution and accuracy aren't necessarily one in the same. Certainly we want to create maps that have fine resolution and are accurate. However, data that are easier to collect for providing finer maps are easier because the methodology used to collect them is quick, inexpensive and prone to error.

General Problem

- Inference at the level of one layer may be desired using information gathered from other levels
- Question may be "How do soil attributes measured at point locations relate to weed infestation measured on rectangular units?"

(Gotway 2002) gives several examples where data is on one scale but inference is desired at another. Individual level inference is wanted but because of privacy issues data is only available at some aggregate level. Data from Standard Metropolitan Statistical areas may be available but information at the county level may be needed.

Gotway, C. A. w. Young, L.J. (2002). "Combining Incompatible Spatial Data." Journal of the American Statistical Association **97**(458): 632-648(17).

Focus

- Use topographic and soil attributes to predict crop yield
- Predict at aggregate levels
- Mapping applications
- Use available geostatistical (kriging) methods
  - R programs
    - Packages (Edzer Pabesma, 2005)
      - **gstat**
      - **sp**

The focus of this presentation is to examine a few geostatistical methods (mainly kriging systems) that involve problems with combining misaligned spatial data. Definitions of terms related to the problems will also be given.

My intention is to provide information of available tools that can be used to krige data. These are available for free from the R Development Core Team (2005) . All of the kriging systems were fitted using the gstat package (Pebesma 2004) . There is a variety of example code for fitting similar systems in the gstat package. You need to download the package and once downloaded refer to the directory C:\Program Files\R\R-2.2.1\library\gstat\demo for example code. I found these scripts very helpful.

Pebesma, E. J. (2004). "Multivariable geostatistics in S: the gstat package." Computers & Geosciences **30**(7): 683.

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Example 1: Field scale study relating elevation to yield

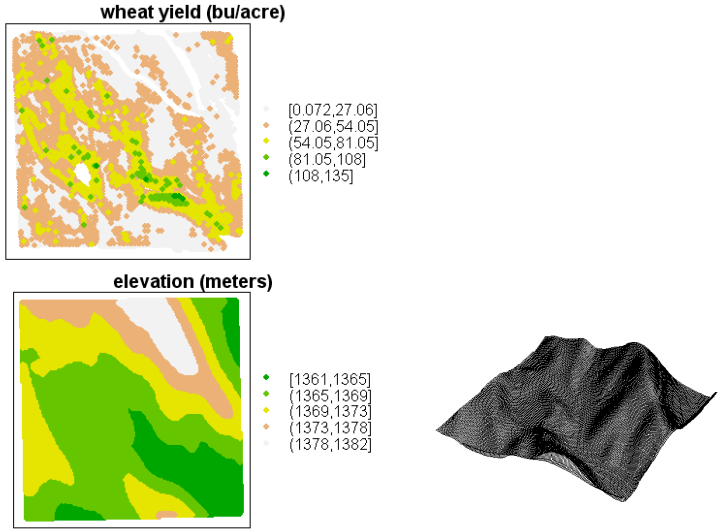
Models were fit that use elevation to predict spatial crop yield values. (Green & Erskine, 2004)

Can elevation help predict yield for large plots (blocks) in the field?

(Green 2004) addresses quantification of spatial variability of crop yield and soil water at farm scales using geostatistical and fractal analyses. His data are used in this example to demonstrate kriging methods for predicting wheat yield at the particular Nothern Colorado farm.

Timothy R. Green, R. H. Erskine. (2004). "Measurement, scaling, and topographic analyses of spatial crop yield and soil water content." Hydrological Processes **18**(8): 1447-1465.

Yield and elevation maps



Yield data (Green 2004) was collected from a field of roughly 800 square meters area using a combine mounted with a calibrated monitoring device to measure yield in bushels per acre. A GPS system was used to mark the 6701 points on the field where yield was recorded. Each yield value represents around 10 square meters of area. Because of various factors (e.g. the combine will not move at a regular speed) the actual area represented at each point will vary and yield values can be expected to be quite noisy.

Elevation data was collected over the same area using an all terrain vehicle. These data were interpolated to a regular grid of 5 meter spacing. These interpolated values of elevation should be very accurate representing small deviations from the true elevation (~ 0.05 m).

Yield and elevation maps were generated using the **sp** (Pebesma 2005) package in R.. See 'wheat yield and elevation plots.R' in the wheat folder on the ftp site.

Green, T. R., Erskine, Robert H. (2004). "Measurement, scaling, and topographic analyses of spatial crop yield and soil water content." Hydrological Processes **18**(8): 1447-1465.

Pebesma, E. J., Bivand, Roger S. (2005). S Classes and Methods for Spatial Data:the sp Package.

Example 2: Precision agriculture

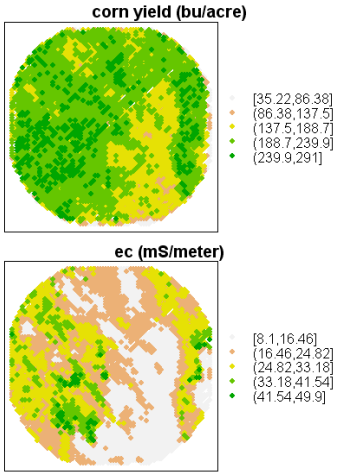
Soil electroconductivity (EC) mapping to explain yield variability for a center pivot cropping system in Northern Colorado.

Can EC help predict yield for large blocks in the center pivot system?

