Statistical Tools for Systems Research

NCSU | NCA&TSU | NCDA&CS

www.cefs.ncsu.edu

Chris Reberg-Horton
Organic Cropping Systems Specialist

NC STATE UNIVERSITY

a priori vs. *a posteriori* use of data

- *a priori* - agricultural disciplines
  - most of ANOVA and regression use
  - pairwise comparisons and exploratory regression analyses are notable exceptions
- *a posteriori* – many fields of ecology
  - Ordination, multiple regression – particularly stepwise type procedures, correlation
- Over the last 20 year, agricultural disciplines have improved at *a posteriori* and ecology has improved at *a priori*

Multiple Regression Example

Schomberg, H.H. et al. 2009. Assessing Indices for Predicting Potential Nitrogen Mineralization in Soils under Different Management Systems. Soil Sci. Soc. Am. J. Volume 73: 1575-1586

- Study of 9 sites with various tillage practices. These sites are part of systems trials.
- Measured many different soil N pools in many different ways.
- $$N_t = N_0(1 - e^{-kt})$$
- What is the best combination of tools for predicting mineralizable N?

Multiple Regression Example

Table 6. Equations† for predicting N_0, N_0*, and k from N indices.

	N index	Intercept	Standard error‡	Slope	Standard error	WLS r^2§
N_0¶	TC	2.36	0.23	1.12	0.09	0.65
	TN	−2.15	0.63	1.05	0.09	0.64
	POMC	4.33	0.12	0.64	0.08	0.52
	POMN	2.11	0.41	0.57	0.08	0.50
	Cold_N	4.77	0.11	0.34	0.07	0.35
	Hot_N	2.58	0.23	1.10	0.10	0.62
	Hyd_N	3.05	0.21	1.01	0.10	0.60
	NaOH_N	−0.02	0.60	1.03	0.12	0.55
	PB_N	1.60	0.43	1.05	0.13	0.54
	Ana_N	2.54	0.22	0.78	0.07	0.63
	Nmin_24	2.71	0.20	0.80	0.07	0.62
	FI_CO2	0.78	0.43	0.91	0.09	0.58
	Ca_hypcl	3.69	0.16	0.19	0.02	0.57

Multiple Regression Example

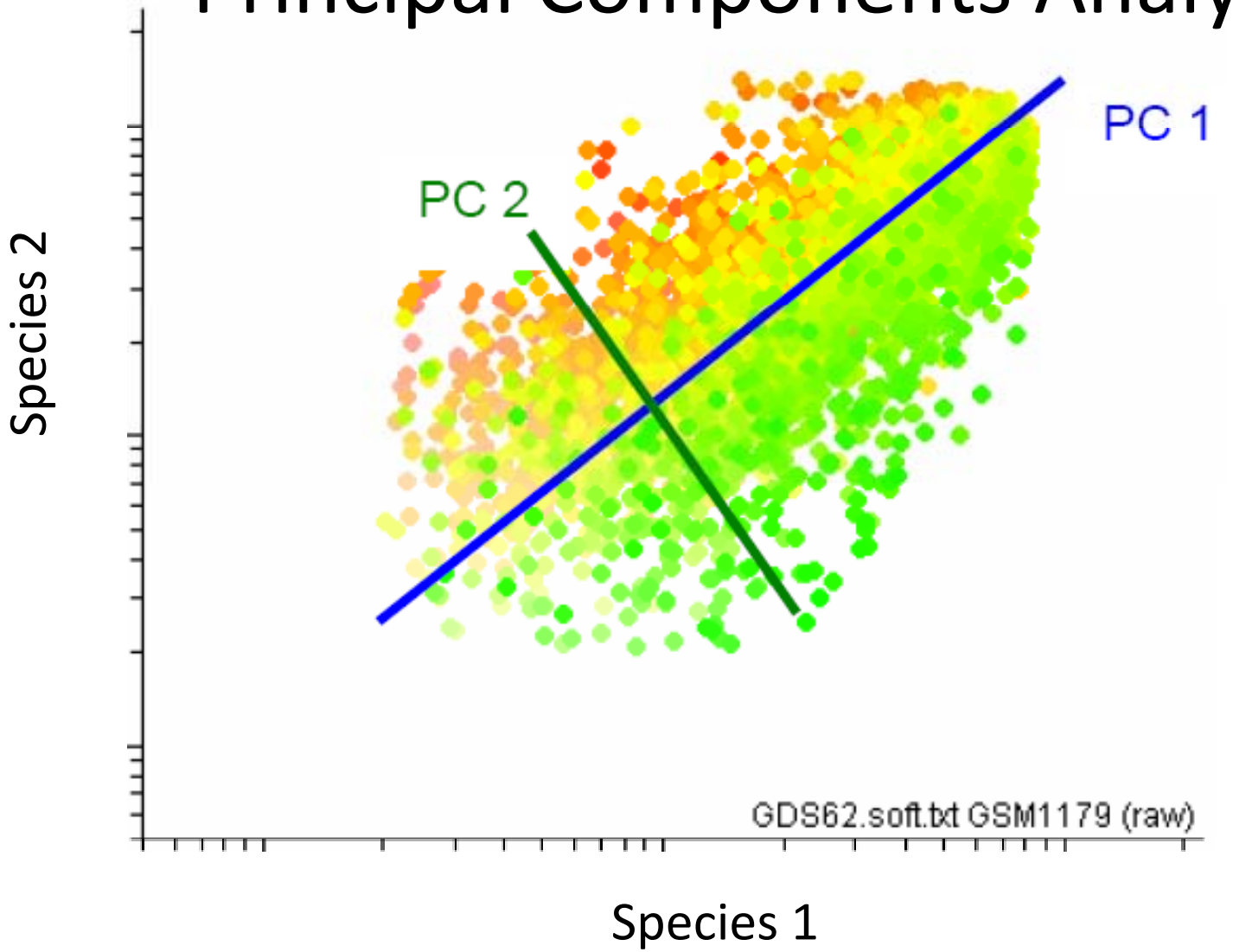
Table 7. Multiple indices equations† for predicting N_0, N_0^*, and k.

