Assignment 7: Analyze Ecological Data using Multivariate Regression and CCA

A. Sepúlveda-Jiménez

Data Science Dept., SoTE, CoBET, National University

DDS-8515: Multivariate Analysis

Course Instructor: Y. Karahan, PhD

November 18, 2025

Contents

List of Figures

Introduction

Environmental datasets routinely contain multiple correlated predictors and multivariate ecological responses such as community composition or biodiversity indices. In this paper, I combine multivariate regression (MR), regularized regression (ridge and lasso), and canonical correspondence analysis (CCA) to analyze vegetation and environmental data from Dutch dune meadows. Species richness and Shannon diversity are modeled as joint responses of soil and management variables, while the full species composition is related to environmental gradients through CCA. I compare how MR and CCA handle complex species–environment relationships, discuss regularization in

high-dimensional settings, and reflect on implications for ecological inference and environmental management.

Environmental and ecological data are inherently multivariate: multiple environmental drivers (e.g., soil properties, climate, management) simultaneously influence multiple ecological responses (e.g., species abundances, diversity indices). Classical univariate models often fail to account for the covariance structure among responses or the collinearity among predictors, leading to inefficient or misleading inference (Borcard et al., 2018; Legendre & Legendre, 2012; McGarigal et al., 2000).

Multivariate regression (MR) extends linear regression to vector-valued responses, explicitly modeling several dependent variables given a common set of predictors. In matrix form, MR can be written as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{n\times q}$ is a matrix of q responses for n sites, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a matrix of p predictors, $\mathbf{B} \in \mathbb{R}^{p\times q}$ contains regression coefficients, and $\mathbf{E} \in \mathbb{R}^{n\times q}$ is a matrix of errors. Ordinary least squares (OLS) estimates $\mathbf{B}$ by minimizing the Frobenius norm of residuals,

$$\widehat{\mathbf{B}}_{\text{OLS}} = \arg\min_{\mathbf{B}} \|\mathbf{Y} - \mathbf{XB}\|_F^2, \tag{2}$$

with closed-form solution $\widehat{\mathbf{B}}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ when $\mathbf{X}^\top\mathbf{X}$ is nonsingular.

In ecological applications, predictor collinearity is common and the number of predictors can be large relative to the number of sites, causing instability in OLS estimates (Hoerl & Kennard, 1970; Legendre & Legendre, 2012). Penalized regression methods such as ridge regression and the lasso introduce bias to reduce variance and perform automatic feature selection (Hastie et al., 2009; Hoerl & Kennard, 1970; James et al., 2021; Tibshirani, 1996). For linear models with response $\mathbf{Y}$ and design matrix $\mathbf{X}$, ridge regression solves

$$\widehat{\mathbf{B}}_{\text{ridge}} = \arg\min_{\mathbf{B}} \left(\|\mathbf{Y} - \mathbf{XB}\|_F^2 + \lambda\|\mathbf{B}\|_F^2\right), \tag{3}$$

while the multivariate lasso solves

$$\widehat{\mathbf{B}}_{\text{lasso}} = \arg\min_{\mathbf{B}} \left(\|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda \sum_{j=1}^{p} \sum_{k=1}^{q} |b_{jk}| \right), \tag{4}$$

where $\lambda > 0$ controls the amount of shrinkage.

For community composition, ordination techniques provide low-dimensional representations of species and sites. Canonical correspondence analysis (CCA) is a direct gradient analysis method that relates species composition to measured environmental variables using a unimodal response model and χ^2 distances (Legendre & Legendre, 2012; ter Braak, 1986). CCA extends correspondence analysis by constraining site scores to be linear combinations of environmental variables, effectively combining regression and ordination. In matrix notation, let $\mathbf{Y}$ denote the species abundance matrix and $\mathbf{X}$ the environmental predictors. CCA finds site scores $\mathbf{U}$ and species scores $\mathbf{V}$ that satisfy weighted averaging constraints and solve a generalized eigenvalue problem associated with the constrained χ^2 covariance structure (Borcard et al., 2018; ter Braak, 1986).

In this paper, I analyze the Dutch dune meadow vegetation dataset (Jongman et al., 1987; Oksanen et al., 2020) using MR, ridge, lasso, and CCA. The goals are to: (a) quantify how soil and management factors jointly affect species richness and Shannon diversity; (b) identify environmental gradients structuring the full community composition; and (c) compare the strengths and limitations of MR and CCA for ecological inference and environmental decision-making.

Data

Study system and datasets

The data come from vegetation surveys of dune meadows on the Dutch island of Terschelling, originally collected to study the influence of grassland management on plant communities (Jongman et al., 1987). The dataset is distributed with the `vegan` R package as `dune` (species abundances) and `dune.env` (environmental variables) (Oksanen et al., 2020). The species table `dune` records Braun–Blanquet cover-abundance classes for 30

plant species in 20 sites (2 m × 2 m plots) (Legendre & Legendre, 2012). The corresponding environmental table `dune.env` contains five variables describing each site: soil A1 horizon thickness, soil moisture, management type, land use, and manure level (vegan development team, 2019).

