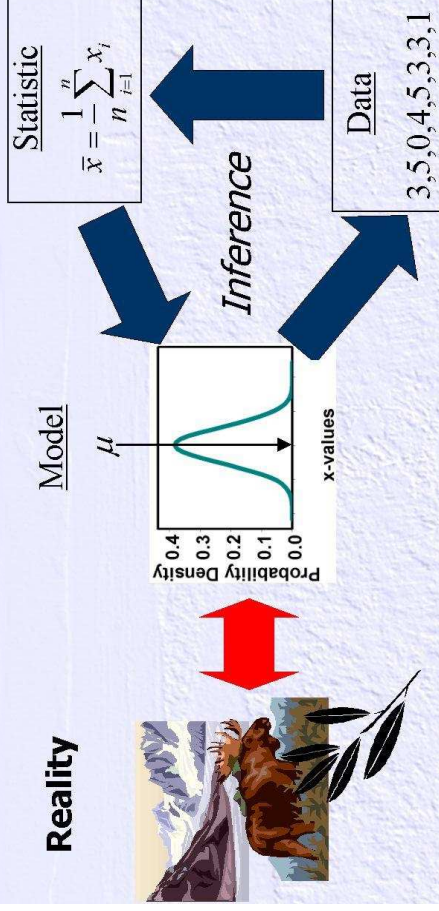


An Introduction to Statistical Models for Spatial Data in Ecology

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

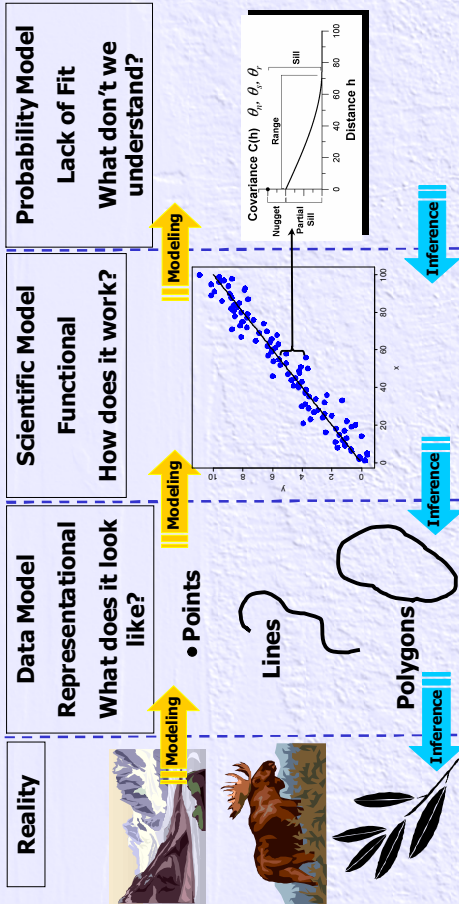
How does it work?



Functional

3

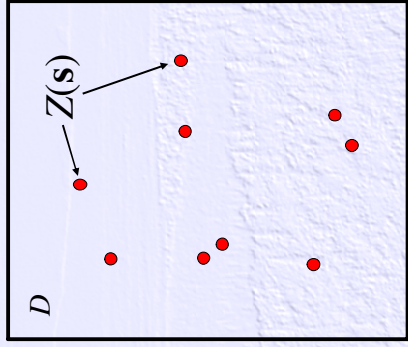
What are Spatial Statistics



"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

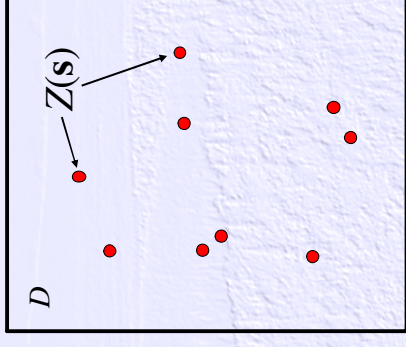
Notation



- 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

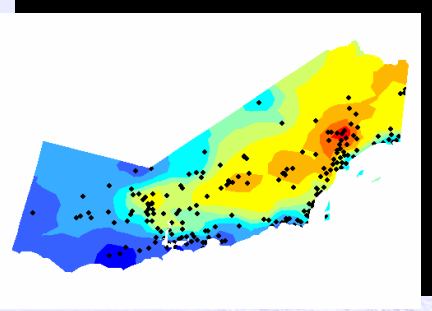


- $\{Z(s) : s \in D\}$
- **Geostatistical Data:** Z random; D fixed, infinite, continuous
- **Lattice Data:** Z random; D fixed, finite, (ir)regular grid
- **Point Pattern Data:** $Z \equiv 1$; D random, finite