Dependent	Variable	Parameter estimate	Standard error	Pr > |t|	95% Confidence limits		RMSE	Model R^2	Model Adj R^2
N_0	Intercept	−1.655	0.509	0.0025	−2.686	−0.623	0.288	0.86	0.85
	TN	0.682	0.114	0.0001	0.452	0.913			
	FI_CO2	0.432	0.104	0.0002	0.221	0.642			
N_0^*	Intercept	−0.930	0.444	0.0436	−1.831	−0.028	0.185	0.94	0.94
	TN	0.820	0.178	0.0001	0.459	1.181			
	Cold_N	0.128	0.028	0.0001	0.072	0.185			
	NaOH_N	−0.336	0.167	0.0524	−0.676	0.004			
	FI_CO2	0.421	0.071	0.0001	0.277	0.564			
k	Intercept	−3.563	0.390	< .0001	−4.356	−2.771	0.408	0.36	0.30
	TC	−1.380	0.373	0.0008	−2.139	−0.622			
	POMN	0.500	0.148	0.0019	0.199	0.802			

Multiple Regression

- Pros
  - Can be used in both *a priori* and *a posteriori* analyses
  - Can be used for empirical and mechanistic analyses
- Cons
  - Automatic procedures can produce models that only fit the dataset in hand
  - Multicollinearity and determining cause and effect are real problems
- New things to know: Full and reduced model F testing has given way to "information criterion". (AIC, AICC, BIC most common)

“Ordination”
Principal Components Analysis



Detrended Correspondence Analysis (DCA): Example

Davis, A.S, K.A. Renner, and K.L. Gross. 2005. Weed seedbank and community shifts in a long-term cropping systems experiment. Weed Science 53:296-306.