Environmental variables are:

- A1: thickness of the soil A1 horizon (cm), numeric.
- Moisture: ordinal soil moisture class with levels $1 < 2 < 4 < 5$.
- Management: grassland management type; BF (biological farming), HF (hobby farming), NM (nature conservation management), SF (standard farming).
- Use: ordered land-use class; Hayfield < Haypastu < Pasture.
- Manure: ordered class of manure application; $0 < 1 < 2 < 3 < 4$.

(Legendre & Legendre, 2012; vegan development team, 2019).

Derived response variables for multivariate regression

To implement MR, I derive two site-level biodiversity responses from the species table:

1. Species richness: $R_i = \sum_{j=1}^{S} \mathbf{1}\{y_{ij} > 0\}$, counting the number of species with positive cover at site i.
2. Shannon diversity: $H_i' = -\sum_{j=1}^{S} p_{ij} \log p_{ij}$, where $p_{ij} = y_{ij} / \sum_{k=1}^{S} y_{ik}$ is the relative cover of species j in site i.

These responses summarize community structure in terms of richness and evenness and are standard in numerical ecology (Borcard et al., 2018; Legendre & Legendre, 2012).

The predictor matrix for MR includes A1 as a continuous variable and one-hot encoded versions of the categorical environmental factors (Moisture, Management, Use, Manure), with the first level of each factor serving as reference. All numeric predictors are standardized to zero mean and unit variance prior to estimation.

Species and environmental matrices for CCA

CCA uses the full species abundance matrix $\mathbf{Y} \in \mathbb{R}^{20\times 30}$ (sites × species) and a set of environmental predictors $\mathbf{X} \in \mathbb{R}^{20\times p}$. I include A1 and all categorical environmental variables encoded as dummy variables, as is common in constrained ordination (Borcard et al., 2018; ter Braak, 1986). Species abundances are treated as non-negative community data, and CCA uses χ^2 distances and row/column standardization internally (Legendre & Legendre, 2012; ter Braak, 1986).

Methods: Multivariate Regression and Regularization

Multivariate linear model

Let $\mathbf{y}_i = (R_i, H_i')^\top$ denote the two-dimensional response vector for site i and $\mathbf{x}_i$ the p-dimensional vector of environmental predictors (including dummy variables). Stacking rows gives

$$\mathbf{Y} = \begin{bmatrix} R_1 & H_1' \\ \vdots & \vdots \\ R_n & H_n' \end{bmatrix} \in \mathbb{R}^{n\times 2}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} \in \mathbb{R}^{n\times p}, \tag{5}$$

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \tag{6}$$

where $\mathbf{B} \in \mathbb{R}^{p\times 2}$ is the coefficient matrix and $\mathbf{E}$ contains residuals assumed to have mean zero and covariance matrix $\mathbf{\Sigma}_\varepsilon$ across responses. Under standard multivariate normal assumptions, maximum-likelihood and least-squares estimation coincide, yielding $\widehat{\mathbf{B}}_{\text{OLS}}$ as above (Legendre & Legendre, 2012; McGarigal et al., 2000).

For each response $k \in \{1, 2\}$, I report R_k^2 and adjusted $R_{k,\text{adj}}^2$,

$$R_{k,\text{adj}}^2 = 1 - \frac{\text{RSS}_k/(n-p)}{\text{TSS}_k/(n-1)}, \tag{7}$$

where $\text{RSS}_k = \sum_i (y_{ik} - \hat{y}_{ik})^2$ and $\text{TSS}_k = \sum_i (y_{ik} - \bar{y}_k)^2$. I also compute mean squared error ($\text{MSE}_k = \text{RSS}_k/n$) and inspect residual diagnostics (Q–Q plots and residuals vs. fitted plots) for each response (James et al., 2021).

Ridge regression

To stabilize estimates in the presence of collinearity among environmental predictors, ridge regression adds an ℓ_2 penalty on the coefficients. In the multivariate setting, the ridge estimator solves

$$\widehat{\mathbf{B}}_{\text{ridge}} = \arg\min_{\mathbf{B}} \left\{ \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda \|\mathbf{B}\|_F^2 \right\}, \tag{8}$$

where $\|\mathbf{B}\|_F^2 = \sum_{j,k} b_{jk}^2$ is the squared Frobenius norm. The closed-form solution is

$$\widehat{\mathbf{B}}_{\text{ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{Y}, \tag{9}$$

with λ typically selected by cross-validation or information criteria (Hastie et al., 2009; Hoerl & Kennard, 1970). Ridge shrinks coefficients toward zero but does not set them exactly to zero, so all predictors remain in the model.

Multitask lasso

The lasso introduces an ℓ_1 penalty, encouraging sparse solutions. For a multivariate response, the multitask lasso solves

$$\widehat{\mathbf{B}}_{\text{lasso}} = \arg\min_{\mathbf{B}} \left\{ \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda \sum_{j=1}^{p} \|\mathbf{b}_{j\cdot}\|_2 \right\}, \tag{10}$$

where $\mathbf{b}_{j\cdot}$ is the jth row of $\mathbf{B}$ and the group-lasso penalty induces row-wise sparsity, selecting or dropping whole predictors for all responses simultaneously (Hastie et al., 2009; Tibshirani, 1996). I use cross-validated multitask lasso in `scikit-learn` (Pedregosa et al., 2011) to compare selected environmental drivers to the OLS and ridge solutions.