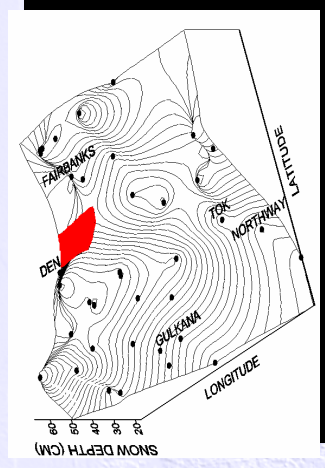
6

Examples of Geostatistical Data

Ozone Predictions



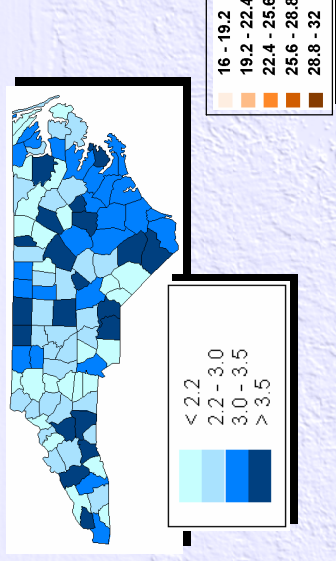
Average Snow Depth



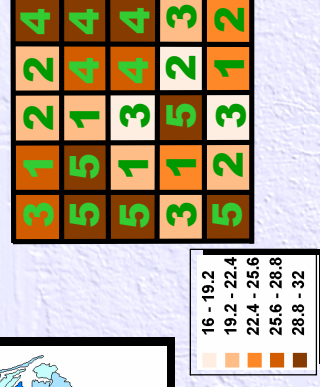
7

Examples of Lattice Data

Transformed SIDS rates



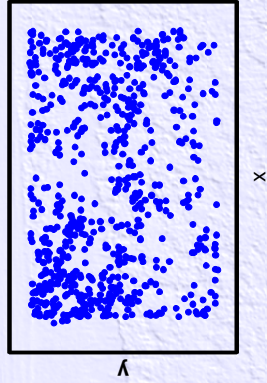
Plots in a Designed Experiment



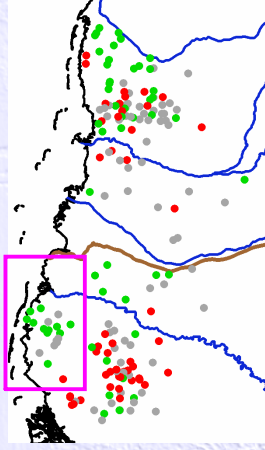
8

Examples of Point Patterns

Lansing Woods Hickory Locations



Arctic Caribou Calving Locations



Statistical Models

Linear Model

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

Nonlinear Model

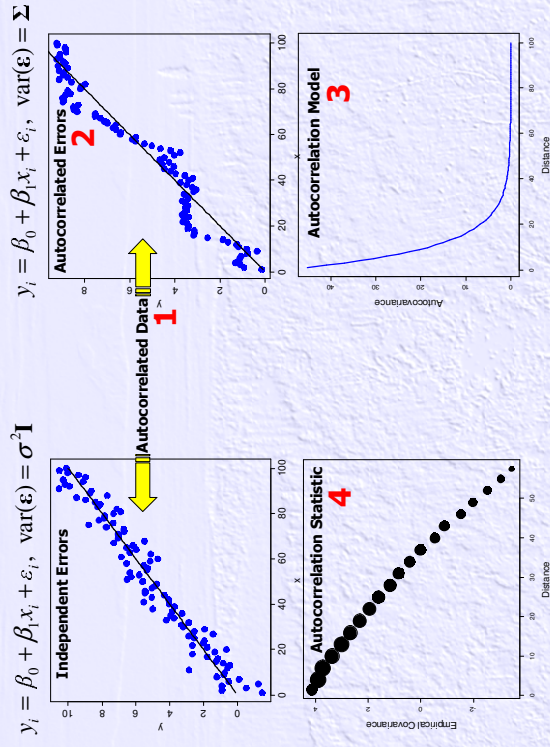
$$\begin{pmatrix} \mathbf{Z}_{observed} \\ \mathbf{Z}_{unobserved} \end{pmatrix} = g(\mathbf{X}, \boldsymbol{\varepsilon}, \boldsymbol{\theta})$$

Prediction Estimation

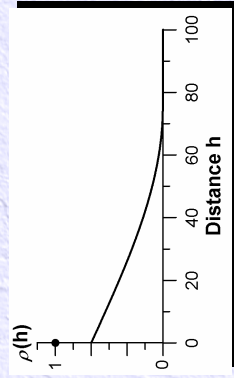
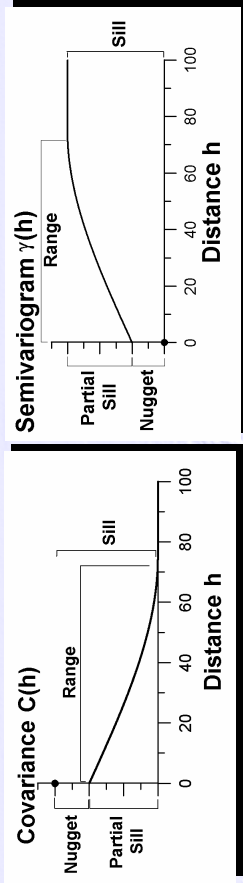
Five Meanings of Autocorrelation