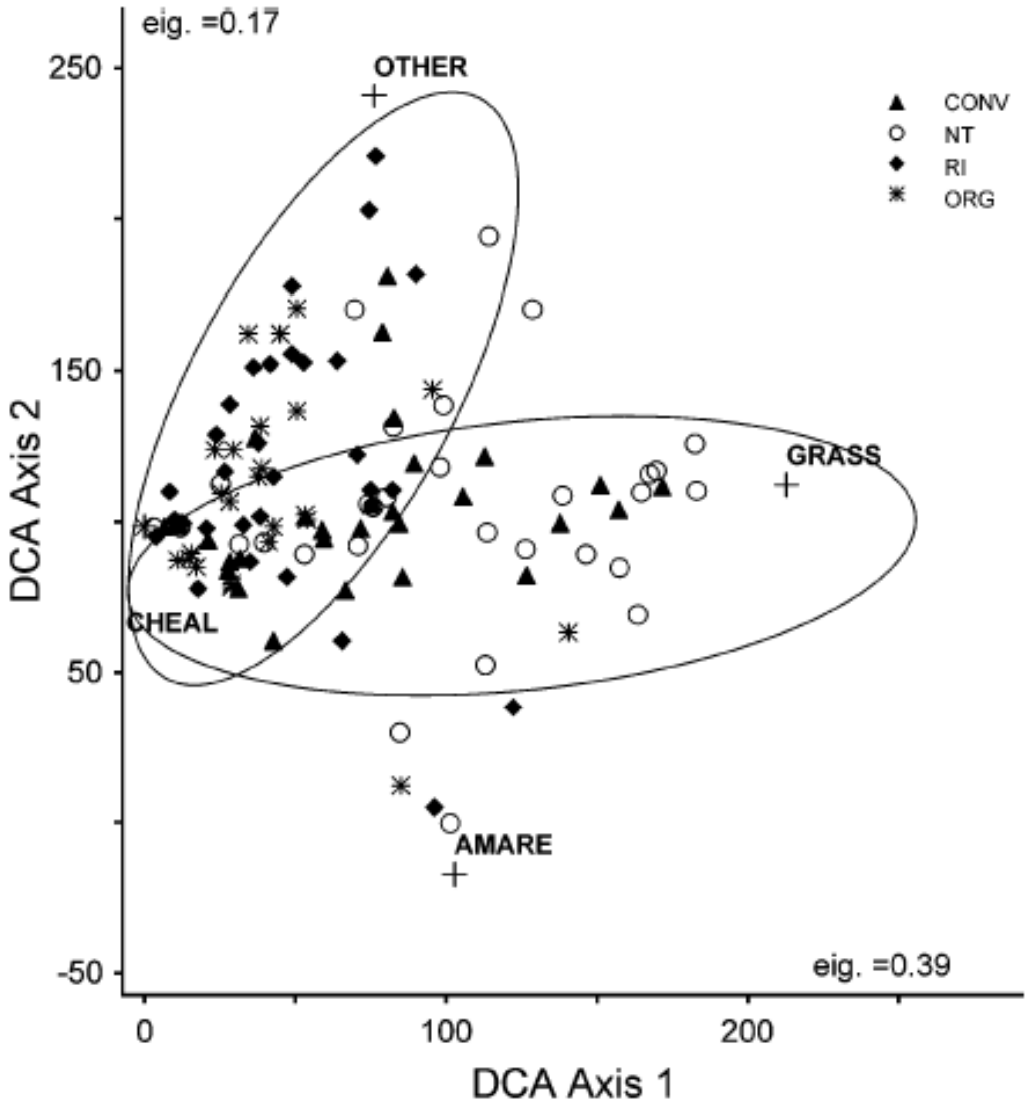
- Long-term cropping systems trial in Michigan (LTER)
- Four systems: conventional, no-till, reduced input, organic
- Collected weed seedbank data in five years out of a 12 year time period
- Collected weed biomass data every year

Detrended Correspondence Analysis (DCA): Example

TABLE 2. Multivariate analysis of variance of the effects of site heterogeneity, management system, and time on weed community composition[a] within the Long Term Ecological Research plots at Kellogg Biological Station, Hickory Corners, MI.

Community type	Source	Pillai trace	*F*	df	P value
Seedbank	Rep	0.55	3.01	20, 380	< 0.001
	Management (M)	0.93	10.49	12, 282	< 0.001
	Year (Y)	1.32	11.77	16, 380	< 0.001
	MY	0.88	2.23	48, 380	< 0.001
Aboveground	Rep	0.27	3.05	20, 84	< 0.001
	Management (M)	0.83	20.10	12, 627	< 0.001
	Year (Y)	1.33	10.49	40, 840	< 0.001
	MY	1.52	4.28	120, 840	< 0.001

Detrended Correspondence Analysis (DCA): Example



Detrended Correspondence Analysis (DCA): Example

TABLE 3. Spearman correlations between plot scores for Detrended Correspondence Analysis (DCA) Axes 1 and 2 of Long Term Ecological Research weed community species composition and selected environmental variables.