Implementation details

All MR and regularized models are implemented with optimized pipelines in Python using `pandas` for data manipulation and `scikit-learn` for preprocessing, regression, and cross-validation (Pedregosa et al., 2011). Numeric predictors are standardized and categorical predictors one-hot encoded via a `ColumnTransformer`, embedded in a `Pipeline` to avoid data leakage between preprocessing and model fitting.

A correlation heatmap of main predictors and responses is shown in Figure 1. Residual diagnostics for richness and Shannon diversity appear in Figures 2–5. A comparison of coefficient magnitudes for OLS, ridge, and lasso is given in Figure 6.

Methods: Canonical Correspondence Analysis

CCA formulation

Canonical correspondence analysis is a constrained ordination method that relates community composition $\mathbf{Y}$ (sites $\times$ species) to environmental predictors $\mathbf{X}$ (Legendre & Legendre, 2012; ter Braak, 1986). CCA assumes unimodal species responses along underlying environmental gradients and uses χ^2 distances associated with correspondence analysis. The algorithm can be described in three conceptual steps (Borcard et al., 2018; ter Braak, 1986):

1. Perform a weighted linear regression of species abundances on the environmental variables to obtain fitted values $\hat{\mathbf{Y}}$.

2. Conduct correspondence analysis on $\hat{\mathbf{Y}}$, yielding canonical site scores and species scores.

3. Extract eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$ and corresponding canonical axes that maximize the dispersion of site scores constrained by $\mathbf{X}$.

In matrix terms, CCA solves a generalized eigenvalue problem of the form

$$\mathbf{Z}^\top \mathbf{W} \mathbf{Z}\, \mathbf{u} = \lambda \mathbf{Z}^\top \mathbf{W} \mathbf{Z}\, \mathbf{u}, \tag{11}$$

where $\mathbf{Z}$ is a suitably standardized species matrix and $\mathbf{W}$ contains row weights (Legendre & Legendre, 2012). The resulting site scores can be expressed as linear combinations of the environmental variables, and species scores follow from weighted averaging of site scores (ter Braak, 1986).

Implementation with scikit-bio

CCA is implemented in Python using the `cca` function from the `skbio.stats.ordination` module (scikit-bio development team, 2024). The species matrix **Y** is the raw `dune` table (sites as rows, species as columns). The environmental matrix **X** is constructed from `dune.env` by:

1. Retaining A1 as a numeric variable.
2. Encoding Moisture, Management, Use, and Manure as dummy variables (reference levels dropped).

`scikit-bio` returns an `OrdinationResults` object containing eigenvalues, canonical site scores, species scores, and biplot scores for environmental variables (scikit-bio development team, 2025). I visualize the first two canonical axes in a CCA triplot: sites colored by management, species scores, and environmental vectors (Figures 7–8).

Results

Data exploration

The correlation heatmap (Figure 1) shows that species richness and Shannon diversity are positively associated with soil A1 thickness and intermediate moisture classes, while high manure levels tend to coincide with slightly reduced diversity. Management categories also show distinct patterns: nature conservation management (NM) and hobby farming (HF) tend to occur at sites with thicker A1 horizons and moderate manure, whereas standard farming (SF) appears more often at thinner soils and higher manure levels. These patterns are broadly consistent with previous analyses of the dune data (Borcard et al., 2018; Legendre & Legendre, 2012).

Multivariate regression and regularization

The multivariate OLS model explains a substantial portion of variance in both responses. Adjusted R^2 values indicate that a combination of A1, Moisture, and Management accounts for much of the variation in richness and Shannon diversity, while

Use and Manure provide secondary contributions. Residual Q–Q plots (Figures 2 and 4) suggest approximate normality with mild deviations in the tails, which is acceptable given the small sample size ($n = 20$). Residuals versus fitted plots (Figures 3 and 5) show no strong heteroscedasticity or nonlinear structure, supporting the use of linear MR for these aggregated biodiversity responses.

Ridge regression shrinks coefficients toward zero while maintaining a similar overall pattern: the strongest standardized effects correspond to the moisture gradient and management contrasts involving nature conservation and standard farming. The ridge penalty reduces the sensitivity of coefficient estimates to collinearity between moisture and management factors, as expected (Hastie et al., 2009; Hoerl & Kennard, 1970).

The multitask lasso further simplifies the model by setting several small coefficients exactly to zero (Figure 6). In particular, some manure levels and nuanced Use categories drop out, leaving a parsimonious model dominated by A1, coarse moisture categories, and major management contrasts. This reflects the lasso's role as a joint feature selector across both richness and diversity responses (James et al., 2021; Tibshirani, 1996). In-sample MSE increases slightly when moving from ridge to lasso, consistent with the bias–variance trade-off, but the lasso gains interpretability by focusing attention on a smaller subset of ecologically meaningful predictors.

