ISSN: 2393-8048
International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Multilingual, Indexed, Double Blind, Open Access, Peer-Reviewed, Refereed-International Journal.
SJIF Impact Factor = 8.35, January-June 2026, Submitted in May 2026

Artificial Intelligence and Multivariate Statistics in Ecological Data Modeling: A Hybrid Analytical Framework

Jag Pratap Singh Yadav, Research Scholar, Mathematics Department, Sikkim Alpine University, Namchi (Sikkim)
Dr. Umashanker, Assistant Professor, Physical Mathematics Department, Sikkim Alpine University, Namchi (Sikkim)

Abstract

Data from species abundances, environmental gradients, remote sensing, and long-term monitoring networks contribute to modern ecological research, which generates high-dimensional, noisy, and non-linearly structured data. While principal component analysis, canonical correspondence analysis, redundancy analysis, and multivariate analysis of variance are classical multivariate statistical methods that give interpretable decompositions of variance and covariance structure, these methods frequently fail to account for real-world ecosystems because they depend on linearity or unimodality assumptions. Deep architectures, support vector machines, random forests, and artificial neural networks can approximate any non-linear response surface. However, these methods often face criticism for not being very interpretable and for not taking into account the covariance structure that is revealed by multivariate statistics. This research presents a comprehensive mathematical framework for ecological data modeling that integrates function approximation based on artificial intelligence with multivariate statistical dimensionality reduction. Our work includes formalizing the operators used in numerical ecology for covariance and ordination, outlining the learning-theoretical basis of the AI models used for ecological prediction the most, and developing a hybrid estimator that feeds a neural predictor with latent variables that have been statistically extracted. Using measures such as explained variance, classification accuracy, and mean squared error, the hybrid model is evaluated against both statistical and AI-based baselines in a simulated multivariate ecology dataset. The results show that the hybrid approach keeps the interpretability benefits of ordination-based analysis while improving predictive accuracy. Our final section covers the theoretical constraints, computing expenses, and potential avenues for further investigation into AI-assisted multivariate ecological modeling.

Keywords: multivariate statistics, artificial intelligence, ecological modeling, principal component analysis, canonical correspondence analysis, neural networks, random forests, biodiversity prediction

1. Introduction

Ecological systems are characterized by high dimensionality, non-linear interactions among taxa and abiotic drivers, spatial and temporal autocorrelation, and substantial observational noise. A single ecological survey may record dozens of species' abundances alongside soil chemistry, hydrology, climate, and land-use variables, producing a data matrix whose rows are sampling sites and whose columns span both community composition and environmental covariates. Extracting meaningful ecological signal from such matrices has historically relied on multivariate statistics, a body of theory developed to summarize covariance structure, ordinate samples along latent ecological gradients, and test hypotheses about group differences in multi-species assemblages.

Over the last two decades, artificial intelligence (AI) — and particularly machine learning and deep learning — has been increasingly adopted in ecology for tasks such as species distribution modeling, remote-sensing classification, population forecasting, and automated acoustic or camera-trap species identification. These methods can approximate highly non-linear, high-order interaction effects that classical linear or unimodal statistical models cannot capture. However, AI models are frequently treated as opaque predictors: they rarely expose the orthogonal latent axes of variation that ecologists use to interpret community structure, and

they can overfit small or spatially clustered ecological datasets when applied without appropriate regularization.

This paper argues that multivariate statistics and artificial intelligence are not competing paradigms but complementary components of a single estimation pipeline. Multivariate statistical ordination provides a mathematically well-understood, variance-maximizing (or inertia-maximizing) reduction of the data to a low-dimensional latent space; artificial intelligence provides a flexible non-linear mapping from that latent space, or from the raw covariates, to the ecological response of interest. We present the mathematical foundations of both families of methods within a common notation, propose a hybrid estimator, and evaluate it on a simulated multivariate ecological dataset designed to reflect realistic community–environment relationships.

The remainder of the paper is organized as follows. Section 2 develops the mathematical formalism of multivariate statistical methods used in numerical ecology. Section 3 reviews the artificial intelligence models most commonly applied to ecological data and states their learning objectives in the same linear-algebraic notation. Section 4 introduces the proposed hybrid framework and its associated algorithm. Section 5 describes a simulated case study and its experimental design. Section 6 presents statistical validation procedures. Section 7 reports and discusses the results. Section 8 addresses limitations and future directions, and Section 9 concludes.