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

Four Meanings of Autocorrelation

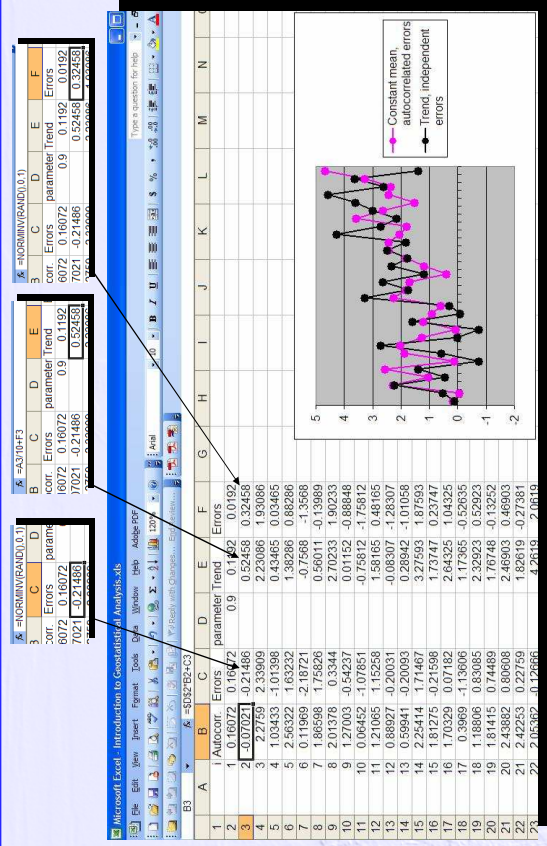


Autocorrelation Models



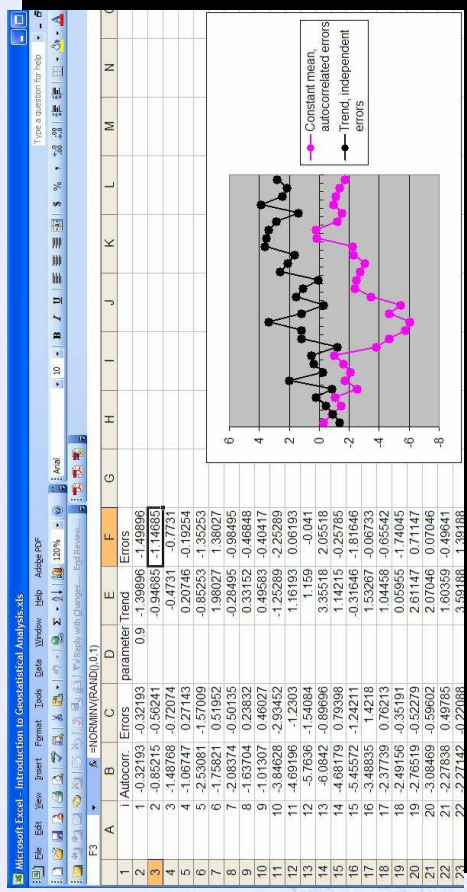
13

Autocorrelation



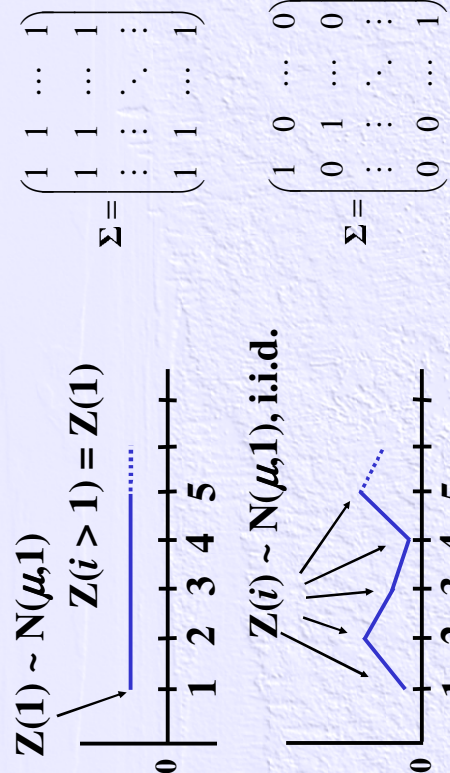
14

Try it! F9



15

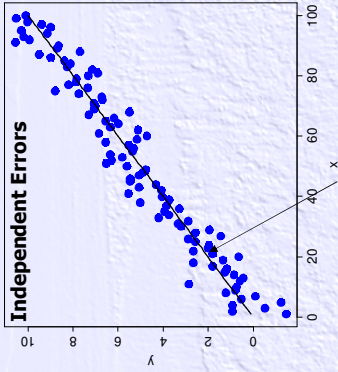
Why Spatial Statistics?



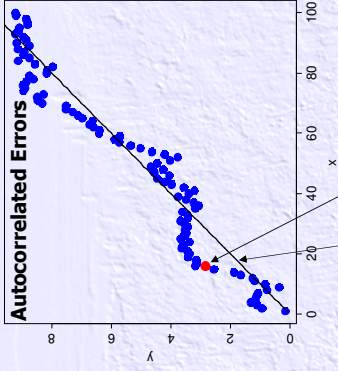
16

Fits vs. Prediction

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



Fit = Prediction



Fit Prediction

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

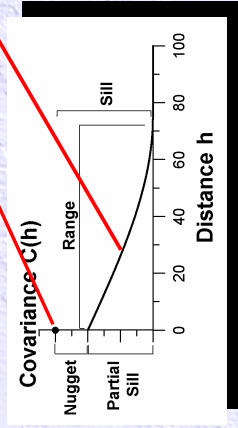
Estimation

- Regression
- Designed Experiments

Why Do We Need Autocorrelation Models?

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

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix}$$



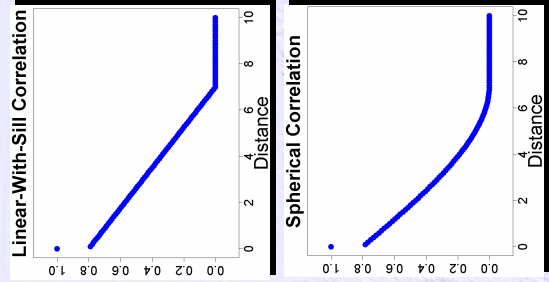
Estimation?!

Leave it to the Statisticians!

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

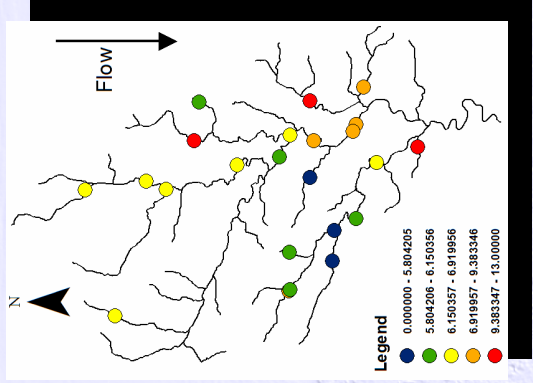
$$\begin{pmatrix} \mathbf{Z}_{observed} \\ \mathbf{Z}_{unobserved} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