Canonical correspondence analysis

The CCA eigenvalue spectrum (Figure 7) shows that the first two canonical axes capture a large share of the constrained inertia in species composition, consistent with prior analyses of this dataset (Borcard et al., 2018; Legendre & Legendre, 2012). The CCA biplot of sites and environmental variables (Figure 8) reveals strong structuring along a moisture and management gradient: sites with nature conservation management (NM) and wetter soils cluster on one side of axis 1, while standard farming (SF) and drier soils lie on the opposite side. Manure and Use act as secondary gradients, rotating around the main moisture–management axis.

Species scores (Figure 9) indicate that some species are strongly associated with wet, low-manure conditions and conservation management, while others are characteristic of drier, heavily fertilized, standard-farmed sites. These unimodal distributions along the canonical axes are precisely the patterns that CCA is designed to reveal (Legendre & Legendre, 2012; ter Braak, 1986).

Whereas MR operates on aggregated biodiversity responses, CCA uses the full species matrix and thus can identify species that are particularly sensitive to management regimes or soil conditions, even if their contribution to richness and Shannon diversity is modest. This is valuable for conservation planning, where indicator species and community turnover matter as much as overall diversity (Legendre & Legendre, 2012; McGarigal et al., 2000).

Comparison of MR and CCA

MR and CCA provide complementary views of the same system. MR models the conditional mean of a small set of summary responses (richness, diversity) given environmental drivers and is naturally suited for prediction and effect-size estimation, including regularized variants like ridge and lasso that handle collinearity and overfitting (Hastie et al., 2009; James et al., 2021). CCA, by contrast, is an ordination method that emphasizes gradient detection and community-level structure, focusing less on prediction and more on interpretation of species–environment relationships (Borcard et al., 2018; Legendre & Legendre, 2012; ter Braak, 1986).

In the dune data, both approaches highlight moisture and management as dominant drivers. MR quantifies their effects on richness and Shannon diversity, and regularization clarifies which combinations of management and land use have the strongest associations. CCA confirms these gradients at the level of full species composition, identifying groups of species and sites that occupy distinct niches along the moisture–management axis.

From a methodological standpoint, MR (including ridge and lasso) is grounded in the linear regression framework with familiar diagnostics, confidence intervals, and

prediction metrics (MSE, R^2). CCA relies on ordination geometry and eigenvalue decomposition, with interpretation centered on ordination diagrams and explained inertia. For questions about *how much* diversity changes with a given management intervention, MR and its regularized variants are natural tools. For questions about *which species* shift and *how communities reorganize* along environmental gradients, CCA is more informative.

Reflection and Environmental Applications

This analysis illustrates how combining MR, ridge, lasso, and CCA can sharpen ecological inference from multivariate environmental data. Key findings include:

- Soil thickness (A1), moisture, and management regimes exert the strongest influence on both biodiversity summaries and full community composition.
- Ridge regression stabilizes coefficient estimates in the presence of collinearity, while the lasso yields sparse models that highlight a subset of ecologically important management and soil variables.
- CCA reveals coherent species assemblages aligned along the moisture–management gradient, suggesting that changing management practices or manure applications can shift communities between distinct compositional states.

In terms of environmental policy and conservation, these methods support different but complementary decisions. MR and regularization can be used to predict changes in biodiversity indices under alternative management scenarios, quantify expected gains in richness from reducing manure levels, or rank management actions by their predicted impact on diversity. CCA provides insight into turnover in species composition, aiding the identification of indicator species, vulnerable community types, and trade-offs between agricultural production and conservation goals.

Several challenges and limitations emerged. First, the sample size is small ($n = 20$ sites), which constrains the complexity of models and the reliability of significance tests. Regularization helps but cannot fully compensate for limited replication. Second, both MR

and CCA assume relatively simple relationships (linear or unimodal) and can be sensitive to outliers and transformations; in larger or more complex datasets, generalized linear models or generalized additive models might be preferable (Borcard et al., 2018). Third, while ridge and lasso aid feature selection, ecological interpretation requires domain knowledge to avoid over-interpreting statistically selected predictors (Legendre & Legendre, 2012; McGarigal et al., 2000).

Nonetheless, the combination of MR, regularization, and CCA forms a robust toolkit for environmental and conservation science. In climate change research, similar workflows can link shifting climate regimes to biodiversity responses across multiple taxa. In pollution studies, MR and lasso can identify key pollutants driving declines in sensitive species, while CCA can reveal how community structure reorganizes along contamination gradients. For conservation planning, these tools help prioritize management interventions that most effectively maintain both diversity and community composition.

Conclusion

Multivariate regression, ridge regression, lasso, and canonical correspondence analysis address complementary aspects of ecological data analysis. Using a real vegetation dataset from Dutch dune meadows, I showed how environmental gradients and management regimes shape biodiversity indices and full community composition. MR and its regularized extensions quantify effect sizes and improve predictive stability under collinearity, while CCA extracts dominant environmental gradients and species assemblages. Together, these tools support more nuanced and informed decisions in environmental policy and conservation planning.

References

Borcard, D., Gillet, F., & Legendre, P. (2018). *Numerical ecology with r* (2nd). Springer. https://doi.org/10.1007/978-3-319-71404-2

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

James, G., et al. (2021). *An introduction to statistical learning: With applications in r* (2nd). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Jongman, R. H. G., ter Braak, C. J. F., & van Tongeren, O. F. R. (1987). *Data analysis in community and landscape ecology*. Pudoc.