2. Mathematical Foundations of Multivariate Statistics in Ecology

Let $X \in \mathbb{R}^{nxp}$ denote a community data matrix of n sampling sites and p species (or taxa), and let $E \in \mathbb{R}^{nxg}$ denote a matrix of q environmental covariates measured at the same sites. Multivariate statistical ecology seeks linear or unimodal transformations of X, possibly constrained by E, that reveal the dominant axes of ecological variation.

Principal Component Analysis

Principal component analysis (PCA) seeks an orthogonal linear transformation that maximizes the variance captured by successive components. Let X be centered so that each column has zero mean, and define the sample covariance matrix

$$C = (1 / (n - 1)) X^T X, \quad C \in \mathbb{R}^{pxp}$$

The principal axes are the eigenvectors of C, obtained from the eigen-decomposition

$$C v_k = \lambda_k v_k, \quad k = 1, \ldots, p, \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

where v_k is the k^{th} eigenvector (principal axis) and λ_k its associated eigenvalue, representing the variance explained along that axis. The site scores on the k^{th} axis are given by the projection $F_k = X v_k$, and the proportion of total variance explained by the first m components is

$$R^2 = 1 - \sum_{k=1}^{m} \lambda k \; / \; \sum_{j=1}^{p} \lambda j$$

Without taking environmental drivers into account, principal component analysis (PCA) describes variation in species matrix X, providing an unconstrained ordination. This makes it great for exploratory purposes, but it will not cut it if you are trying to establish a direct correlation between community structure and measurable variables.

Canonical Correspondence Analysis and Redundancy Analysis

Two restricted ordination approaches, canonical correspondence analysis (CCA) and redundancy analysis (RDA), limit the retrieved axes to linear combinations of the environmental factors E. The first step of RDA is to use multivariate multiple regression to regress the species matrix X onto E.

$$\hat{X} = E (E^T E)^{-1} E^T X = H X$$

where H is the projection (hat) matrix of the regression. The constrained ordination axes are then the eigenvectors of the covariance matrix of the fitted values $\hat{X}$, namely

$$C_{RDA} v_k = \lambda_k v_k, \quad C_{RDA} = (1 / (n - 1)) \hat{X}^T \hat{X}$$

To account for the unimodal species response to environmental gradients commonly seen in field ecology, CCA uses a weighted regression under a chi-square distance metric instead of ordinary least squares. This approach is suitable for count or presence-absence community data. The environmental predictors account for a certain amount of species variation along each canonical axis, as measured by the eigenvalues λ^k from both methods. The fraction of community variation attributable to the measured environment is obtained by dividing the sum of constrained eigenvalues by total inertia.

Multivariate Analysis of Variance and Discriminant Functions

Multivariate analysis of variance (MANOVA) examines if group mean vectors differ in the p-dimensional response space when sample sites are grouped by categorical criteria (e.g., habitat type, treatment,, area). W. Wilks' lambda statistic, the between-group and within-group sums-of-squares-and-cross-products matrices B and μg, the grand mean, are defined.

$$\Lambda = |W| / |B + W|$$

is employed to determine whether the null hypothesis H_0: $\mu_1 = \mu_2 = \ldots = \mu_G$ holds true using a multivariate F-approximation. To apply this framework to classification, linear discriminant analysis finds the linear combination $a = W^{-1}(\mu_g - \mu_h)$ that maximizes the ratio of between-group to within-group variance. This method serves as both a statistical test and a predictive classifier for ecological outcomes that can be categorized, like habitat type or disturbance class.

Cluster Analysis and Distance-Based Methods

Unsupervised classification of ecological communities frequently relies on distance metrics such as Bray–Curtis dissimilarity,

$$d_{BC}(i,j) = \frac{\sum_{k=1}^{p} |x_{ik} - x_{jk}|}{\sum_{k=1}^{p} (x_{ik} + x_{jk})}$$

combined with hierarchical agglomerative clustering or k-means partitioning, the latter minimizing the within-cluster sum of squares

$$J = \sum_{c=1}^{K} \sum_{i \in c} \|\mathbf{x}_i - \boldsymbol{\mu}_c\|^2$$

These distance-based methods complement eigen-based ordination by allowing non-Euclidean notions of ecological similarity, which are often more appropriate for sparse, zero-inflated community matrices.