Pitfalls: Valid Autocorrelation Models

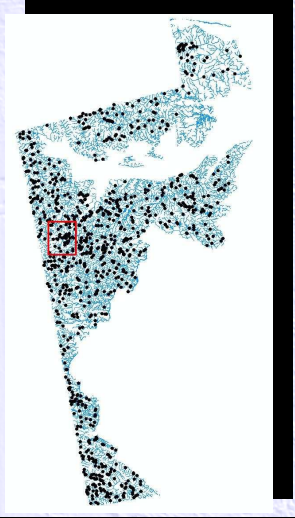


21

Stream Network Models

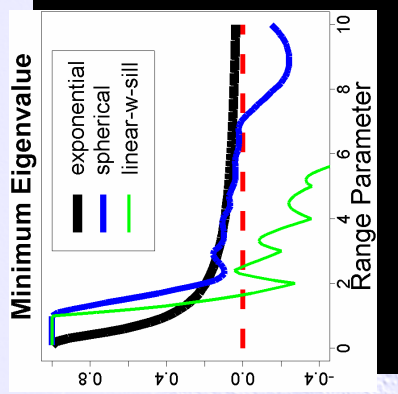
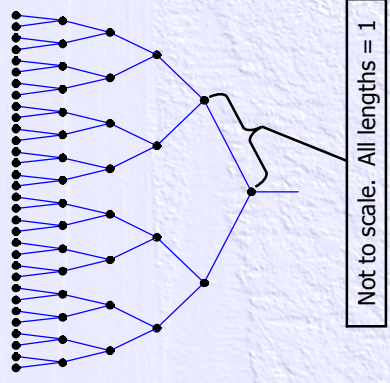