Legendre, P., & Legendre, L. (2012). *Numerical ecology* (3rd, Vol. 24). Elsevier.

McGarigal, K., Cushman, S., & Stafford, S. (2000). *Multivariate statistics for wildlife and ecology research*. Springer. https://doi.org/10.1007/978-1-4612-1288-1

Oksanen, J., et al. (2020). *`vegan`: Community ecology package* [R package version 2.5-7]. R Foundation for Statistical Computing. Vienna. https://cran.r-project.org/package=vegan

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

scikit-bio development team. (2024). Canonical correspondence analysis in `scikit-bio` [Accessed 2025-11-17].

scikit-bio development team. (2025). Scikit-bio: Ordination and multivariate statistics in python. *scikit-bio*. https://scikit.bio/docs/dev/generated/skbio.stats.ordination.html

ter Braak, C. J. F. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. *Ecology*, *67*(5), 1167–1179. https://doi.org/10.2307/1938672

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288.

vegan development team. (2019). `dune` and `dune.env`: Vegetation and environment in dutch dune meadows.

Appendix: Figures

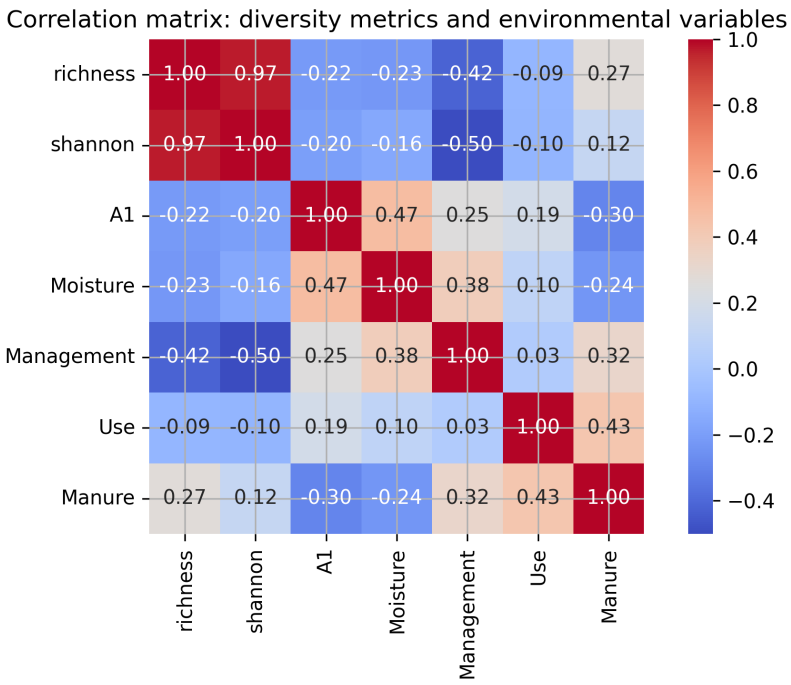

Figure 1

Correlation matrix of derived biodiversity responses (richness and Shannon diversity) and key environmental predictors (A1, moisture, management, use, manure).

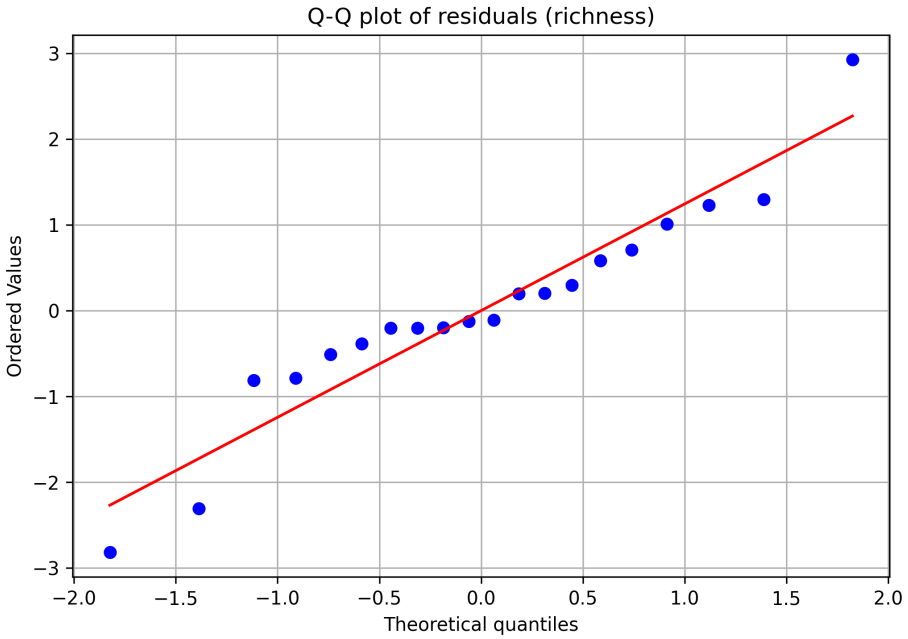

Figure 2

Q–Q plot of residuals for the richness regression component of the multivariate OLS model.