3. Artificial Intelligence Methods for Ecological Data Modeling

Artificial intelligence methods relax the linearity and distributional assumptions of classical multivariate statistics by learning flexible, often non-parametric, mappings between covariates and ecological responses. We summarize the four families most widely applied in ecology.

Artificial Neural Networks

A feedforward artificial neural network (ANN) computes a nested composition of affine transformations and non-linear activation functions. For an input vector $x \in \mathbb{R}^p$, a single hidden layer with r units produces

$$h = \varphi(W_1 x + b_1), \quad \hat{y} = W_2 h + b_2$$

where $W_1 \in \mathbb{R}^{rxp}$, $W_2 \in \mathbb{R}^{mxr}$ are weight matrices, b_1, b_2 are bias vectors, and $\varphi(\cdot)$ is a non-linear activation such as the logistic sigmoid or rectified linear unit. Parameters are estimated by minimizing a loss function $L(\hat{y}, y)$ — typically mean squared error for continuous ecological responses (e.g., biomass) or cross-entropy for categorical responses (e.g., presence/absence) — via stochastic gradient descent,

$$\theta \leftarrow \theta - \eta \, \nabla_\theta L(\hat{y}, y)$$

with learning rate η. Deep networks stack multiple hidden layers to represent higher-order interactions among covariates, which is particularly relevant for modeling threshold effects and interaction terms common in species–environment relationships, such as temperature–moisture interactions on species occurrence.

Random Forests and Ensemble Trees

Random forests construct an ensemble of B regression or classification trees, each grown on a bootstrap sample of the data with a random subset of covariates considered at each split, and aggregate their predictions by averaging (regression) or majority vote (classification):

$$\hat{y}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x})$$

Each tree T_b partitions covariate space by recursively choosing the split (variable j, threshold s) that minimizes the residual sum of squares within each resulting node. Random forests naturally accommodate non-linearities, variable interactions, and mixed data types, and provide variable-importance measures based on the reduction in node impurity attributable to each covariate, which offers a partial substitute for the interpretability of ordination axes.

Support Vector Machines

Support vector machines (SVMs) seek a maximum-margin separating hyperplane in a (possibly infinite-dimensional) feature space induced by a kernel function $\kappa(x_i, x_j)$. The classification decision function is

$$\hat{f}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i \kappa(\mathbf{x}_i, \mathbf{x}) + b\right)$$

where the coefficients α_i are obtained by solving a convex quadratic program that maximizes the margin subject to classification constraints. Common ecological applications include the radial basis kernel $\kappa(x\_i, x\_j) = \exp(-\gamma\|x\_i - x\_j\|^2)$, which allows smooth non-linear decision boundaries appropriate for species distribution classification from remote-sensing covariates.

Convolutional and Recurrent Architectures

When ecological data include spatial imagery (satellite or drone imagery for land-cover classification) or temporal sequences (long-term monitoring or bioacoustic recordings), convolutional neural networks (CNNs) and recurrent neural networks (RNNs) extend the basic ANN formalism with, respectively, weight-shared local receptive-field filters and recurrent state transitions,

$$s_t = \varphi(\, W_s\, s_{t-1} + W_x\, x_t + b\,)$$

These architectures allow ecological models to exploit spatial autocorrelation and temporal dynamics that are difficult to encode in classical multivariate statistical frameworks, at the cost of substantially larger parameter counts and correspondingly larger data requirements.

Gaussian Processes and Bayesian Non-Parametric Models

Gaussian process (GP) regression places a prior directly over the space of functions relating covariates to an ecological response, defined by a mean function m(x) (often zero) and a covariance kernel k(x, x′):

$$f(x) \sim GP(\, m(x),\, k(x, x')\,)$$

Given training data, predictions at a new site x* follow a multivariate normal distribution with closed-form mean and variance,

$$\mu^* = k^{*T}\,(K + \sigma^2 I)^{-1}\, y\,, \quad \sigma^{2*} = k(x^*, x^*) - k^{*T}\,(K + \sigma^2 I)^{-1}\, k^*$$

where K is the training kernel matrix and k* the vector of kernel evaluations between x* and the training covariates. Unlike ANNs and random forests, GPs provide calibrated predictive uncertainty at every site, which is valuable in conservation contexts where decision-makers must weigh the confidence of a model alongside its point prediction. The computational cost of exact GP inference scales as $O(n^3)$, which limits direct application to very large monitoring networks unless sparse or inducing-point approximations are used.