The Water Management Research Unit in Fort Collins develops irrigation, agricultural chemical, and other management practices that protect water quality for all Americans while improving the husbandry of natural resources and the irrigator's economic viability. Research covers precision farming with center pivot sprinklers, remote sensing, and weed management for reduced applications of chemicals.

The data given here was collected in 1999 for relating various soil properties with soil electroconductivity (EC). Yield data was collected in 1999. Each point roughly represents around 11 to 12 sq. meters (the swath length is 20 ft.; the distance between points is around 6 ft).

Yield and electroconductivity (EC) maps



This data is similar the wheat and elevation data. The data given here was collected in 1999 for relating various soil properties with soil electroconductivity (EC). Yield values represent around 11 to 12 sq. meters. The EC data are measured can be expected to be much noisier than the elevation data.

Yield and EC maps were generated using the **sp** (Pebesma 2005) package in R.. See 'corn yield and EC plots.R' in the corn folder on the ftp site.

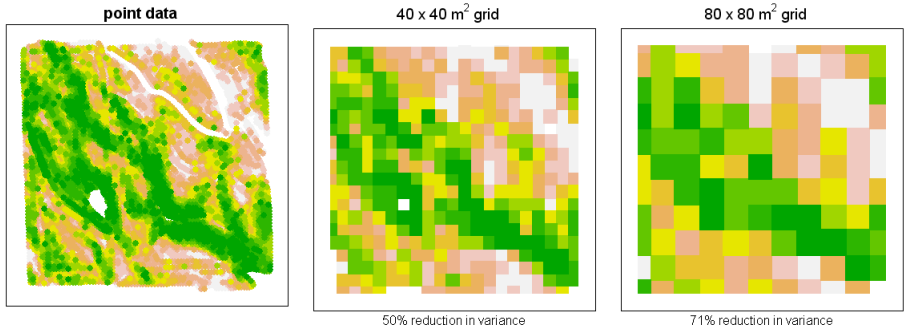
Pebesma, E. J., Bivand, Roger S. (2005). S Classes and Methods for Spatial Data:the sp Package.

Methods

- Aggregation
- Change of Support (COSP)
- Kriging
  - Point Kriging
  - Cokriging models
  - Block kriging
- Spatial joins

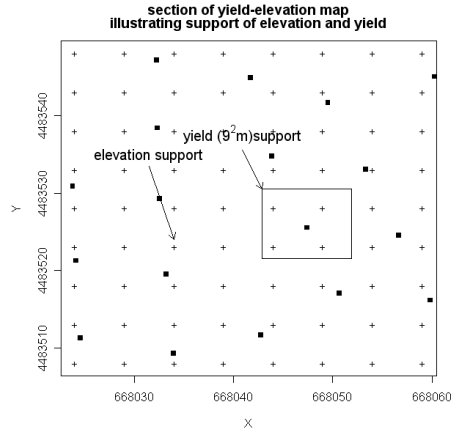
Methods used to combine multi-scale spatial data include aggregation, various kriging methods and those that involve what is referred to as a change of support (block kriging). We may be interested in changing from a point system to a system of blocks, from a system of blocks to a system of points., or from a system of blocks to another system of blocks.

Aggregation –
averaging over point values to form areal units



It is a well-known fact of statistics that averaging reduces variance. The apparent spatial variation also changes with aggregation.

Support of data



This slide is provided to point out data can be recorded at points but may have areal support. The yield data in both examples are geo-referenced at points but because the grain collected by the combine is collected over a region it represents yield over some small area (around 10 square meters for the wheat yield and around 11 square meters for the corn yield).

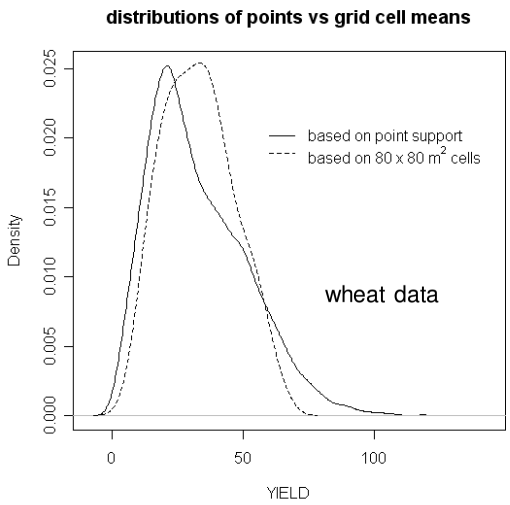
Why Aggregation?

- **Prediction:** prediction is wanted at larger scales
- **Different support:** aggregation transforms a variable from point support to areal support.
- **Smoothing:** aggregation smoothes out noise to detect trends

It isn't uncommon where data is collected at one scale and inference is desired at another. Making inferences on block averages whose support is different from those of the data is called a change of support problem. (Isaaks and Srivastava 1989) give an example of a mining operation where data are collected at points but mining operations involve only large blocks of material extracted from the mine. Having only point data on hand the problem here is to estimate the distribution of the average tonnage of ore contained in blocks. To estimate this sampled point data need to be aggregated to the size of blocks and the distribution of values associated with blocks may then be used to base decisions. By aggregation we mean obtaining a weighted average. To estimate the average tonnage of ore Z_B for a block B we need to come up with an estimate based on sampled values Z_i in the neighborhood of block. The estimator $\hat{Z}_B = \sum_i \lambda_i Z_i$ is derived by choosing weights λ_i that account for the spatial variation in the Z_i and the estimated spatial variation occurring on the block scale. Spatial variation for the change of support problem is modeled through a variogram $\gamma(h)$ of the Z_i.