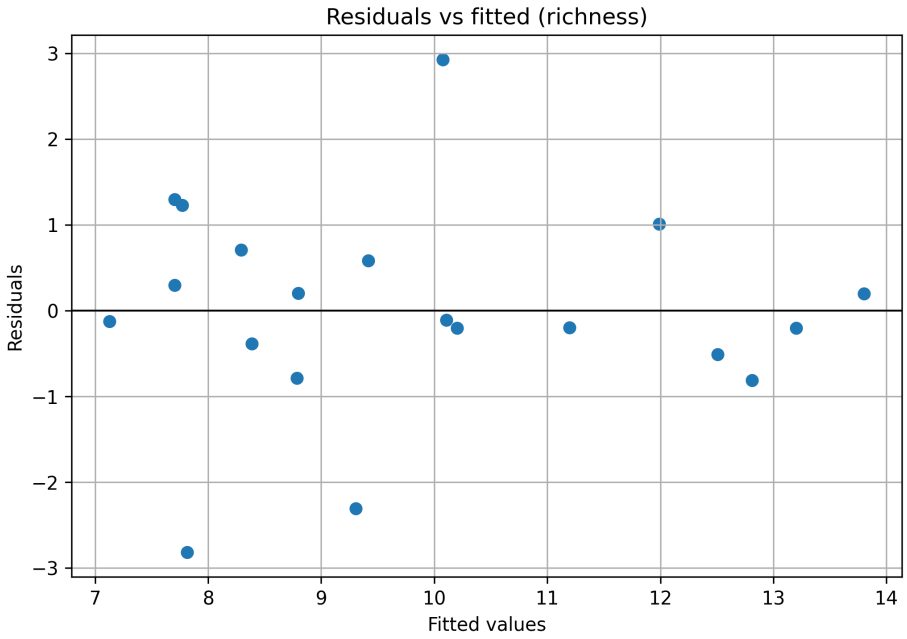

Figure 3

Residuals versus fitted values for the richness regression component.

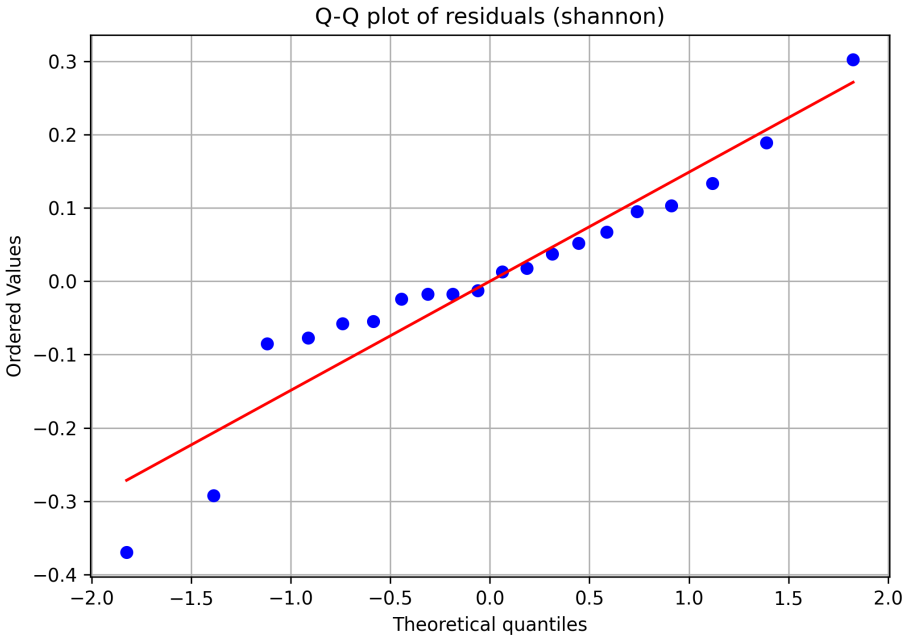

Figure 4

Q–Q plot of residuals for the Shannon diversity regression component.