SO₄ Concentration



22

Pitfalls: Valid Models for Stream Networks

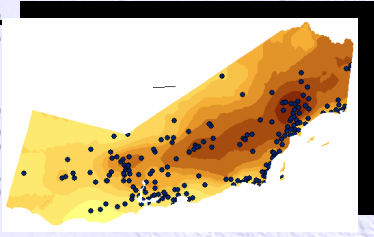


23

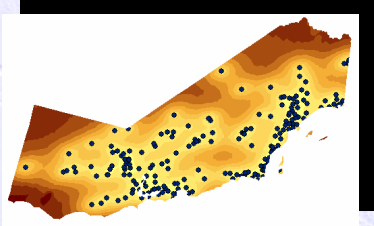
Prediction - Mapping

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

Prediction Map

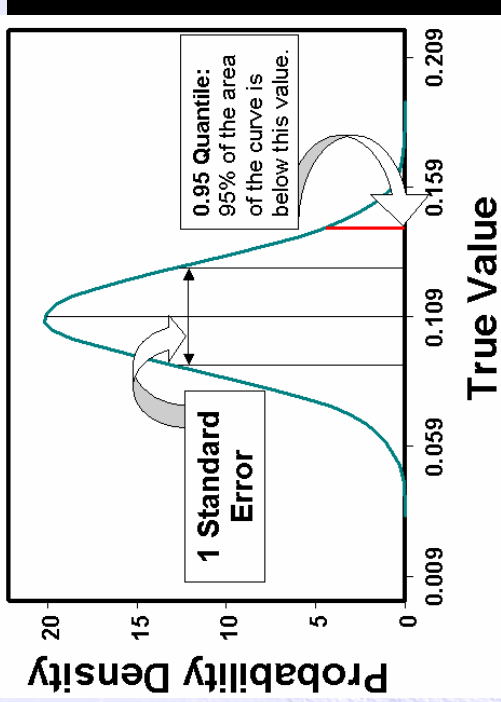


Standard Error Map



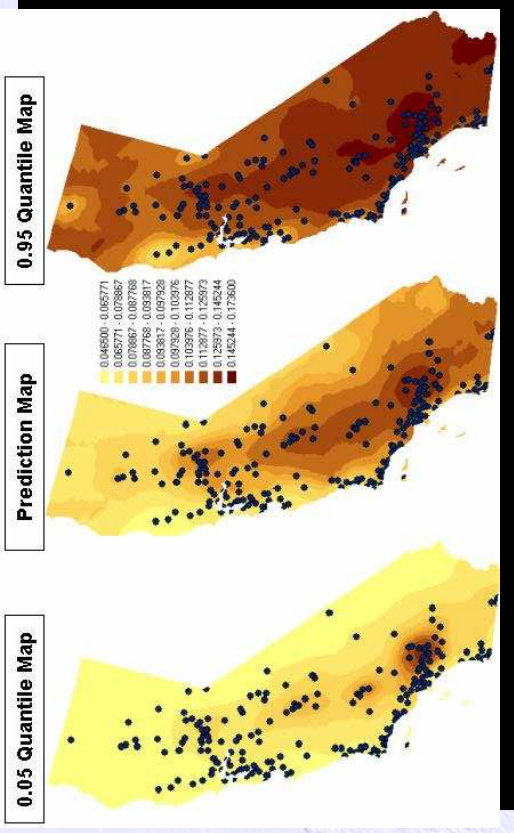
24

Mapping - Quantiles



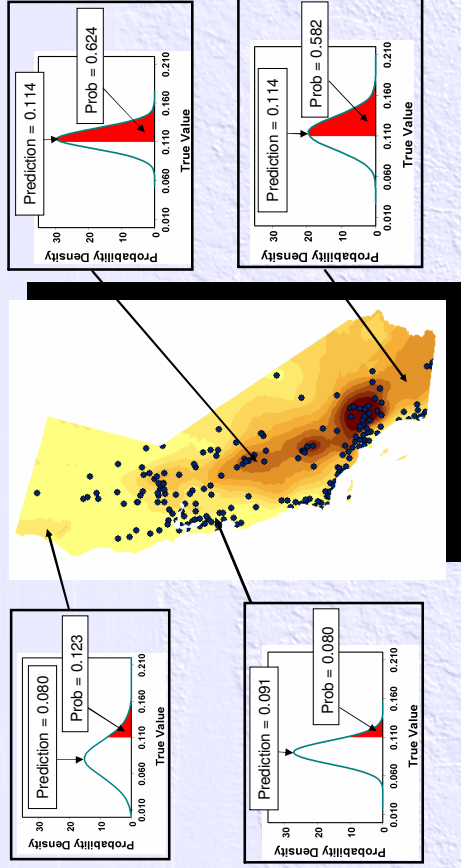
25

Quantile Maps



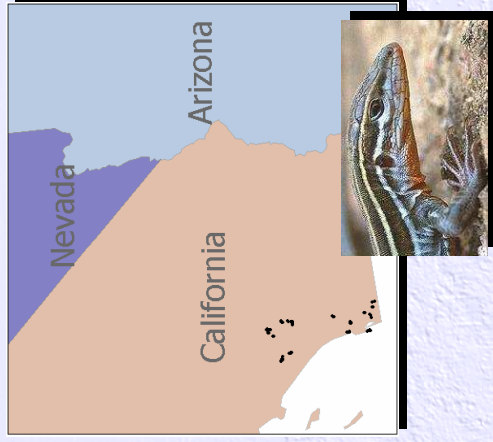
26

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

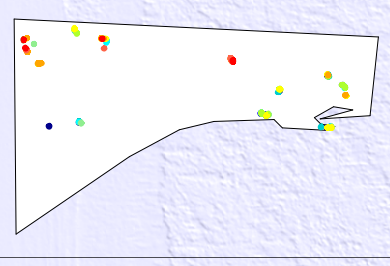


29

California Lizard Data

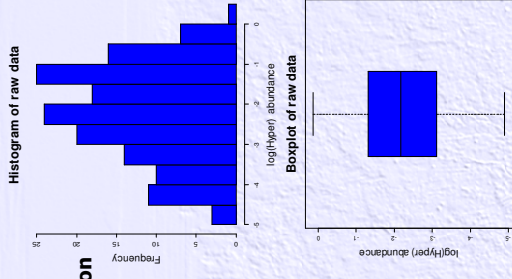
$$\begin{pmatrix} z_{observed} \\ z_{unobserved} \end{pmatrix} = X\beta + \epsilon, \text{ var}(\epsilon) = \Sigma(\theta)$$

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



- -4.9053 to -4.4023
- -4.4023 to -3.9992
- -3.9992 to -2.9951
- -2.9951 to -2.3901
- -2.3901 to -1.9851
- -1.9851 to -1.5801
- -1.5801 to -1.1751
- -1.1751 to -0.7701
- -0.7701 to -0.3651
- -0.3651 to 0.1251

Estimation: REML followed by GLS



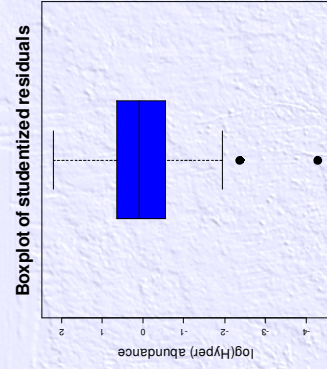
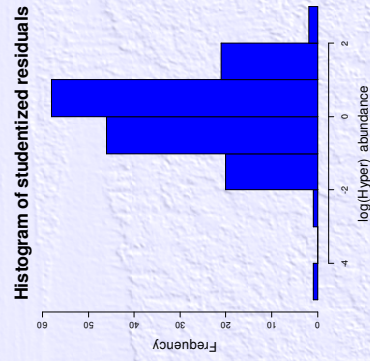
30

Exploratory Data Analysis on Residuals

Studentized Residual:

$$\frac{z_i - \hat{\mu}_i}{\sqrt{MSE(1 - h_{ii})}}$$

$$H = \Sigma^{-1/2} X(X'\Sigma^{-1}X)^{-1} X'\Sigma^{-1/2}$$



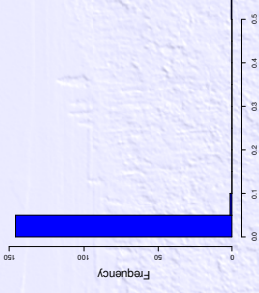
31

Model Diagnostics

Based on deleting observations

Cook's D

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

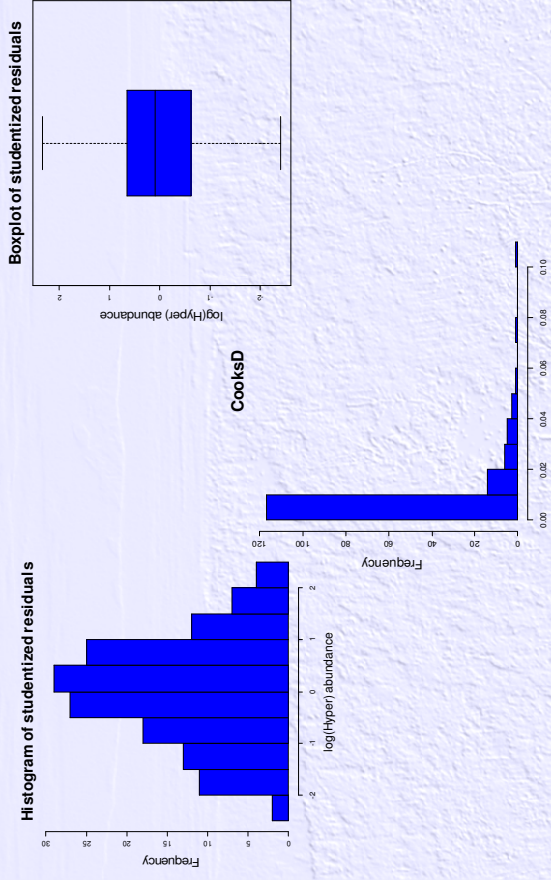


Mostly for outlier detection

$$(z_{observed-1}) = X\beta + \epsilon, \text{ var}(\epsilon) = \Sigma(\theta)$$

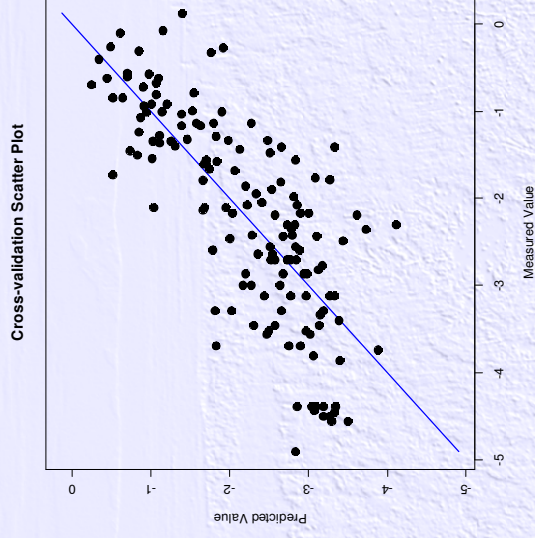
32

Remove Outlier and Re-fit Lizard Data



33

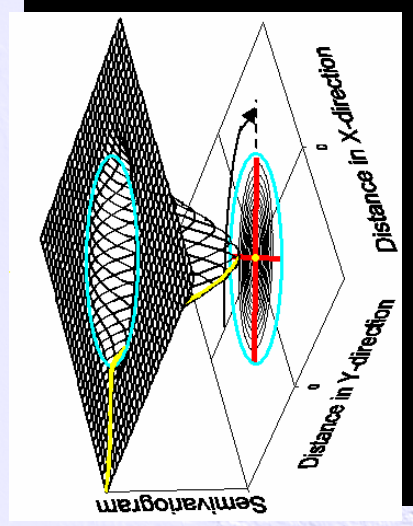
Cross-validation



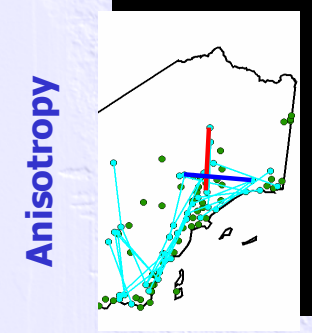
34

Directional Autocorrelation

5 Parameters



Isotropy vs. Anisotropy



35

Cross-validation

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

Standardized Bias :

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

Root Mean Squared Prediction Error :

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

Prediction Interval Coverage :

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

- 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

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

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

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

Estimate

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

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

Prediction

- Mapping
- Sampling

Estimation

- Regression
- Designed Experiments

44

Linear Models

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

Covariance Models

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

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

Exponential, Spherical, etc.

Lattice Models

CAR, SAR, etc.

Estimating Treatment Effects

Contrast	True Value	OLS		Freq Geo		Freq Lat		Bayes Geo		Bayes Lat	
		Est	se	Est	se	Est	se	Est	se	Est	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
		parsill	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

Designed Experiment Experiment

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

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

1600 times Estimate

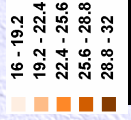
Designed Experiment Results

	ANOVA	GLS-Variogram	Spatial ML Estimation	Spatial REML Estimation
MSE	C1 1.810 C2 1.822 C3 1.139 C4 2.364 C5 2.433	1.166 1.183 0.766 1.546 1.608	1.103 1.105 0.724 1.457 1.551	1.037 1.040 0.687 1.380 1.461
Coverage	C1 0.9450 C2 0.9495 C3 0.9575 C4 0.9500 C5 0.9435	0.9440 0.9415 0.9495 0.9390* 0.9385*	0.9300* 0.9240* 0.9365* 0.9250* 0.9215*	0.9505 0.9500 0.9560 0.9490 0.9465
Power	C1 0.8255 C2 0.9960 C3 1.0000 C4 0.2155 C5 0.0565	0.9750 1.0000 1.0000 0.3370 0.0615	0.9845 1.0000 1.0000 0.4095 0.0785	0.9825 1.0000 1.0000 0.3560 0.0535

Designed Experiments References

Plots in a Designed Experiment

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

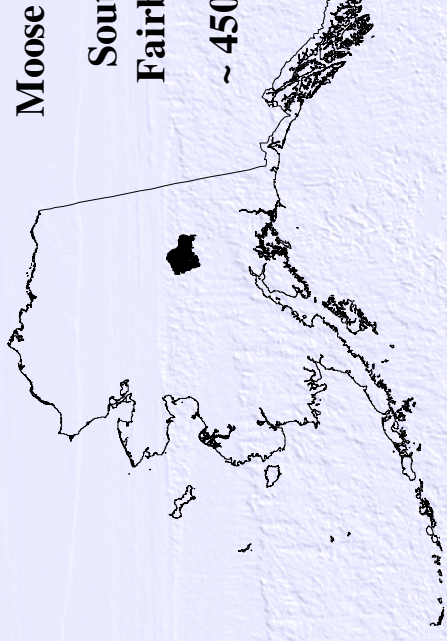


- Ver Hoef, J.M. and Cressie, N. 2001. Spatial statistics: Analysis of field experiments. In Scheiner, S.M. and Gurevitch, J. (eds.), *Design and Analysis of Ecological Experiments, Second Edition*, Oxford University Press, p. 289-307.
- Lenart, E.A., Bowyer, R.T., Ver Hoef, J., and Ruess, R.W. 2002.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

49

Spatial Sampling

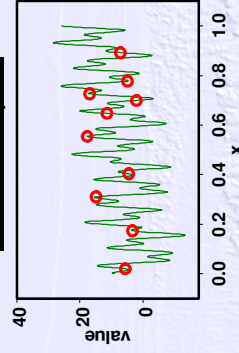
Moose Survey
South of
Fairbanks
~ 4500 mi²