Isaaks, E. H. and R. M. Srivastava (1989). Applied geostatistics. New York, Oxford University Press.

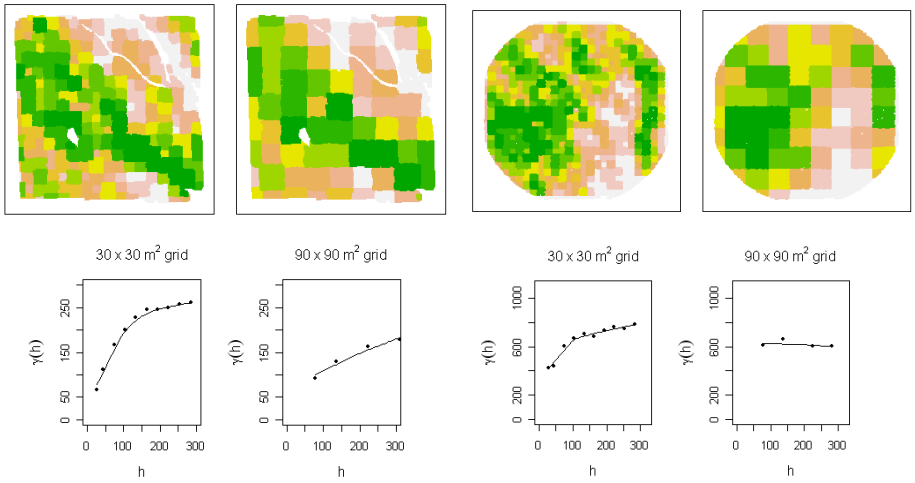
Support effect



The support effect is the change of distribution of statistics that results when data are aggregated. Quoting from (Gotway 2002), 'Changing the support of a variable (typically by averaging or aggregating) creates a new variable. This new variable is related to the original one, but has different statistical and spatial properties,'

Gotway, C. A. w. Y., L.J.) (2002). "Combining Incompatible Spatial Data." Journal of the American Statistical Association **97**(458): 632-648(17).

Scale Problem



The effect of the change of support realized by increasing the areas for which aggregation is performed is called the scale effect. The differences in the statistical properties of the variograms are obvious with scale changes for both examples.

Problems of Aggregation

- Change of support problem (COSP)
  - How can spatial variation at the point support scale be used to estimate spatial variation at an aggregate scale?
  - COSP modeled through variogram
  - Similar to using population variance to form inferences using sample means.

$$\sigma_y^2 \rightarrow \sigma_{\bar{y}}^2 = \sigma_y^2 / n$$

Making inferences on block averages whose support is different from those of the data is called a change of support problem. (Isaaks and Srivastava 1989) give an example of a mining operation where data are collected at points but mining operations involve only large blocks of material extracted from the mine. Having only sampled point data available a big problem is to estimate the distribution of the average tonnage of ore contained in blocks.

Isaaks, E. H. and R. M. Srivastava (1989). Applied geostatistics. New York, Oxford University Press.

Aggregation Methods

- **Generic** : weighted average of values Z_i for estimating average for an area B of size |B|

  $\hat{Z}_B = \sum_i \lambda_{B_i} Z_i, \ \sum_i \lambda_{B_i} = 1$
- **Arithmetic means:** simple averages ($\lambda_{B_i} = 1/n$, n= sample size) ignore spatial structure
- **Kriging:** averages use weights λ_{B_i} derived from spatial structure $\gamma(h)$ - variogram

By aggregation we mean obtaining a weighted average. To estimate the average Z_B of a variable Z for a block B we need to come up with a weighted average $\hat{Z}_B = \sum_i \lambda_{B_i} Z_i$ based on sampled values Z_i in the block neighborhood. Later more detail will be given when the method of block kriging is described.

Some Notation

- S = point where an observation is made
- Z(S) = value of observation at S
- $\delta(\mathbf{s})$ = error from mean value at S
- μ = mean value at for any S
- $\mu(\mathbf{s})$= mean value that depends on location S and/or predictors at S

Notation added to clarify expressions to follow.

Ordinary and Universal Kriging

Ordinary Kriging

- Model: $Z(\mathbf{s}) = \mu + \delta(\mathbf{s})$

- Universal Kriging
  - Model: $Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s})$

- Predictor: $Z(S) = \sum_i \lambda_i \cdot Z(S_i)$
  - λ_i **weight of ith value, derived from variogram of $\delta(S)$ and/or predictors**

Kriging is a method of spatial prediction. The predictors are in the form of a weighted average $\hat{Z} = \sum_i \lambda_i \cdot Z_i$. The differences in these two kriging methods are their underlying models.

For ordinary kriging, the underlying model for the Z is a constant mean plus error where errors are spatially autocorrelated. The spatial autocorrelation of errors doesn't depend on location. The λ_i are derived using the model assumptions to give the minimum mean-squared prediction error. For ordinary kriging, the λ_i are a function of the variogram $\gamma(h)$ that describes the autocorrelation of errors.

For universal kriging, the underlying model for the Z is a mean that depends on location and/or other predictor variables plus error where the errors are spatially autocorrelated. Again, the spatial autocorrelation of errors doesn't depend on location. For universal kriging, the λ_i are a function of the variogram $\gamma(h)$ that describes the autocorrelation of errors and the predictors that are modeling the mean.