Environmental variable	Weed community ordination Axes			
	Seedbank		Aboveground biomass	
	DCA1	DCA2	DCA1	DCA2
%Sand	− 0.17	0.23*[a]	0.11	0.08
%Silt	0.17	− 0.22*	− 0.11	− 0.09
%Clay	0.11	− 0.15	− 0.05	− 0.03
Bulk density	− 0.01	0.16	− 0.20*	0.02
pH	− 0.11	− 0.36*	0.10	0.02
Replication	0.23*	0.25*	0.21*	− 0.13
Management	-0.42	− 0.02	0.28**	0.09
Year	0.22†	0.33**	− 0.23*	0.07

Detrended Correspondence Analysis (DCA)

- Pros
  - Great for situations where species response to "latent" variables is not linear.
  - Distills many variables into few dimension for graphing and finding patterns.
  - Can deal with lots of zeroes, more dependent variables that observations, non-normality.
- Cons
  - "Environmental variables" you measured may not explain much of the variation in your "species" data.
  - Hypothesis testing

Redundancy Analysis: Example

Reberg-Horton, C., E.R. Gallandt, and T. Molloy. 2006. Measuring community shifts in a weed seedbank study with the use of distance-based redundancy analysis. Weed Science 54:861-866.

- Long term systems trial: conventional, reduced input, and biointensive systems
- Subplots: amended vs. unamended with composts and manures
- Two entry points

