

# Connecting Correlated Predictors Using Graphical Models

Alix I Gitelman with K. Georgitis  
Statistics Department  
Oregon State University  
DAMARS, STARMAP

September 2005; SARMMM, Corvallis

## 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 songbird survey data.

## Talk Outline

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

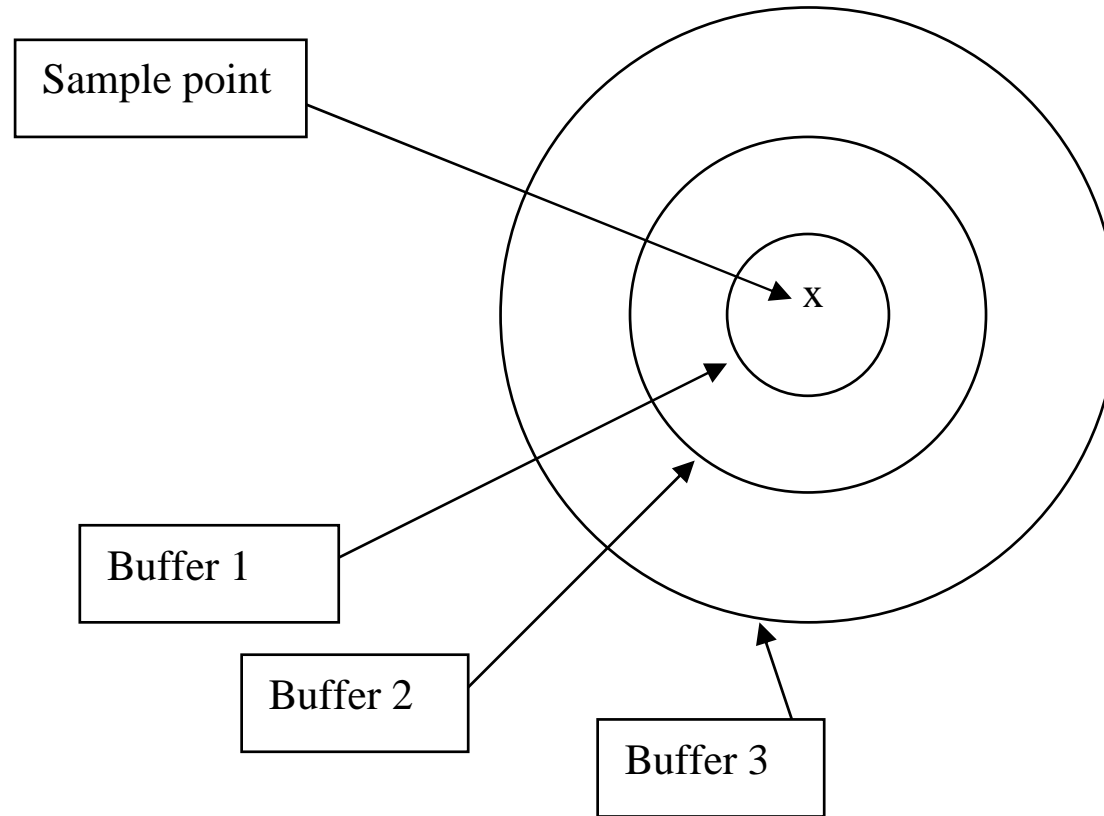
## Habitat Association

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/recorded using GIS:

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

These decisions are part of “the answer.”

## An Example

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

- ten-minute unlimited radius point count at three subsamples per stand (in the end—a presence/absence response)
- four circular buffers of different radii (i.e. different spatial extents): 100m, 500m, 1000m, and 5000m (we drop the 5000m buffer)
- 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 (we drop four of these)

## Predictors *ad libitum*

These land cover predictors are...

...correlated within buffers:

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

where  $x_{ij}$  is the proportion of the  $j$ th buffer covered by the  $i$ th land cover type

...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?