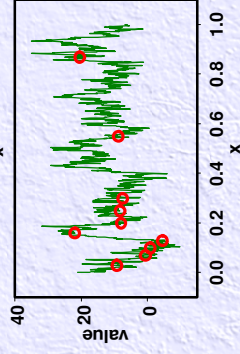
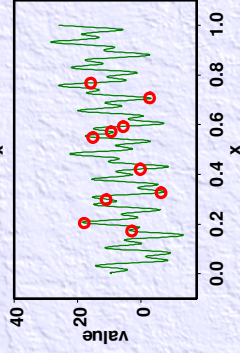
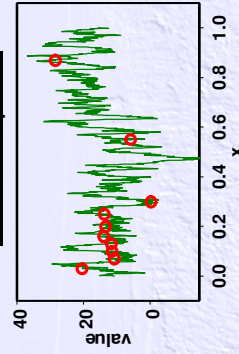
50

Sources of Randomness

Fixed Pattern,
Random Samples



Random Pattern,
Fixed Samples



51

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

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

52

Sampling and Geostatistics

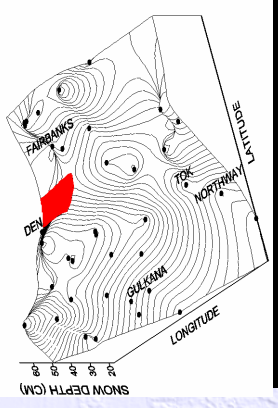
		Method	
Population	Infinite (Spatially Continuous)	Classical Sampling Methods	Geostatistics Block Kriging
		Finite (Spatially Discrete)	Finite Population Block Kriging
	Classical Sampling Methods		Geostatistics Block Kriging
	Classical Sampling Methods		Finite Population Block Kriging

53

Infinite Population Parameters

Total $\tau = \int_A z(s) ds$

Mean $\alpha = \int_A z(s) ds / A$

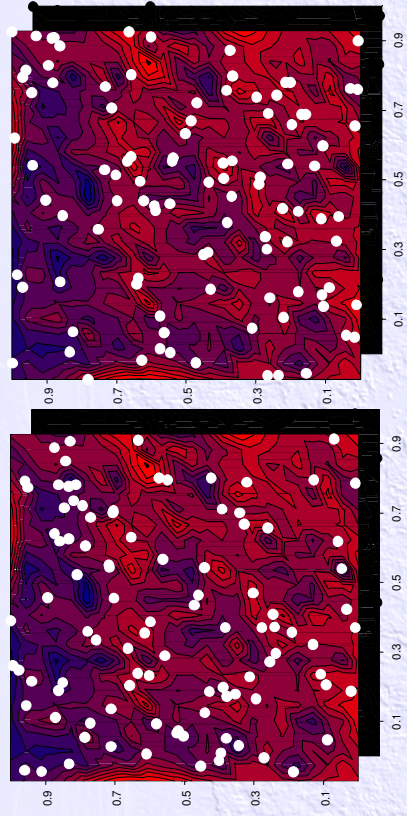


$$\begin{pmatrix} \mathbf{Z}_{observed} \\ \mathbf{Z}_{unobserved} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

54

Simulation Study

Fixed Pattern, Random Samples



55

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
RAE ⁴	1.29	1.00
80%CF 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

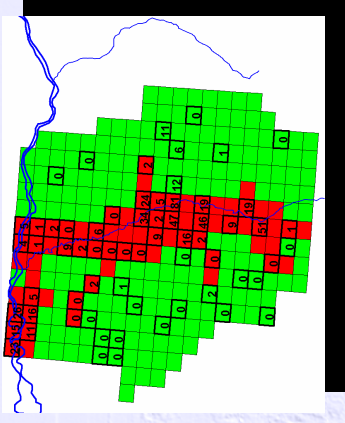
56

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

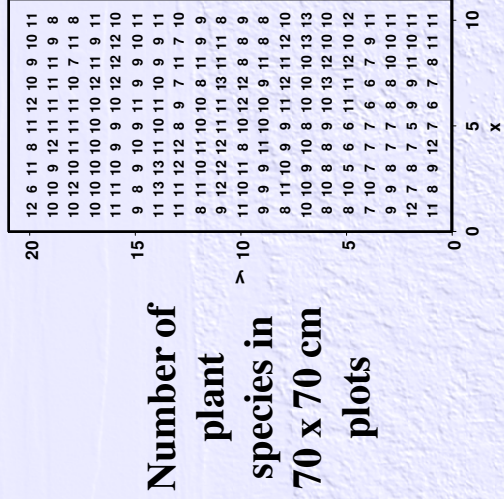
Finite Population Parameters

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



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

Simulation Study



Fixed Population,
 $N = 200$

Random Sample,
 $n = 100$

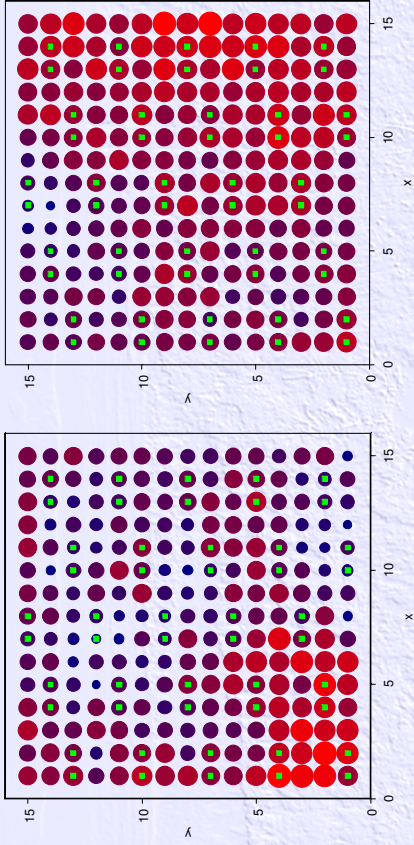
Simulation Results

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
RAE ⁴	0.122	0.105
80%CF 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

