Ecological Model Building

Smit, A. J.
University of the Western Cape

2024-07-06

Table of contents

Developing ecological models, whether multiple regression models or constrained ordinations with numerous environmental predictors, requires a synthesis of biological and ecological theory (the substance of your undergraduate education) with defensible statistical analysis to achieve an integrated, acceptable outcome.

I shall guide you through the process of selecting the most appropriate model to explain ecological outcomes, using the dataset concerning seaweed species composition (β-diversity) along the South African coastline as our exemple. You should be well-acquainted with this analysis by now. The principles we explore are broad applicability across various regions and scales.

1 About the Seaweed Data

In my examples (Gradients Example, Multiple Regression and db-RDA), I employ environmental variables derived from an extensive daily time series of seawater temperature. These data were transformed into various statistics that capture components of the thermal regime along the South African coastline:

- the annual mean climatology (`annMean`)
- the climatological mean for February (warmest month) and August (coldest month) (`febMean` and `augMean`)
- the climatological SD for those months (`febSD` and `augSD`)
- the temperature range (from daily climatology) for February and August (`febRange` and `augRange`)

Each variable was calculated as Euclidean distances between site pairs, thus establishing a foundation for β-diversity assessment. To this end, I apply the Sørensen dissimilarity to the corresponding species data between those site pairs. Given that these summary statistics derive from the same initial dataset, issues of non-independence among predictor data become inevitable. Multicollinearity is virtually guaranteed (even theoretical considerations make this apparent). Nevertheless, these data retain utility for understanding whether seaweed flora composition responds primarily to mean annual temperature, minimum values, or maximum values. The analysis might further reveal which aspects of temperature variability along the coast exert greater influence in structuring species composition.

I have also incorporated geographical distance (recalculated as Euclidean distance) between site pairs along the coast, which thus corresponds to the same spatial grid as the β-diversity and environmental distance data. Here lies a conceptual snare: does distance function as a genuine predictor, or merely serve as a convenient descriptor of the spatial domain across which species gradients unfold? The same question arises for `bio`, the bioregional classification of the seaweed flora established by Professor John Bolton.

2 Theoretical Understanding of Environmental Drivers

Begin by examining how specific environmental variables (*e.g.*, `dist`, `bio`, `augMean`, `febRange`, `febSD`, `augSD`, `annMean`) might influence seaweed community structure along the coast. Temperature is a fundamental driver, affecting biological processes and impacting species differentially. When selecting relevant predictor variables, consider how temperature metrics (such as the mean values, fluctuations, and ranges) influence reproductive timing, growth rates, and physiological tolerances. Develop your thinking from your ecological understanding of these principles. You'll find that some variables may lack theoretical importance and warrant removal on conceptual grounds before modelling commences.

New Concepts

1. Where such data exist, consider incorporating **trait-based approaches** into your analysis. Functional traits of seaweeds (for example, thallus morphology, photosynthetic pigments, or reproductive strategies) may respond differentially to environmental gradients. This approach

can illuminate mechanisms driving community composition beyond simple species presence/absence data. While I have not yet explored this avenue, I anticipate pursuing it as a future research project, probably through **Fourth-corner analysis** or **RLQ ordination**.

Fourth-corner analysis

Fourth-corner analysis addresses questions about which species traits respond to which environmental gradients. The method derives its name from the "fourth corner" of a conceptual data matrix linking three data tables:

- R: **Site-by-species abundance matrix** (the species table with the community data)
- L: **Site-by-environment matrix** (the temperature variables, distance, bioregions, *etc.*, in the environmental table)
- Q: **Species-by-traits matrix** (morphological, physiological, or life-history characteristics)

The "fourth corner" is the missing direct link between traits and environment, which the analysis infers through the observed community patterns. The method tests whether some trait-environment combinations occur more or less frequently than expected by chance, given the observed species distributions. The statistical underpinning relies on permutation tests to assess significance, typically using three different null models:

- Permuting sites (tests for environmental filtering)
- Permuting species (tests for trait convergence)
- Permuting both simultaneously (combines both ecological processes)

RLQ ordination

RLQ ordination extends the fourth-corner concept into multivariate space and provides an integrated view of trait-environment relationships. It does a double ordination that maximises the covariance between environmental variables (table **R** × **L**) and species traits (table **L** × **Q**), with the species composition table (**L**) bringing in the new concept.