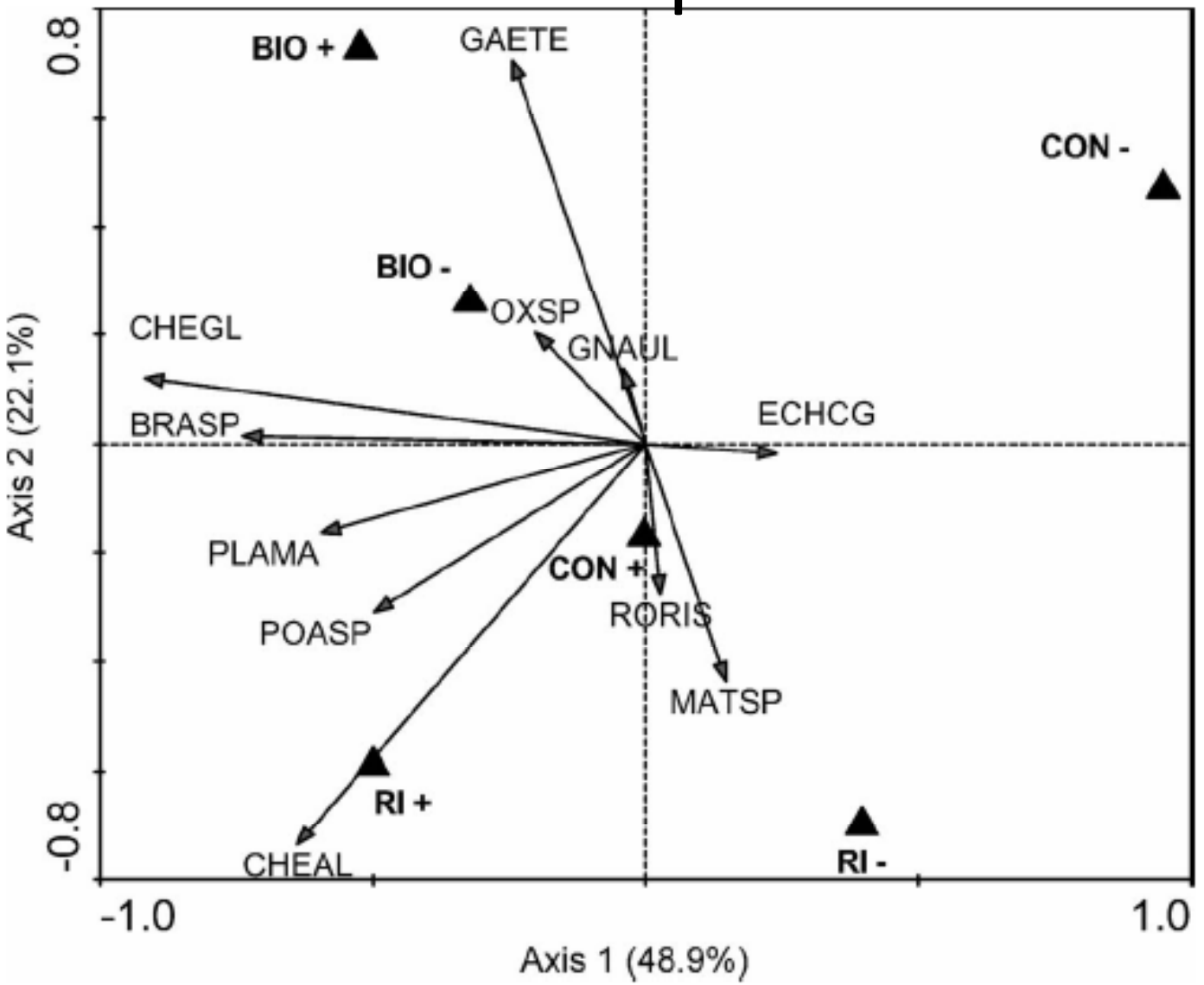
Redundancy Analysis: Example

TABLE 2. Effects of pest management system (P), soil management system (S), and rotation entry point (C) on weed seed density of all species occuring in t

Latin name	Common name	Bayer code	P	C	S	P by C	C by S
			Prob. > *F*				
Multivariate analysis			0.002	< 0.001	< 0.001	0.208	0.048
Univariate analyses							
Brassica sp.	Mustard	BRASP[a]	0.037	0.148	0.169	0.812	0.014
Capsella bursa-pastoris (L.) Medik.	Shepherd's-purse	CAPBP	0.741	0.220	0.028	0.496	0.220
Chenopodium album L.	Common lambsquarters	CHEAL	< 0.001	0.077	0.038	0.019	0.043
Chenopodium glaucum (L.)	Oakleaf goosefoot	CHEGL	0.120	0.375	< 0.001	0.804	0.401
Echinochloa crus-galli (L.) Beauv.	Barnyardgrass	ECHCG	0.145	< 0.001	0.363	0.370	0.441
Galeopsis tetrahit L.	Common hempnettle	GAETE	0.005	0.027	0.542	0.462	0.298
Gnaphalium uliginosum L.	Low cudweed	GNAUL	0.835	0.051	0.195	0.165	0.485
Matricaria sp.	Chamomile	MATSP	0.299	0.338	0.898	0.892	0.809
Oxalis sp.	Woodsorrel	OXSP	0.906	0.650	0.349	0.188	0.101
Plantago major L.	Broadleaf plantain	PLAMA	0.272	0.002	0.001	0.577	0.606
Poa sp.	Bluegrass	POASP	0.341	0.138	0.138	0.865	0.106
Polygonum aviculare L.	Prostrate knotweed	POLAV	0.820	0.255	0.406	0.281	0.922
Polygonum lapathifolium L.	Pale smartweed	POLLA	0.950	0.608	0.608	0.510	0.582
Rorippa palustris (L.) Bess.	Marsh yellowcress	RORIS	0.291	0.541	0.890	0.879	0.998
Spergula arvensis L.	Corn spurry	SPRAR	0.232	0.408	0.505	0.746	0.505
Taraxacum officinale G.H. Weber ex Wiggers	Dandelion	TAROF	0.980	0.864	0.728	0.369	0.082

[a] *Brassica* spp., *Matricaria* spp. and others were so categorized because the species within each group were often indistinguishable at the time data were collected. Acronyms codes in these cases.