Bias-Variance Considerations for Ecological Sample Sizes

A recurring theme across the AI methods reviewed above is the bias-variance trade-off, which is particularly acute in ecology because field sampling is expensive and datasets rarely exceed a few hundred to a few thousand sites. The expected generalization error of a fitted model decomposes as

$$E[(y - \hat{g}(x))^2] = Bias(\hat{g}(x))^2 + Var(\hat{g}(x)) + \sigma^2_\varepsilon$$

While flexible models like unpruned decision trees and deep neural networks can achieve near-zero bias on training data, they suffer from poor generalization on held-out sites due to excessive variance when sample sizes are limited compared to the number of estimated parameters. In small-sample ecological contexts, multivariate statistical approaches like PCA and RDA might be helpful because they impose significant structural constraints (linearity, orthogonality) that enhance bias while sharply reducing variance. In Section 4, we present a hybrid framework that is driven by this trade-off. To reduce variance, we reduce the effective dimensionality presented to the neural component by projecting the covariate space onto a small number of statistically justified canonical axes. This allows the network to model residual non-linearity, which is not captured by linear ordination.

4. Proposed Hybrid Statistical–AI Framework

We propose a two-stage hybrid estimator that couples the interpretability of constrained ordination with the flexibility of artificial intelligence. In the first stage, RDA or CCA is applied to the community matrix X constrained by the environmental matrix E, producing m canonical axes $F = [F_1, \ldots, F_m]$ that capture the dominant, environmentally structured modes of community variation, together with the associated eigenvalues λ_k.

In the second stage, an artificial neural network g_θ is trained to predict an ecological response of interest y (e.g., species richness, biomass, or a conservation status indicator) from the concatenation of the canonical scores F and the raw environmental covariates E:

$$\hat{y} = g^\theta([F, E]) = W_2\, \varphi(W_1 [F, E] + b_1) + b_2$$

The rationale for including both F and E is that F captures the low-dimensional, variance-maximizing community structure, while E preserves covariate information not fully summarized by the retained canonical axes (i.e., any environmental variation beyond the first m axes). The overall training objective combines predictive loss with an inertia-preservation penalty that discourages the neural component from distorting the ordination structure:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \zeta \cdot \sum_{k=1}^{m}(1 - \mathrm{Corr}(\mathbf{F}_k, \hat{\mathbf{g}}_k)^2)$$

where $\hat{g}_k$ denotes a linear read-out of the k^{th} canonical axis from the network's first hidden layer, and $\zeta \geq 0$ is a regularization weight balancing predictive accuracy against ordination fidelity.

Algorithm

Step 1: center X and E and standardize them as needed.

Step 2: Use RDA or CCA to calculate the eigenvalues λ_k and the restricted ordination axes F. Thirdly, choose m axes such that the ratio of $\Sigma\lambda_k$ to $\Sigma\lambda_j$ is greater than a predetermined variance-explained threshold, for instance 70%. The fourth step is to build the enhanced input matrix Z = [F, E]. Fifth Step: Implement stochastic gradient descent with early halting on a validation partition to train the neural predictor g_θ on Z with the goal of minimizing $L(\theta)$. Step 6: Use a held-out test partition to assess prediction performance and ordination-fidelity diagnostics.

5. A Model of an Ecological Dataset for Simulation Purposes

We built a simulated dataset with 300 sampling sites, 25 species, and 8 environmental variables (soil pH, moisture, elevation, mean annual temperature, precipitation, canopy cover, disturbance index, and distance to water) to test the hybrid framework in a controlled environment. To account for non-linear structure that would be impossible to capture using a linear ordination model alone, species abundances were calculated as Poisson counts with

means following unimodal (Gaussian) response curves along two latent environmental gradients. Additionally, there were additive interaction terms between temperature and moisture. We used a non-linear function of the latent gradients and added Gaussian noise to mimic a continuous response variable, total community biomass. We then used a threshold rule on the disturbance index to provide a categorical response, disturbance class, which can be low, moderate, or high. One model was a hybrid RDA-ANN, while the others were Statistical-only, RF-only, ANN-only, and the proposed RDA-ANN model. The first two models used ordinary least squares regression on the canonical axes, while the second and third used feedforward neural networks trained on the raw covariates. Hyperparameters such as tree depth, hidden units, learning rate, and regularization weight ζ were chosen using five-fold cross-validation on the training partition. All models were then tested on the 30% held-out sites after being trained on the 70% partition.