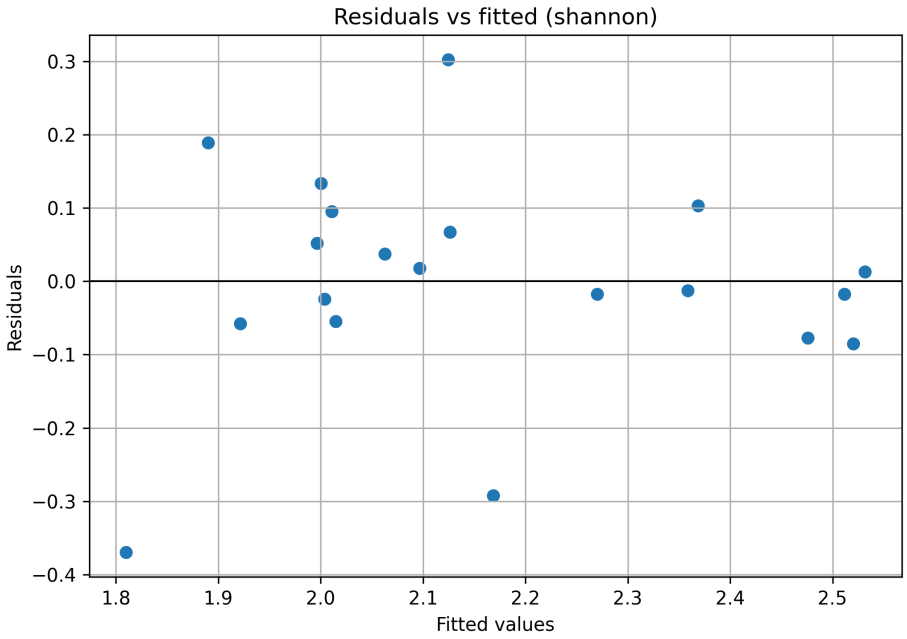

Figure 5

Residuals versus fitted values for the Shannon diversity regression component.

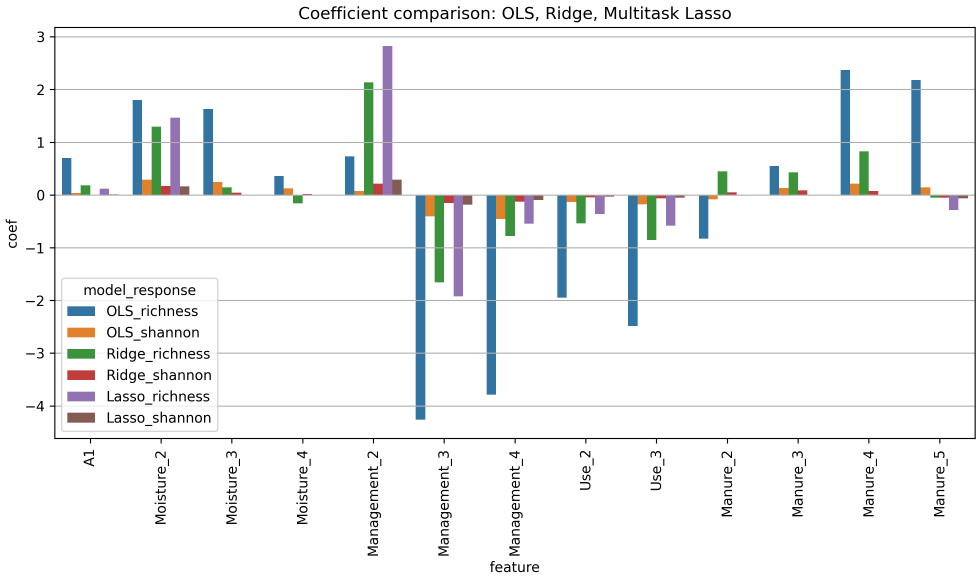

Figure 6

Standardized coefficient magnitudes for OLS, ridge, and multitask lasso models, comparing the inferred influence of each environmental predictor across methods.

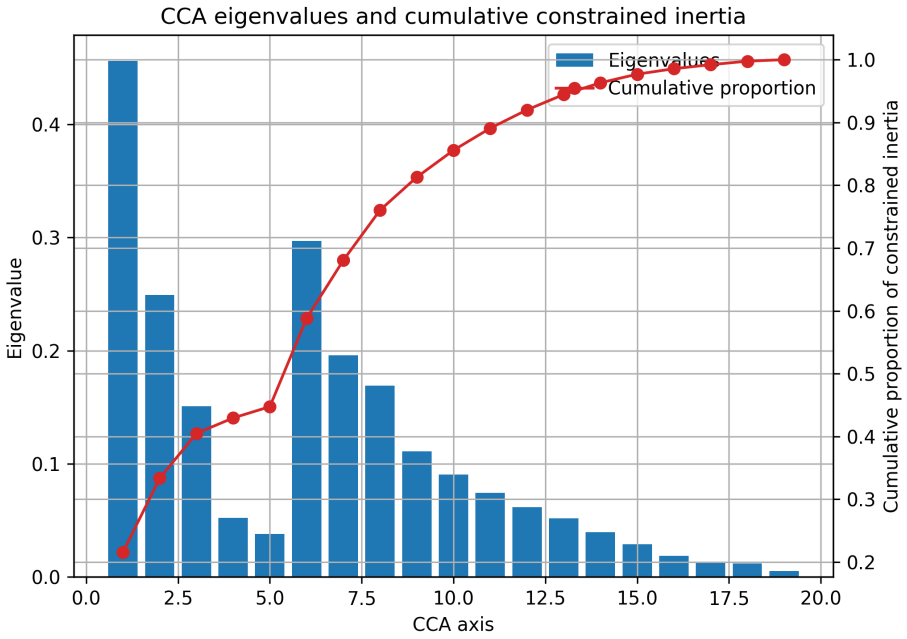

Figure 7

Canonical correspondence analysis eigenvalues and cumulative proportion of constrained inertia for the first few axes.