Redundancy Analysis: Example



Redundancy Analysis

- Pros
  - Can both hypothesis test (*a priori*) and look for pattern (*a posteriori*)
  - Can deal with zeroes, more variables than observations, non-normality
  - updated versions can handle non-Euclidean distance measures (distance based redundancy analysis)
- Cons
  - pain in the neck
  - still lots of debate about whether non parametric MANOVA is more powerful

Observational Studies: Example

Funes-Monzote F.R. et al. 2009. Agro-Ecological Indicators (AEIs) for Dairy and Mixed Farming Systems Classification: Identifying Alternatives for the Cuban Livestock Sector. J. of Sustainable Agr. 33:435-460

- Selected 93 farms classified as experimental mixed livestock, commercial mixed livestock, or commercial specialized dairy
- Used a structured questionnaire that was jointly filled out by farmers and researchers

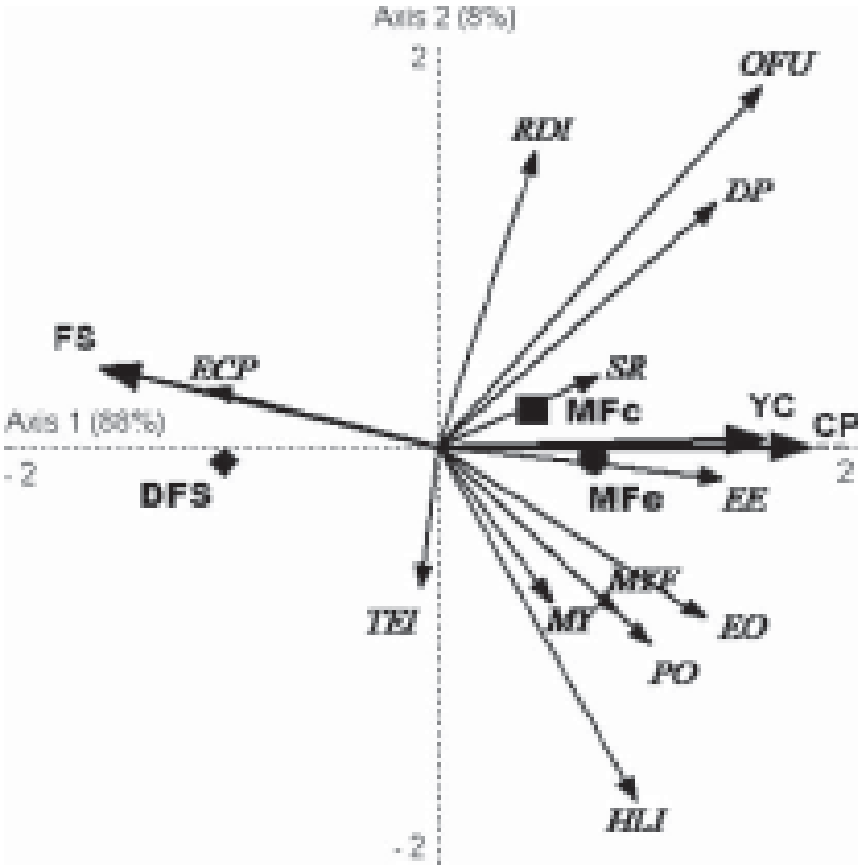
Observational Studies: Example

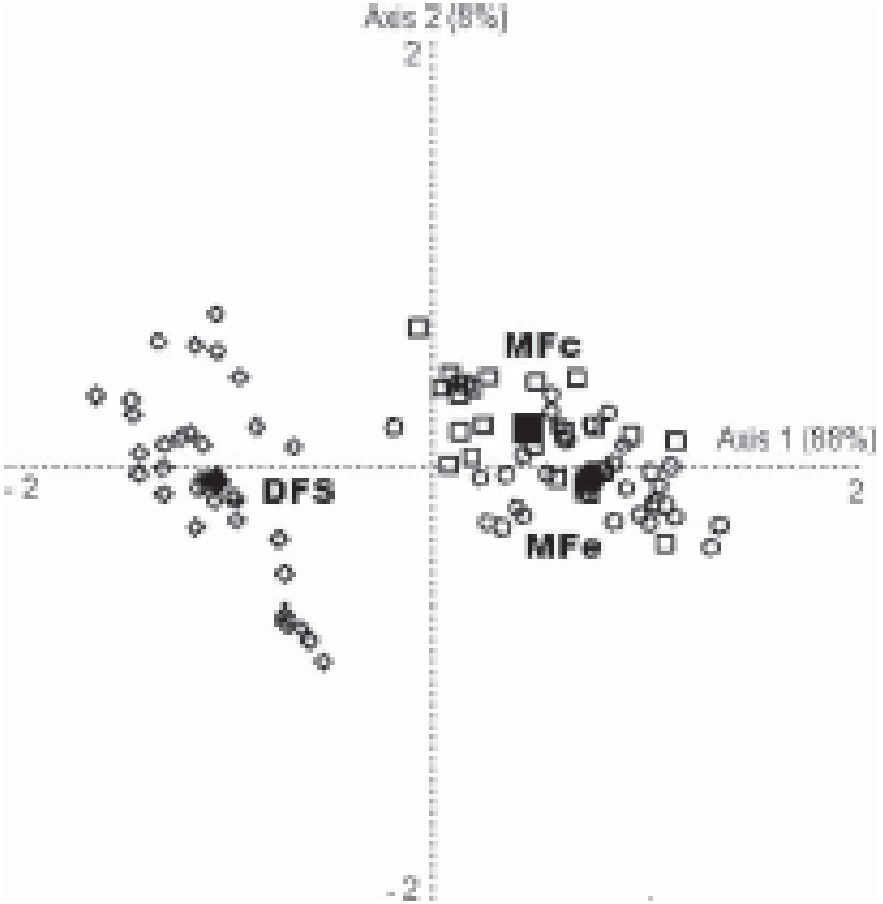
TABLE 5a Geometric Means of Agro-Ecological Indicators (AEIs) According to Type and Years Since Conversion (n = 93) a
Geometric Standard Deviation