Cokriging

- Simultaneously krige two or more variables

  $Z(s) = \sum_i \lambda_i \cdot Z(s_i) + \sum_j \omega_j \cdot X(u_j)$ **$Z(s_i)$ yield at locations s_i**

  $X(u_j)$ EC at locations u_j

- Not only requires fitting of variograms for each variable but also requires fitting of the cross-variogram for each pair of variables

Cokriging is a method originating from the need for predicting a primary variable Z that is undersampled (because it may be expensive to sample) but another secondary variable X is available that is related to Z and more heavily sampled (because X it is less expensive/difficult to sample). Both X(S) and Z(S) are fitted to a model simultaneously. This is a form of multivariate prediction modeling. The estimator for an unknown Z is of the form $\hat{Z} = \sum_i \lambda_i \cdot Z_i + \sum_j \omega_j \cdot X_j$. The usefulness of the secondary variable for predicting the primary variable is enhanced when the primary is undersampled. See (Isaaks and Srivastava 1989) for a more complete description.

Isaaks, E. H. and R. M. Srivastava (1989). Applied geostatistics. New York, Oxford University Press.

Block Kriging

- estimate the mean value of an attribute for a local area **B** using points in the neighborhood of **B**
- used with either ordinary, universal or cokriging
- variogram is adjusted to handle the scale effect
- estimator: $\hat{\mathbf{Z}}(\mathbf{B}) = \sum_i \lambda_{B_i} \cdot \mathbf{Z}(\mathbf{s}_i) \approx \frac{1}{|B|}\int_B Z(s)ds$

Block kriging is an aggregation method for estimating or predicting an average value $\hat{Z}_B$ over an area B. The estimator $\hat{Z}_B = \sum_i \lambda_{B_i} Z_i$ is derived by choosing weights λ_{B_i} that account for the spatial autocorrelation in the Z_i and the estimated spatial autocorrelation occurring on the block scale. Therefore we need to know how the autocorrelation among units on the point scale changes to autocorrelation among units on the block scale. Spatial variation for the change of support problem is modeled through a variogram $\gamma(h)$ of the Z_i.(Cressie 1993) describes the needed calculations to modify the point support variogram $\gamma(h)$ to the block support variogram $\gamma(B)$ (pages 124-125,) for block kriging (aggregation over an area B).

Cressie, N. A. C. (1993). Statistics for spatial data. New York, J. Wiley.

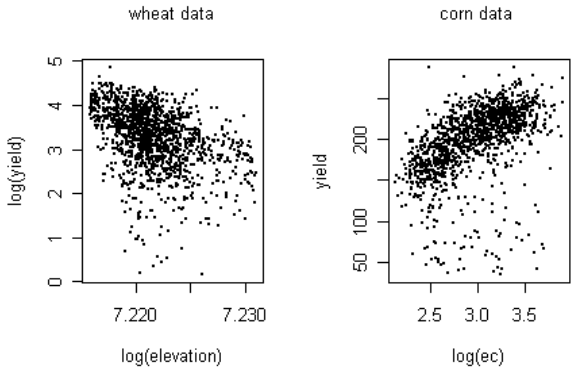
Spatial Join of datasets

- Combine two or more datasets with different attributes measured at different locations by translating them to same location.
- Problems
  - Trans-locating errors in variables problem
  - Ad Hoc approach, descriptive purposes
- Benefits
  - plotting techniques reveal relationships and needed transformations for other more legitimate methods
  - attributes can be studied together

Data measured at different locations can be joined many different ways. Consider two geostatistical datasets A and B each with different attributes. One approach would be to conduct a search of points in dataset B for each data point in dataset A. The attribute values corresponding to the points in B nearest in distance to those in A are joined with those of A. Another approach would be to lay a grid over the intersection of the areas from which datasets are formed. For each point in the grid, a search is conducted to find the points in A and in B that are closest and these two are joined. Yet another way would be to spatially interpolate all the points in B to those in A. Many Geographic Information Systems provide software for joining misaligned data but the capabilities of the software is limited to descriptive purposes.

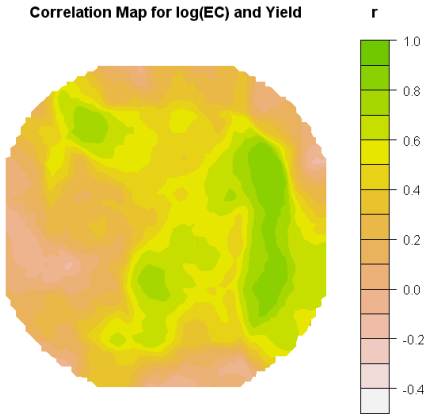
See **joindata.R** for an R program that quickly joins two datasets.

Scatterplots of spatially joined data



When kriging methods involve predictors, they are linear functions of the predictors. Spatial joins were used to construct scatterplots to study the relationships between yield and predictors to find transformations that ensure linearity. In both cases, suitable transformations were found.

A spatial join may be used to see if relationships are stationary



After creating a spatial join, Pearson's r was calculated as a moving window statistic to consider how the relationship between EC and YIELD may change throughout the region

Using spatially joined data from the center pivot example, Pearson's correlation coefficient was used as a 'moving-window' statistic. A 12x12 m grid was overlaid on the center-pivot area. For each point on the resulting grid, all points in the spatially joined dataset within a 100 m radius were selected and the correlation coefficient was computed, and then mapped.

(Carroll and Oliver 2005) give details of this technique in their study of EC and soil properties.

Carroll, Z. L. and M. A. Oliver (2005). "Exploring the spatial relations between soil physical properties and apparent electrical conductivity." Geoderma **128**(3-4): 354.