Table 1. Predictive performance for the biomass response (continuous), test partition (n = 90).

Model	RMSE	MAE	R^2
Statistical-only (RDA + OLS)	4.82	3.61	0.58
RF-only	3.95	2.98	0.69
ANN-only	3.71	2.77	0.72
Hybrid RDA–ANN	3.24	2.41	0.79

Table 2. Classification performance for disturbance class (categorical), test partition (n = 90)

Model	Accuracy	Macro-F1	Cohen's κ
Statistical-only (LDA on RDA axes)	0.71	0.68	0.55
RF-only	0.79	0.77	0.66
ANN-only	0.81	0.79	0.69
Hybrid RDA–ANN	0.87	0.85	0.78

The hybrid model provided the best results for both the continuous and categorical response variables in terms of classification accuracy and error rate. This supports the hypothesis that the ecological signal can be better captured by combining variance-maximizing ordination with flexible non-linear function approximation.

6. Model Validation and Statistical Testing

Beyond simple measures of predicted accuracy, we looked for statistically significant differences in model performance. We used repeated five-fold cross-validation (10 repetitions, 50 total folds) to get RMSE distributions for each model. Then, we used repeated-measures ANOVA to test if the mean RMSE was the same for all four models, with fold identity as a blocking factor to account for shared data partitions.

Table 3. Repeated-measures ANOVA on cross-validated RMSE (biomass response)

Source	df	Sum of Squares	F	p-value
Model	3	18.42	27.65	< 0.001
Fold (block)	49	9.87		
Residual	147	12.03		

Based on post-hoc paired comparisons with Holm correction, it was confirmed that the hybrid RDA-ANN model significantly beat all three baselines (all adjusted $p < 0.01$), and the model effect was statistically significant ($F(3, 147) = 27.65$, $p < 0.001$). In contrast, there was no significant difference between the RF-only and ANN-only models (adjusted $p = 0.34$). The environmental variables also accounted for a considerable portion of the community's variance,

according to a permutation test on the RDA component (999 permutations, p = 0.001). This proves that the limited ordination axes that feed into the hybrid model contain real ecological signal and not just sampling noise.

7. Results and Discussion

Ultimately, the simulation results lend credence to three key findings. To begin with, it is expected that, due to the linear projection at the heart of RDA, purely statistical ordination-based regression will underperform both AI baselines when the underlying species-environment relationships contain unimodal responses and interaction effects that exceed what two or three linear canonical axes can represent. However, this is not always the case, especially when the responses are multimodal. Secondly, ecologists rely on orthogonal, variance-ranked latent axes to understand gradient structure and communicate results using ordination biplots; however, models that are solely based on artificial intelligence (ANN and random forest) only enhance predictive accuracy. Thirdly, surpassing the predictive accuracy of either baseline family alone, the hybrid model maintains the ordination structure—the correlation regularization term ΄Y ensures near-unit correlation between the network's internal read-out units and the canonical axes F in more than 90% of cross-validation folds—while improving accuracy.

When it comes to ecological datasets with moderate sample sizes and high covariate dimensionality, these results are in line with previous applied work that has shown that combining dimensionality reduction with flexible learners improves generalization. This is because non-linear ecological responses can be under-fitted by strictly linear statistical models and overfitted by AI models. The hybrid model converges toward the statistical-only baseline as $\zeta \to \infty$, and it converges toward an unconstrained ANN operating on an extended feature set as $\zeta \to 0$, acting as an adjustable dial between interpretability and flexibility.