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

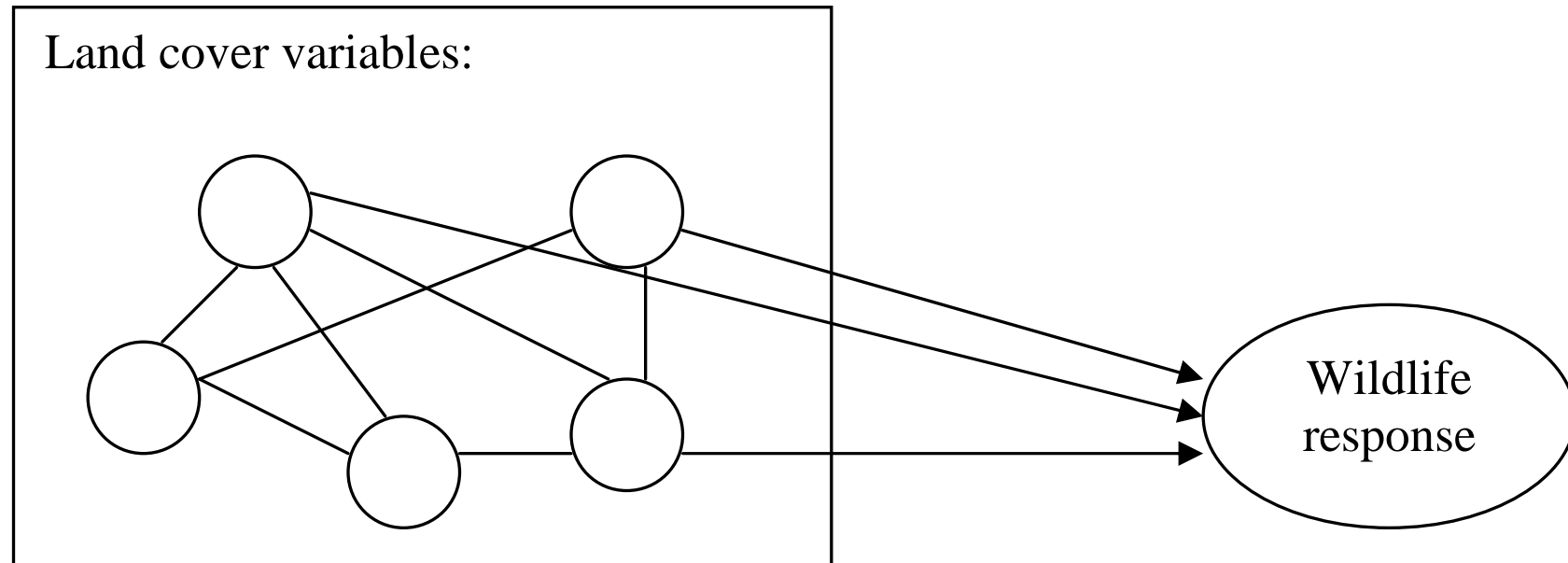
# Partial Least Squares

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: linearity, direction of correlations.

1. better as a predictive model
2. understanding the latent mechanisms
3. can be fairly similar to PCR

# Graphical Models



Pearl 2000; Shipley 2000

# Graphical Models

Some features:

1. Take a holistic approach to modeling the ecological system.
2. Reduce or eliminate multicollinearity problems by explicitly modeling dependence between predictor 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  $\phi_Y$  and  $\phi_Z$  denote parameters corresponding to the 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  $Y_s$  given the  $\mathbf{Z}_{sj}$ 's.

## Example Revisited

Between buffer correlations (for some land cover types):

<b>Land Cover Type</b>	$r(B_1, B_2)$	$r(B_1, B_3)$	$r(B_2, B_3)$
Aspen-Birch (AB)	0.73	0.62	0.93
Conifer Regen (CR)	0.96	0.79	0.87
Lowland Conifer (LC)	0.73	0.56	0.87
Lowland Non-forest (LN)	0.51	0.27	0.33
Pine/Oak-Pine (PO)	0.81	0.70	0.95
Spruce-Fir (SF)	0.60	0.50	0.88

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

## Example Revisited

Within buffer 1 correlations:

	AB	CR	LC	LN	PO	SF
AB	1.00	0.02	-0.28	0.02	-0.46	-0.05
CR		1.00	-0.05	0.08	-0.12	0.00
LC			1.00	0.27	-0.23	-0.18
LN				1.00	-0.24	-0.03
PO					1.00	-0.31
SF						1.00

Similar results for buffers 2 and 3.