Attribute characteristics

- Wheat Yield / Elevation
  - Response: **Yield**
  - Predictor: **Elevation, measured with little error**
  - location: measured at different locations
- Corn Yield / EC
  - Response: **Yield**
  - Predictor: **EC, measured** with a lot of error
  - location: measured at different locations

Before looking at the specific problems for using elevation or EC to help predict yield using kriging methods, this slide is given to motivate why different approaches for predicting yield are taken

For the wheat yield data, elevation has so little error associated with it that it is practical to treat these values as being static. There will be little error incurred by interpolating elevation values to points where wheat yield is observed. Doing this we act as if both yield and elevation are measured at the same points in the field. I view this problem as a univariate regression problem where the predictor, elevation, is known for any point in the field

For the corn yield data, soil EC has a lot of error associated with it to begin with. Interpolating EC value to points where corn yield is observed will add more error. For practical as well as illustrative reasons corn yield and soil EC joined to the same points for analysis. I view this problem differently in that it lends itself to a cokriging application.

Suggested approach using elevation to help predict yield

- Universal Kriging
  - interpolate values of the elevation to locations where yield is recorded
  - use elevation as a predictor

Using the example datasets, two approaches are considered for incorporating the predictor variables with kriging methods.

For the wheat yield example, universal kriging will seems to be a reasonable approach for predicting yield using the model $Z(S) = \mu(S) + \beta \cdot X(S) + \delta(S)$ where $Z(S) =$ log(Yield) at location S, $\mu(S)$ is the mean value of log(Yield) at location S, $X(S) =$ interpolated value of log(Yield) at location S and $\delta(S)$ is the error at location S. Although elevation is not observed at the same locations as yield, interpolating elevation to those points where yield is measured should incur little error.

For the corn yield example, universal cokriging will be used on the basis that a reasonable model for predicting yield is $\underline{Z}(S) = \underline{\mu}(S) + \underline{\delta}(S)$ where $\underline{Z}(S)$ represents the bivariate values of both yield and EC at location S, $\underline{\mu}(S)$ represents the mean of the bivariate values at location S and $\underline{\delta}(S)$ represents the bivariate errors at location S.

Predicting corn yield
Can EC help prediction?

- Hold out thirty 40 x 40 m^2 grid cells for comparison of prediction methods
- Use remaining data to fit prediction models; **ordinary kriging**, **universal kriging**, and **universal cokriging**
- Block krige yield to 40 x 40 m^2 grid using each method & obtain standard errors

Steps used to compare kriging methods for the corn yield data are similar to that of the wheat yield only yield data and EC data weren't spatially joined.

The usefulness of universal cokriging for predicting blocks of unsampled plots was tested by holding out a set of thirty randomly selected 40 x 40 m^2 blocks, predicting their average values and comparing them back to the actual means for those blocks. For comparison purposes, ordinary kriging and universal kriging using locations were included. Using each kriging method, yield was block kriged to predict average values for the 40 x 40 m^2 held out blocks. The abilities of these methods for prediction were evaluated by comparing r^2. Standard errors of the estimates were also compared.

The spatial autocorrelation structure of the errors for the model was fitted to Gausian variogram models. For universal kriging, residuals were obtained by fitting a trend surface of yield over the field. The residuals were then used to obtain an empirical variogram. The empirical variogram was fitted to a Gaussian variogram model by least squares.

Kriging approach using EC to help predict yield values on blocks

- Cokriging
  - Simultaneously krige both yield and EC
  - Fit linear model of coregionalization (LMC)
    - a method for fitting variograms for yield and EC and the cross-covariogram of EC and yield

$$Z(s_0)=\sum_i \lambda_i \cdot Z(s_i) + \sum_j \omega_j \cdot X(u_j)$$

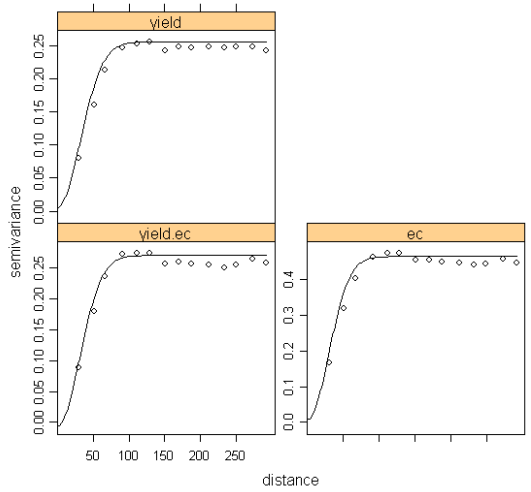
$Z(s_i)$ yield at locations s_i

$X(u_j)$ predictor at locations u_j

I chose to explore the ability of cokriging yield and EC because I felt this to be a reasonable application to cokriging. I felt this to be a more reasonable application than cokriging yield and elevation since the error in elevation was expected to be very small. A linear model of coregionalization was used to fit the variograms and cross-variogram for yield and EC. The method is described in (Isaaks and Srivastava 1989) and as a word of warning can be quite an undertaking.

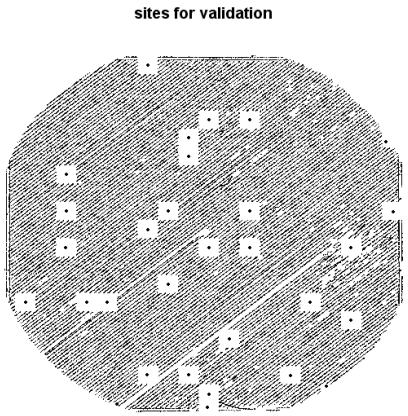
Isaaks, E. H. and R. M. Srivastava (1989). Applied geostatistics. New York, Oxford University Press.