Simulation Study



61

Simulation Results

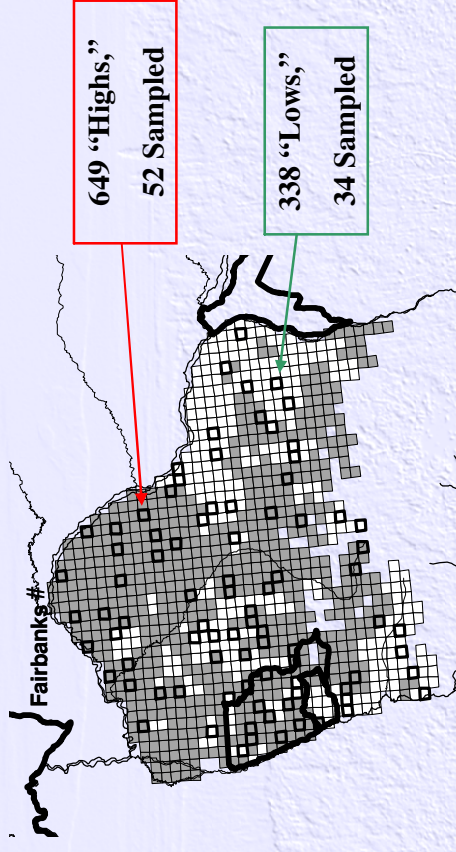
Table 3. Comparison of Random Sampling and Finite Population Block Kriging. 1000 patterns were generated using a spatially autocorrelated stochastic process, and fixed and random samples were taken. Sample sizes were 50. For FPBK, an isotropic exponential covariance model was estimated from the sample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

Validation Statistics	SRS ¹	FPBK ²	FPBK ³
Bias	0.522	-0.181	0.127
RMSPE ⁴	28.0	20.7	17.3
RAE 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



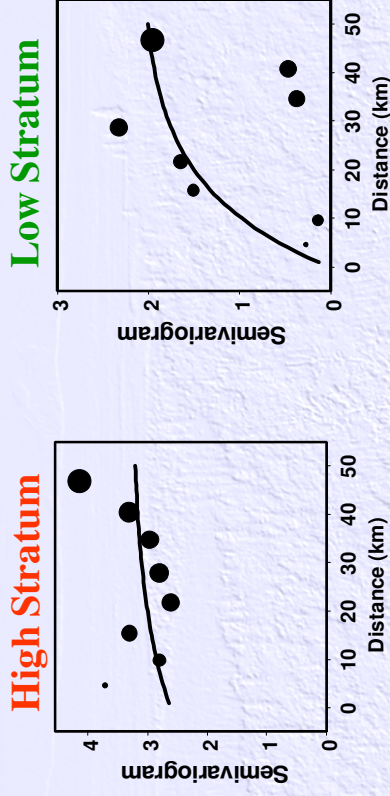
64

Conducting the Survey



65

Modeling Covariance



66

Results

Small Area

FPBK

$$\hat{\tau} = 1437$$

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

SRS (13H, 4L)

$$\hat{\tau} = 1535$$

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

Total Area

FPBK

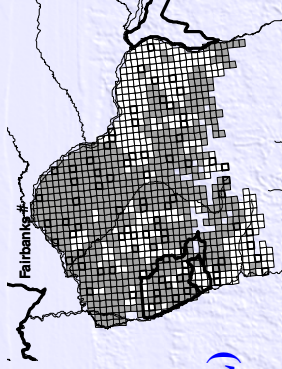
$$\hat{\tau} = 11327$$

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

SRS

$$\hat{\tau} = 11535$$

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



67

Summary

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

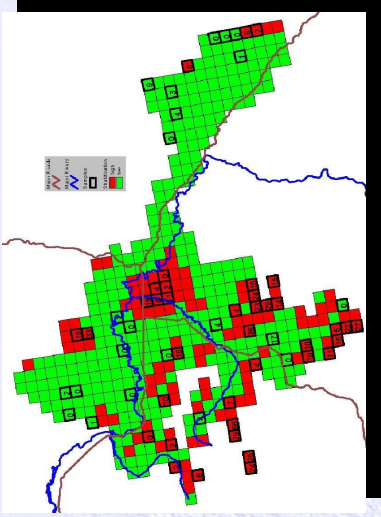
68

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.

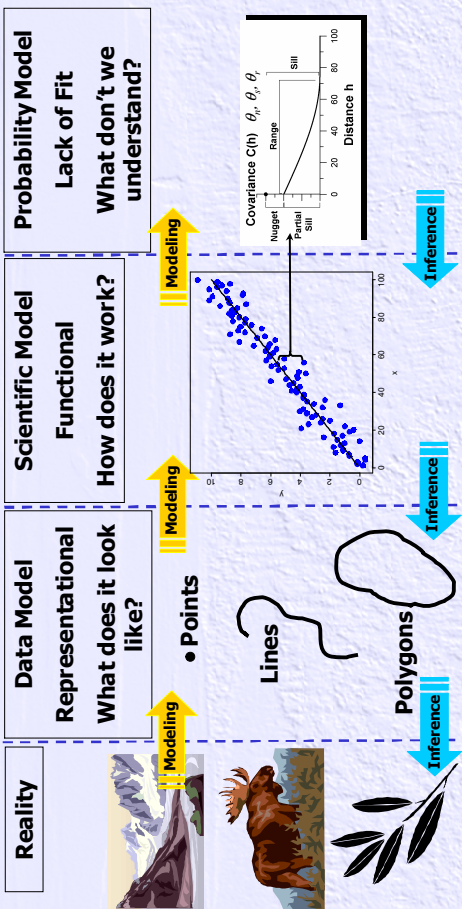
• Ver Hoef, J.M. 2002. Sampling and geostatistics for spatial data. *Ecoscience* 9: 152 – 161.

• 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