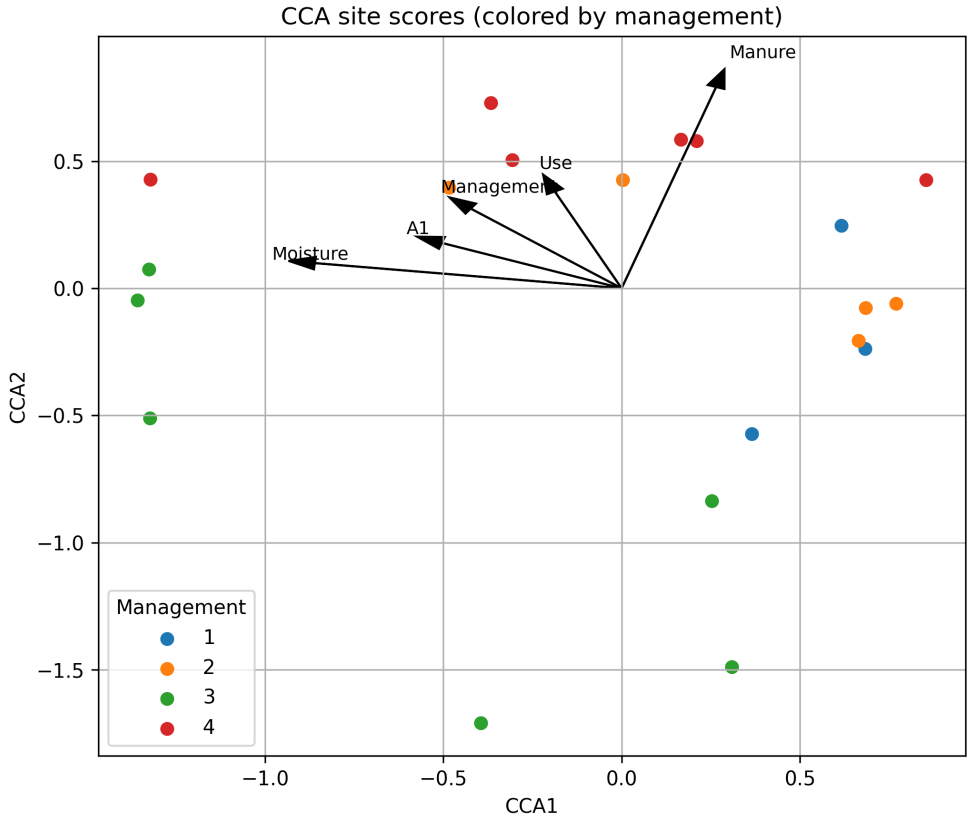

Figure 8

CCA biplot of site scores on the first two canonical axes, colored by management type, with environmental vectors overlaid.

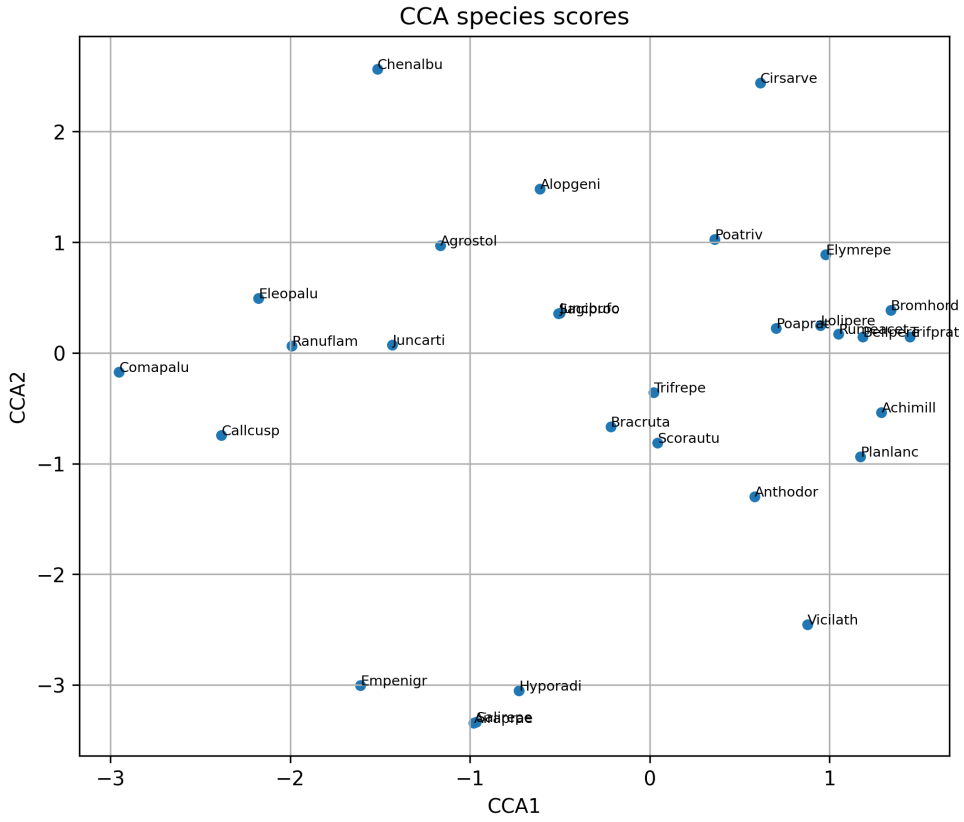

Figure 9

CCA species scores on the first two canonical axes, highlighting species associated with distinct combinations of moisture, management, and manure.