Fitted cross-variogram for cokriging



Cokriging requires fitting of the cross-variograms(semivariograms) as well as the variograms.

Validation Sites



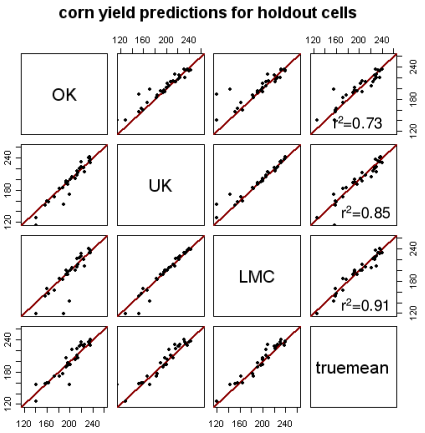
Random samples of thirty 40 x 40 m2 blocks of data were held out in order to compare methods. The plot shows the locations of these plots.

Compare methods

- OK: Ordinary kriging
- LMC: cokriging with linear model of coregionalization
- UK: Universal kriging
- True Mean: Observed average yield

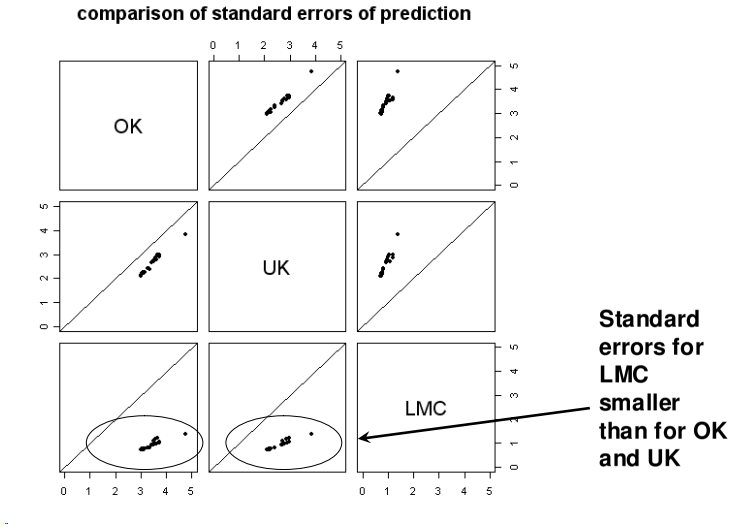
To follow methods of prediction, abbreviations are made. Kriging methods compared for predicting the thirty held out plots in the field are; **OK** – ordinary kriging; **UK** – universal kriging using locations as predictors; **LMC** – universal cokriging of yield and EC using a method of fitting a cross-covariogram called linear model of coregionalization; **truemean** – is used to denote the actual sample average observed for the held out plots. See (Isaaks and Srivastava 1989) for details of the linear model of coregionalization.

Compare predictions



The scatterplot matrix is used to make comparisons among the predictions for the means of the holdout plots. Each of the methods compared are weighted averages of values at points in the neighborhood of the plot being estimated. All of the estimates agree well with the true sample averages for the plots being estimated. r^2 is calculated by squaring the correlation coefficient computed between the true sample average for each of the thirty holdout plots and the corresponding prediction values obtained by each method: OK – ordinary kriging; UK – universal kriging using locations as predictors; LMC – universal cokriging of yield and EC using a method of fitting a cross-covariogram called linear model of coregionalization .See (Isaaks and Srivastava 1989) for further details. Based on the r^2, there appears to be some increased ability of prediction over ordinary kriging using the universal and cokriging methods. However, this may be due to the search region of points used for making the predictions.

Compare standard errors of methods used



Ordering the methods from largest to smallest on the basis of size of standard errors, ordinary kriging is showing the largest standard errors, then the next largest is for universal kriging and the smallest is for universal cokriging. Intuition would lead you to think this would be the order relation for these standard errors. The more information used would lead to more precise estimates. However, I am wary of these estimates because of the personal choices I made in the fitting of the variograms used for developing these universal kriging predictors. As a check I computed approximate 95% confidence intervals by calculating estimate +/- 2 standard errors, and found the proportion of intervals that cover the true means to be 3% with the LMC. An adjustment to the variogram estimates brought this coverage up to 100% with the standard errors still smaller than those for the ordinary kriging method.

Summary

- kriging is spatial prediction tool that uses weighted averages
  - weights depend on autocorrelation structure
  - explanatory variables may adjust the weights for a more accurate prediction - cokriging or universal kriging
- cokriging and universal kriging are ways to incorporate multi-scale data
- aggregation methods compared here give similar predictions but some accuracy and precision may improve with predictors
- block kriging is a useful scaling tool
- joining misaligned spatial data may be useful as a exploratory/descriptive tool
- examples given involve heavily sampled spatial regions– with less heavily sampled data a predictor may have bigger impact
- focus was on geostatistical methods, newer Bayesian methods namely tree-structured hierachical models may be more effective (Zhu 2004, 2005)

To summarize, geostatistical methods covered here mostly revolve around mapping applications. Although difficult, different sources of data can be combined to improve mapping accuracy and precision.

