

Connecting Correlated GIS Predictors in a Bayes Network Model

Alix I Gitelman with Kathryn Georgitis
Statistics Department
Oregon State University
STARMAP, Colorado State University

June, 2005 WNAR/IMS Meeting, Fairbanks AK

Acknowledgement

This presentation was developed under STAR (Science to Achieve Results) Research Assistance Agreements CR-829095 and CR-829095 awarded by the US Environmental Protection Agency (EPA) to Colorado State University and Oregon State University, respectively. The presentation has not been formally reviewed by the EPA. The views expressed here are solely those the authors and respective programs under these two agreements. The EPA does not endorse any products or commercial services mentioned in this presentation.

Thanks to Nick Danz for providing the data.

Talk Outline

- Concentric Circles Design
- GIS Predictors *ad libitum*
- An Example
- Principal Components and Partial Least Squares
- Bayes Networks/Graphical Models
- Example, Revisited

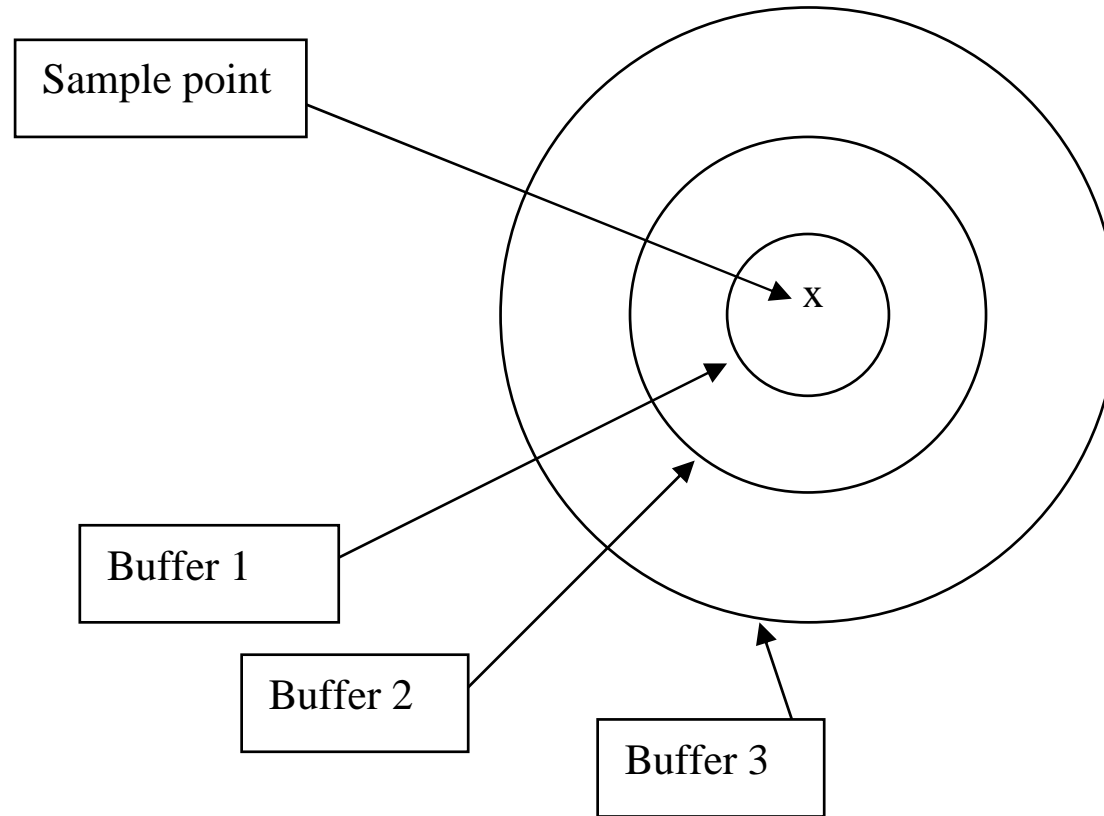
Habitat Association

Questions:

1. What landscape characteristics are associated with habitat selection?
2. Are there different landscape scales associated with habitat selection?

For question 2 Dugan et al. (2002) give a nice discussion of “phenomenon,” “sampling” and “analysis” scales.

Concentric Circle Design



(Bergin et al. 2000; Pearson and Niemi 2000; Hatten and Paradzick 2003; Hostetler and Knowles-Yanez 2003; Martinez et al. 2003; Mayer and Cameron 2003; Holland 2004).