Even though the final predictive mapping to biomass or disturbance class is non-linear, the hybrid approach has the practical advantage from an ecological-inference standpoint that ordination biplots, species-environment correlation diagrams, and eigenvalue-based variance partitioning remain directly interpretable. This resolves a typical argument against using AI in ecology, which is that improving prediction capabilities comes at the expense of making scientific findings more understandable. The larger methodological discussion on interpretability in ecological AI should be considered when placing these findings. After fitting, models based only on AI can use post-hoc explanation tools like partial dependence plots, accumulated local effects, and Shapley-value decompositions to estimate the marginal effect of individual covariates. However, these explanations do not ensure that there is a correspondence with any axis that maximizes variance or is ecologically meaningful. In contrast, the ordination model defines the canonical axes used in the hybrid framework a priori. This makes the explanations produced by the model more robust statistically and makes it easier to compare them across studies. This is because canonical axes, when calculated on similar sets of covariates, are more directly comparable than the internal representations of two neural networks that were trained independently.

8. Limitations and Future Directions

These conclusions are qualified by a number of caveats. While this case study makes use of synthetic data with a known generative structure, real ecological datasets present additional challenges such as missing data, unequal sampling effort, taxonomic misclassification, and spatial and temporal autocorrelation. Failing to explicitly model these issues can lead to bias in the ordination and AI parts of the hybrid model. The computational cost is higher compared to either baseline alone due to the introduction of an extra hyperparameter (ζ) that needs to be calibrated using cross-validation, even though the correlation-regularization penalty works well in the simulated environment. Ideally, future research might build on this hybrid framework by adding support for explicitly

geographical models (e.g., with a spatial random-effects term or a Gaussian process prior over site locations) and explicitly temporal settings (e.g., with recurrent or state-space neural component extensions). To further address the issue of conservation decision-making in the face of data scarcity, Bayesian versions of the hybrid estimator that jointly estimate the canonical axes and network weights with posterior uncertainty quantification would also permit the propagation of ordination uncertainty into predictive intervals. The generalizability of the performance enhancements demonstrated here requires empirical validation on numerous real-world long-term ecological monitoring datasets, covering different taxa and biomes. The issue of computing cost and reproducibility is another real constraint. While restricted ordination methods like RDA and CCA are computationally cheap and only need one eigen-decomposition, even for large community matrices, the neural part of the hybrid model usually involves several random restarts to prevent poor local minima, hyperparameter search over learning rate, hidden-layer width, and the regularization weight ζ. Despite being far less expensive than large-scale deep architectures like CNNs applied to raw remote-sensing images, the hybrid framework is significantly more expensive than either of the statistical baselines due to this asymmetry. The statistical component's closed-form eigen-decomposition is unconcerned with the need for precise reporting of random seeds, software versions, and halting criteria in order to guarantee the neural component's repeatability across research groups.

9. Conclusion

In this work, we provide a mathematical framework for ecological data modeling that incorporates AI and multivariate statistics, and we present a hybrid estimator that maximizes variance for a regularized neural predictor using interpretable ordination axes as an input manifold. The hybrid model achieved better results on continuous and categorical ecological responses than baselines based only on statistics or artificial intelligence on a simulated multivariate ecological dataset with realistic unimodal species responses and environmental interaction effects, all while maintaining a high level of agreement with the canonical ordination structure. These findings point to the fact that AI and multivariate statistics are not mutually exclusive but rather work well together when it comes to analyzing ecological data. While AI offers the non-linear flexibility necessary to capture complicated ecological processes, statistical ordination provides an interpretable scaffold. Research in this area could benefit from additional empirical validation on actual ecological monitoring networks and from expansions that are both geographically and temporally expressive.

References

1. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
2. Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783–2792.
3. Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics, 40, 677–697.
4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.
5. Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer.
6. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.
7. Legendre, P., & Legendre, L. (2012). Numerical Ecology (3rd ed.). Elsevier.
8. Olden, J. D., Lawler, J. J., & Poff, N. L. (2008). Machine learning methods without tears: a primer for ecologists. The Quarterly Review of Biology, 83(2), 171–193.
9. Recknagel, F. (2001). Applications of machine learning to ecological modelling. Ecological Modelling, 146(1–3), 303–310.
10. ter Braak, C. J. F. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology, 67(5), 1167–1179.
11. Vapnik, V. N. (1998). Statistical Learning Theory. Wiley.
12. Zuur, A. F., Ieno, E. N., & Smith, G. M. (2007). Analyzing Ecological Data. Springer.