Thank You

R^2 as a goodness of fit statistic for mixed models

Matt Kramer

kramerm@ba.ars.usda.gov

Biometrical Consulting Service, ARS/BARC/USDA

Outline

- Introduction
- (Desirable) properties of R^2
- Philosophies for extension into mixed models
- R^2 estimates for examples of mixed models data
- Conclusion

Introduction

- R^2 is often quoted as a measure of goodness of fit, typically as the proportion of variance in the dependent variable that is explained by the model
- It is natural to ask how R^2 changes when adding random effects or spatially correlated residuals
- Current packages don't provide an R^2 statistic for an estimated mixed model

(Desirable) properties of R^2

Kvålseth (1985, Am. Statistician 39, 279–285) proposed the following requirements for R^2

- 1. R^2 must possess utility as a measure of goodness of fit and have an intuitively reasonable interpretation
- 2. R^2 ought to be dimensionless
- 3. $0 \leq R^2 \leq 1$, where $\mathrm{R}^2 = 1$ corresponds to perfect fit, and $R^2 \geq 0$ for any reasonable model specification
- 4. Applicable to (a) any type of model, (b) whether effects are fixed or random, and (c) regardless of the statistical properties of the model variables
- 5. R^2 should not be confined to any specific model-fitting technique

(Desirable) properties of R^2

- 6. Values for different models fit to the same data set are directly comparable
- 7. Generally compatible with other acceptable measures of fit
- 8. Positive and negative residuals weighted equally

(Desirable) properties of R^2

Under the usual regression model, various definitions yield the same numeric result, e.g.,

- $1 - \sum(y - \hat{y})^2 / \sum(y - \bar{y})^2$
- $\sum(\hat{y} - \bar{y})^2 / \sum(y - \bar{y})^2$
- $(\sigma_y^2 - \sigma_{y|x}^2)/\sigma_y^2$
- $1 - \sum(e - \bar{e})^2 / \sum(y - \bar{y})^2$, e is a model residual
- Squared multiple correlation coefficient between the regressand and the regressors
- Squared correlation coefficient between y and $\hat{y}$
- Different definitions of R^2 may yield different quantities when the usual regression model is generalized

(Desirable) properties of R^2

Cameron and Windmeijer (1996, JBES 14, 209–220), to extend the definition to count data, suggest

- 1. $0 \leq R^2 \leq 1$
- 2. R^2 does not decrease as regressors are added
- 3. R^2 based on residual SS coincides with R^2 based on explained SS
- 4. There is a correspondence between R^2 and a significance test on all slope parameters and between changes in R^2 as regressors are added and significance tests
- 5. R^2 has an interpretation in terms of information content of the data

Philosophies for extension into mixed models

Philosophy 1: R^2 is a measure of between variable effects and should be free of contamination of within variable effects (e.g., autocorrelation due to repeated measures or geographic proximity), otherwise part of the variance of y is explainable by its own past or its neighbors.

Pierce (1979, JASA 74: 901-910) suggests the following form: $R_*^2 = (\sigma_{y|y_*}^2 - \sigma_{y|y_*,x}^2)/\sigma_{y|y_*}^2$, where y_* denotes past or neighboring y. This is similar to the expression for R^2, $(\sigma_y^2 - \sigma_{y|x}^2)/\sigma_y^2$, except that we are now also conditioning on y_*.

Philosophies for extension into mixed models

Philosophy 2: How much better than the mean is a model that predicts y when conditioned on the set of x variables and on past and neighboring values of y?

Magee (1990, Am. Statistician 44: 250–253) suggests developing general R^2 measures based on Wald and likelihood ratio test statistics.

Wald R^2

Wald test: Buse (1973, Am. Statistician 27: 106–108) modifies R^2 as $1 - \frac{\hat{\boldsymbol{u}}'\boldsymbol{V}^{-1}\hat{\boldsymbol{u}}}{(\boldsymbol{Y}-\bar{\boldsymbol{Y}})'\boldsymbol{V}^{-1}(\boldsymbol{Y}-\bar{\boldsymbol{Y}})}$, where $\hat{\boldsymbol{u}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ (i.e. the spatially correlated residuals), $\boldsymbol{V}$ is the variance-covariance matrix of the residuals, and $\bar{\boldsymbol{Y}} = \bar{y}\mathbf{1}$.

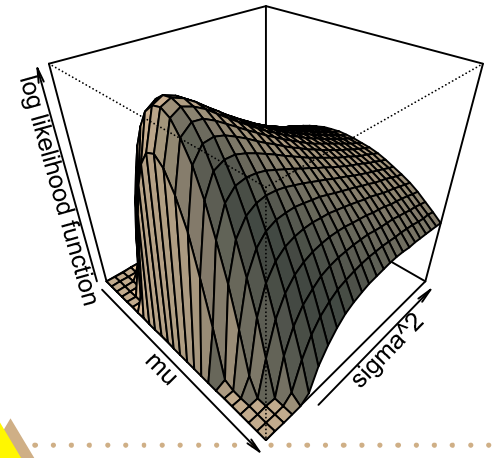
The inverse of $\boldsymbol{V}$ "undoes" the correlation between residuals.

One problem, we don't have $\boldsymbol{V}$, we only have an estimate of it, and it may not be a very good estimate.

A second problem is that software packages don't have this expression pre-programmed, to calculate this R^2 would require some work.