Predictors *ad libitum*

Landscape variables collected using GIS:

- Deciding on classifications
- Deciding on spatial scales (extents)
- Deciding on pixel size
- Deciding on aggregations

An Example

A breeding songbird survey conducted in the Western Great Lakes region of Minnesota and Wisconsin.

- Sampling units were forest stands larger than 40 acres (16 ha).
- 10-minute unlimited radius point count at three subsamples per stand.
- 4 circular buffers of different radii (i.e. different spatial extents): 100m, 500m, 1000m, and 5000m.
- Explanatory variables were derived from a land cover map: aspen-birch; conifer regeneration; hardwood regeneration; lowland conifer; lowland hardwoods; lowland non-forested; northern hardwoods; pine and oak-pine; spruce-fir; upland non-forested.

Predictors *ad libitum*

These land cover predictors are...

...correlated within buffers:

$$\sum_{i=1}^p x_{ij} \leq 100\%$$

where x_{ij} is the percent of the i th land cover type within the j th buffer

...correlated across buffers:

$$\text{corr}(x_{ij}, x_{ik}) \neq 0$$

for some land cover types, i and some buffers, $j \neq k$.

...numerous (e.g., 10 per buffer).

Principal Components

For continuous (Normal) responses, Hwang and Nettleton (2003) give data-driven (i.e., using the responses) methods for principal components regression (PCR).

Schaefer (1985) and others give biased PCR-based estimators in the logistic regression setting.

Is this the best way to address the questions of interest?

Principal Components

Some drawbacks:

1. Interpretability
2. Model selection issues
3. Unless you're lucky, these won't address questions about scale.

Partial Least Squares

Due to Wold (1975), these models seek to combine manifest and latent variables, where the latent variables are intermediate between manifest explanatory variables and manifest responses.

Assumptions:

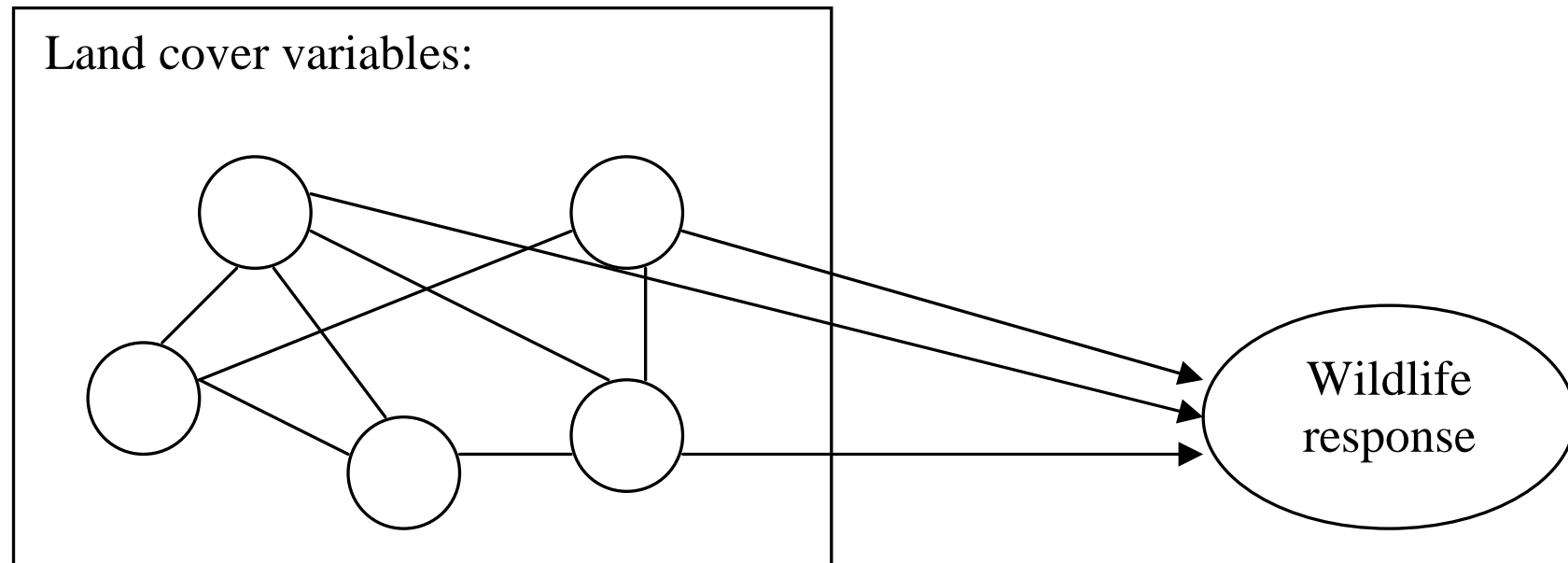
- linear relationships between latent variables
- linear relationships between manifest and latent variables
- direction of correlations between manifest and latent variables