		Type				Years since conversion		
AEIs	Unit	Mixed farming experimental n = 33	Mixed farming commercial n = 25	Dairy farming System n = 35	Geometric SD	3 or more n = 28	1 or 2 n = 30	N con n =
Species richness	Margalef index	8.8^{a}	6.0^{b}	2.5^{c}	1.51	8.6^{a}	6.6^{a}	
Diversity of production	Shannon index	1.8^{a}	2.0^{a}	0.3^{b}	1.23	1.9^{a}	1.9^{a}	
Reforestation index	Shannon index	1.5^{a}	1.8^{a}	0.7^{b}	1.64	1.6^{a}	1.7^{a}	
Milk yield (farm area)	Mg ha^{-1} yr^{-1}	1.5^{a}	1.2^{a}	0.7^{b}	1.79	1.6^{a}	1.1^{b}	
Milk yield (forage area)	Mg ha^{-1} yr^{-1}	2.4^{a}	1.7^{b}	0.7^{c}	1.85	2.6^{a}	1.7^{b}	
Energy output	GJ ha^{-1} yr^{-1}	16.0^{a}	11.7^{a}	2.5^{b}	1.71	16.6^{a}	11.9^{a}	
Protein output	kg ha^{-1} yr^{-1}	118.8^{a}	106.3^{a}	29.5^{b}	1.73	128.9^{a}	100.4^{a}	2
Total energy input	GJ ha^{-1} yr^{-1}	2.6^{c}	3.8^{ab}	3.6^{bc}	1.78	2.8 NS	3.4 NS	
Human labour intensity	hr ha^{-1} d^{-1}	3.6^{a}	1.6^{b}	0.8^{c}	1.87	2.9^{a}	2.2^{a}	
Energy cost of protein	MJ kg^{-1}	22.0^{c}	36.1^{b}	123.2^{a}	1.76	21.8^{c}	33.6^{b}	12
Energy efficiency	GJ output GJ^{-1} input	6.1^{a}	3.0^{b}	0.7^{c}	1.76	5.9^{a}	3.5^{b}	
Organic fertiliser use	Mg ha^{-1}	3.5^{a}	3.8^{a}	0.4^{b}	1.99	3.9^{a}	3.4^{a}	

Geometric means with different letters in superscript differ significantly ($p < 0.01$) between farm systems (Tukey's HSD). Approximate 95% tole within types is [geometric mean / (geometric SD)^2, geometric mean × (geometric SD)^2]. For calculation procedures of Shannon and Margalef in (2001).

Observational Studies: Example





Observational Studies

- Pros:
  - Studies a complicated system in place and allows for integration of economic, sociological, and environmental data
  - Sidesteps issue of whether researchers can recreate complex farming systems on research stations
  - May be the only avenue for studying some issues
- Cons:
  - Best suited for hypothesis generation, not testing
  - Spurious correlations can distract