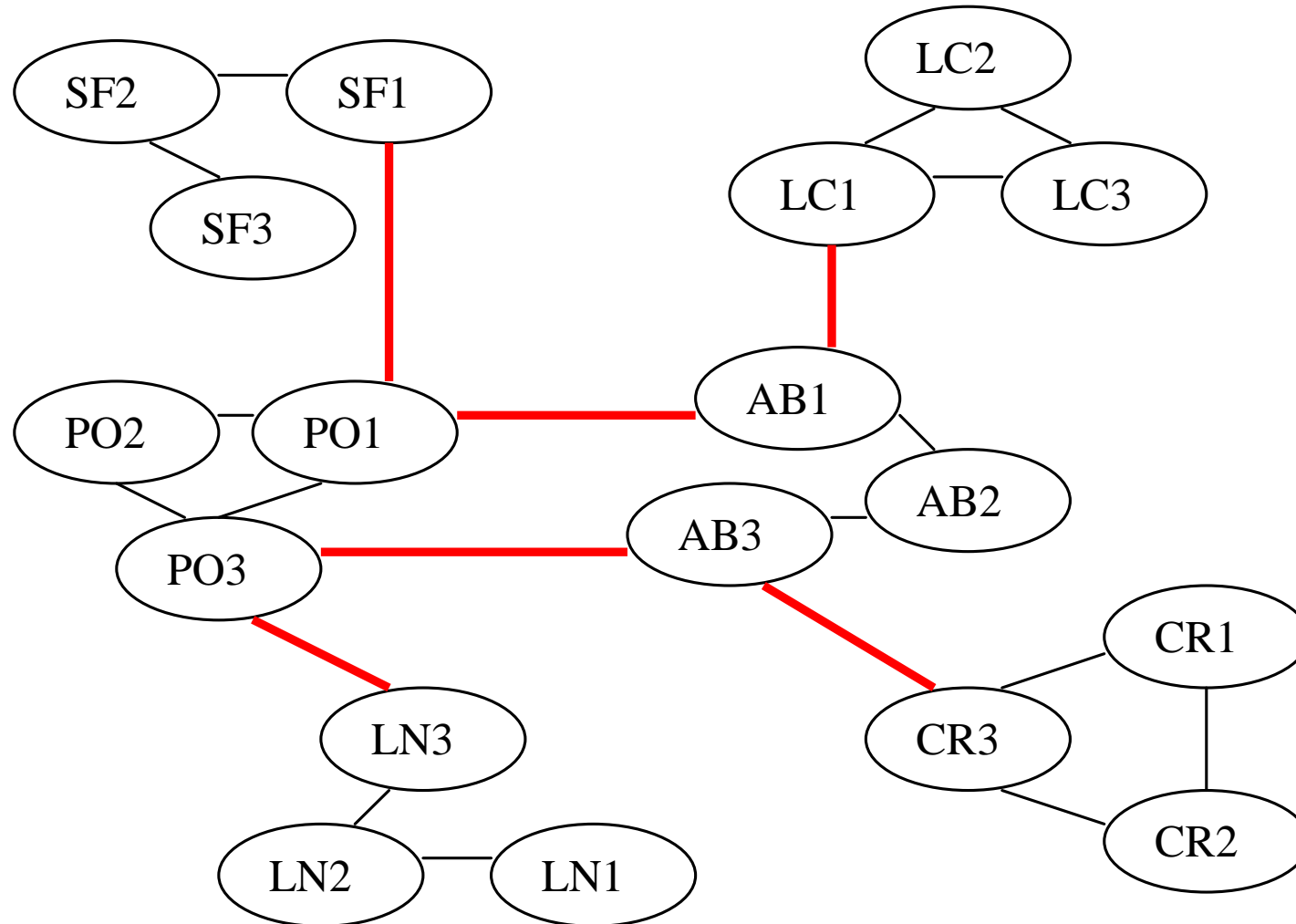
## PCA and PLS Results

A principal component analysis of these same variables gives the following:

- PC1 is a linear combination of the aspen-birch and pine and oak pine cover types
- PC2 is a linear combination of aspen-birch, pine and oak pine and lowland conifer
- Not until PC6 is more than 90% of the overall variation explained.

A PLS analysis finds one hidden factor, and only explains  $\sim 20\%$  of the overall variation.

# Graphical Model Results



## Graphical Model Results

AMP-Separation (Anderssen, Madigan & Perlman 2001) allows factorization of the distribution of the land cover variables as follows:

$$\begin{aligned} f(\mathbf{SF}, \mathbf{LC}, \mathbf{PO}, \mathbf{AB}, \mathbf{LN}, \mathbf{CR}) &= f(\mathbf{SF}, \mathbf{LC}, \mathbf{LN}, \mathbf{CR} | \mathbf{PO}, \mathbf{AB}) f(\mathbf{PO}, \mathbf{AB}) \\ &= f(\mathbf{SF} | \mathbf{PO}, \mathbf{AB}) f(\mathbf{LC} | \mathbf{PO}, \mathbf{AB}) f(\mathbf{LN} | \mathbf{PO}, \mathbf{AB}) \\ &\quad \times f(\mathbf{CR} | \mathbf{PO}, \mathbf{AB}) f(\mathbf{PO}, \mathbf{AB}) \\ &= f(\mathbf{SF} | \mathbf{PO}) f(\mathbf{LC} | \mathbf{AB}) f(\mathbf{LN} | \mathbf{PO}) f(\mathbf{CR} | \mathbf{AB}) \\ &\quad \times f(\mathbf{PO}, \mathbf{AB}) \end{aligned}$$

For example, spruce-fir distributions are independent of aspen-birch distributions once we condition on pine and oak pine distributions; etc.

## Some Comments

- distributional assumptions (proportions aren't Normally distributed; perhaps Beta is better?)
- connections to the wildlife response
- quantitative goodness-of-fit assessments (compare a “graphical regression model” to a graphical model)
- a more realistic interpretation of the land cover variation—can we start to answer our questions about land cover types and scale?