Log-likelihood function for a two parm. model (mean and variance)

100 normally distributed samples were generated ($\mu = 0.5, \sigma^2 = 0.025$) and the log-likelihood function plotted for $\hat{\mu} = [0, 1]$ and $\hat{\sigma}^2 = [0.05, 2]$



Log likelihood R^2

Likelihood ratio: $R^2_{LR} = 1 - \exp(-\frac{2}{n}(\log \mathrm{L}_M - \log \mathrm{L}_0))$, where n is the number of observations, $\log \mathrm{L}_M$ is the log-likelihood of the model of interest, and $\log \mathrm{L}_0$ is the log-likelihood of the intercept-only model.

What is the log-likelihood? The log-likelihood of a statistical model is a function of the data collected and the parameters of the model; the form of this model is assumed known.

It is a special function, the value of the log-likelihood function increases as we reduce the difference between the data and our model for them (we change the value of the log-likelihood function by varying the parameters of the model).

The maximum log-likelihood occurs at those parameter values where this difference is minimized.

Philosophies for extension into mixed models

R^2 based on the likelihood ratio test possesses many desirable properties for a goodness-of-fit statistic

- produces the usual R^2 for ordinary regression (like others)
- since it is based on likelihoods, there is a direct relationship with Kullback-Liebler distance, "information", and information gain $\mathrm{IG} = -\log(1 - R^2_{LR})$ (note that IG is not a linear function of R^2_{LR}) (see Kent (1983), Biometrika 70: 163–174)
- it is easily calculated using output from mixed models software

R^2 examples

Ex. 1: RCBD + covariate (random coefficients, 3 treatments, 4 blocks, 2 obs/block-trt combination, $\sigma^2_{\beta_0} = 4$, $\sigma^2_{\beta_1} = 1$, $\sigma_{\beta_0,\beta_1} = 0$, $\sigma^2 = 1$)

blk 1

A	B
C	A
B	C

blk 2

B	C
B	A
A	C

blk 3

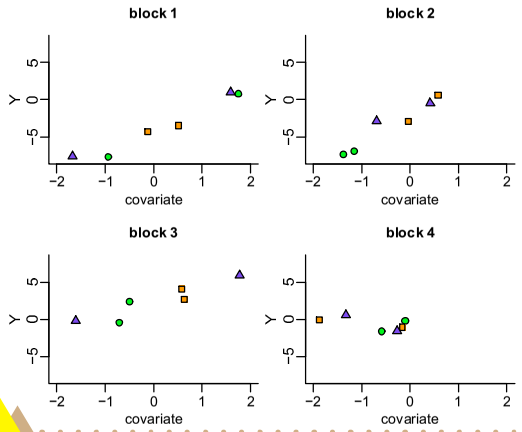
B	C
A	C
A	B

blk 4

B	A
C	C
A	B

R^2 examples

Ex. 1: RCBD + covariate (random coefficients, 3 treatments, 4 blocks, 2 obs/block-trt combination, $\sigma^2_{\beta_0} = 4$, $\sigma^2_{\beta_1} = 1$, $\sigma_{\beta_0,\beta_1} = 0$, $\sigma^2 = 1$)



R^2 examples

R program used for simulating data and estimating the maximum log likelihood (with *nmle* package by Bates and Pinheiro)

model	parms	log likelihood	R^2_{LR}	R^2_W
intercept only	2	-64.45	0	0
trt	4	-63.55	0.07	0.07
trt + cov (f)	5	-59.10	0.36	0.36
trt + blk (r)	5	-60.60	0.27	0.32
trt + blk (r) + cov (r)	7	-42.74	0.84	0.93

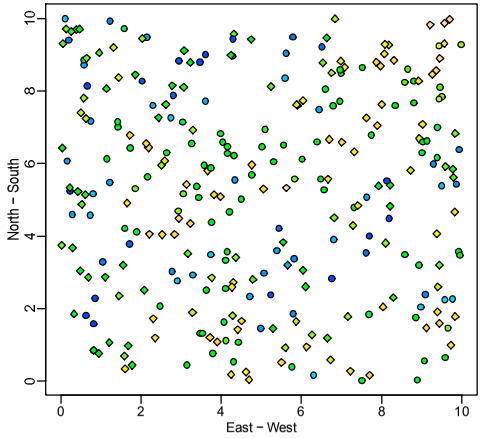
f = fixed
r = random

Spatial exponential correlation

Ex. 2: $\rho = \exp(-d_{i,j}/2)$, $\sigma^2 = 1$, level effect $= 2$, $d =$ distance between i and j

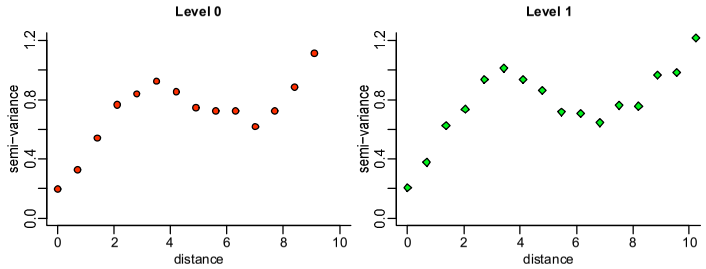
circle = level 0
diamond = level 1

topograhic colors
(blue = lowest values, light brown = highest values)



R^2 examples

Example 2: Semi-variograms



R^2 examples

model	log likelihood	R^2_{LR}
intercept only	-495.94	0
level	-389.68	0.51
level + corr. resid.	-225.27	0.67

Conclusions

- there are various R^2's that can be developed for mixed models, all produce the same value for ordinary regression
- an R^2 based on the likelihood ratio test is easy to calculate from standard mixed models output and has a connection to information theory
- examples were shown demonstrating increases in R^2 when adding random effects or correlated errors to the model