Partial Least Squares

Some drawbacks:

1. Better as a predictive model
2. Understanding the latent mechanisms
3. Fairly similar to PCR

Bayes Networks/Graphical Models



Pearl 2000; Shipley 2000; Gitelman and Herlihy (in press)

Graphical Models

Some features:

1. Take a holistic approach to modeling the ecological system.
2. Reduce or eliminate multicollinearity problems by accounting for dependence among the “explanatory” variables.

Some issues:

1. Model selection: RJMCMC is a good but computationally intensive method.
2. Model evaluation: how to compare graphical models with more traditional approaches?

Specifying a Graphical Model

Let X_{sij} denote the proportion of area in the j th buffer ($j = 1, \dots, k$) covered by the i th land cover type ($i = 1, \dots, p$) at sample point s , $s = 1, \dots, n$.

For the first (innermost) buffer, let $Z_{sij} = X_{sij}$

For each successive buffer, $j = 2, \dots, k$, take

$$Z_{sij} = X_{sij} - \frac{r_{j-1}^2}{r_j^2} X_{si,j-1}$$

where r_j is the radius of the j th buffer.

Model (continued)

So the Z_{sij} 's are the proportions of the i th land cover types in the j th *donut* around the sample point, s .

$\mathbf{Z}'_{sj} = (Z_{sij}, \dots, Z_{spj})$ is a multivariate observation with the constraint that

$$\sum_{i=1}^p Z_{sij} \leq 1$$

for all j and all s .

Furthermore, as the buffer (donut) sizes increase, we ought not to expect these proportions to remain constant.

Indeed, it might be that the “patchiness” is an important habitat association consideration.

Model (continued)

Let Y_s denote the wildlife response at sample point s .

The joint probability distribution of Y_s and $\mathbf{Z}_{s1}, \dots, \mathbf{Z}_{sk}$ can be written:

$$f(Y_s, \mathbf{Z}_{s1}, \dots, \mathbf{Z}_{sk} | \phi) = f(Y_s | \mathbf{Z}_{s1}, \dots, \mathbf{Z}_{sk}, \phi_Y) f(\mathbf{Z}_{s1}, \dots, \mathbf{Z}_{sk} | \phi_Z),$$

Where ϕ_Y and ϕ_Z denote parameters corresponding to the conditional distributions of $Y_s | \mathbf{Z}_{s1}, \dots, \mathbf{Z}_{sk}$ and $\mathbf{Z}_{s1}, \dots, \mathbf{Z}_{sk}$, respectively, where $\phi = (\phi_Y, \phi_Z)$.

Using a graphical model approach, we can factor the joint distribution of the \mathbf{Z}_{sj} 's and eliminate some of them from the conditional distribution of Y_s .

Example Revisited

Between buffer correlations (for some land cover types):

Land Cover Type	$r(B_1, B_2)$	$r(B_1, B_3)$	$r(B_2, B_3)$
Aspen-Birch	0.73	0.62	0.93
Conifer (Reg)	0.96	0.79	0.87
Lowland Conifer	0.73	0.56	0.87
Lowland Non-forest	0.51	0.27	0.33
Pine/Oak-Pine	0.81	0.70	0.95
Spruce-Fir	0.60	0.50	0.88

These estimates are all based on $n = 156$.

Example Revisited

Within buffer 1 correlations:

	A-B	C (R)	LC	LNf	POP	S-F
A-B	1.00	0.02	-0.28	0.02	-0.46	-0.05
C (R)		1.00	-0.05	0.08	-0.12	0.00
LC			1.00	0.27	-0.23	-0.18
LNf				1.00	-0.24	-0.03
POP					1.00	-0.31
S-F						1.00

Similar results for buffers 2 and 3.

Graphical Model: Part 1