Mathematically, RLQ finds linear combinations of environmental variables and linear combinations of traits that produce maximum correlation when projected through the species composition data. It reveals the main gradients along which traits and environmental conditions co-vary. The technique essentially performs a **co-inertia analysis** between two indirect ordinations:

- Environmental variables weighted by species abundances
- Species traits weighted by their occurrence across sites

2. Beyond environmental variables and spatial gradients, incorporating **phylogenetic methods** can yield deeper insights into factors influencing ecological outcomes. Phylogenetic approaches account for evolutionary relationships among species. It might reveal patterns and relationships that complement species composition and trait-based analyses, and can reveal

whether community assembly is driven by environmental filtering (closely related species co-occurring) or competitive exclusion (distantly related species co-occurring). For marine seaweeds, phylogenetic structure can indicate evolutionary constraints on environmental tolerances, niche conservatism vs. adaptive radiation, and historical biogeographic processes.

3 Identifying Spatial Gradients

Assess whether your variables exhibit strong spatial gradients or differences. Consider these examples:

1. The **annual mean temperature (`annMean`)** integrates data from warm and cold seasons to serve as an integrated predictor of global ecosystems. Regionally, it may function as a significant driver due to coastal temperature gradients, though potential collinearity with other variables requires examination.
2. The **mean temperature of the warmest month (`febMean`)** displays a clear gradient from the east coast to Cape Point, but remains relatively stable along the west coast. Variability in this region is captured by `febSD`.
3. The **temperature range of the warmest month (`febRange`)** differentiates the Benguela Current from the Agulhas Current. It exhibits both east-west and north-south gradients.
4. Looking at **temperature variability** such as during the coldest and warmest monsts, `augSD` and `febSD`, respectively, give geographically-linked explanations, maybe showing how to upwelling intensity or current stability affect species structural patterns.
5. Employ **unconstrained ordinations** with environmental vectors (`envfit()`) to guide selection of important structuring predictors.

New Concept

6. Consider incorporating specific **oceanographic features** into your analysis. Ideas that come to mind include upwelling intensity, current velocity, or nutrient availability. These factors can markedly influence seaweed distribution and may enhance the explanatory power of your models. A more advanced approach would be to use **oceanographic models** to derive spatially explicit variables that capture the dynamic nature of coastal environments. For example, Langrangian models will allow one to track the directionality of water movement and its influence on species dispersal and connectivity.

Lagrangian Models

Current patterns, upwelling dynamics, and water mass movements may be more relevant than Euclidean distance for understanding seaweed community patterns. Lagrangian oceanographic variables (tracking water movement) will outperform static Eulerian (as used in the present seaweed analysis) measures for marine species distributions.

4 Assessing Environmental Gradients

Environmental variables with strong spatial gradients likely exert the most significant impact on seaweeds, indicating plausible environmental filtering (niche mechanisms). To quantify these gradients:

1. Conduct **multiple linear regressions** using continuous predictor variables as functions of dist (distance between site pairs).
2. Create **thematic maps** (spatially implicit) of temperature variables and vary symbol size or colour intensity by magnitude. Maybe use GIS tools to interpolate values between sampling points for more comprehensive visualisation.
3. Assess **spatial autocorrelation** in your variables using techniques such as Moran's I or Geary's C. This helps identify the scale at which environmental factors operate and informs model structure.

5 Model Building and Variable Selection

Throughout the model-building process, make informed decisions about variable selection based on both theoretical knowledge and data-driven approaches:

1. Select variables based on **ecological understanding** of species' responses to environmental drivers, considering both direct and indirect effects.
2. Choose variables that reflect **significant known environmental gradients** influenced by factors such as ocean currents, coastal topography, and climate patterns.
3. Use **unconstrained ordinations** (*e.g.*, Principal Component Analysis, PCA) to explore the relationships between environmental variables and species composition. This can help identify key gradients and inform variable selection. Running such ordinations side-by-side on the species and environmental data can reveal how well the environmental variables explain the observed species patterns. This method becomes especially useful when you superimpose the environmental vectors onto the ordination plot using `envfit()` or a GAM or something similar.
4. Also use **data-driven decision making**. Employ statistical methods such as Variance Inflation Factors (VIFs) or forward selection (*e.g.*, `stepAIC()`) to address multicollinearity and refine model selection. Consider modern techniques like elastic net regression or random forests for variable importance ranking.

New Concepts

5. Consider also spatial autocorrelation in the models. Beyond simply identifying spatial autocorrelation, one might develop explicit strategies for incorporating it into models. One such is **Moran's Eigenvector Maps (MEMs)**, which can be used to account for spatial structure in species-environment relationships. MEMs are derived from distance matrices and can be included as predictors in regression models to capture spatial patterns.

Moran’s Eigenvector Maps (MEMs)

Moran’s Eigenvector Maps (MEMs) can decompose spatial patterns at multiple scales. It allows one to partition variation into broad-scale environmental trends and fine-scale spatial processes. I have already done this for the seaweed analysis and you can read about it in the paper. For coastal seaweed data, it can reveal spatially explicit account for:

- Oceanographic connectivity between sites
- Distance-decay relationships in species similarity
- Scale-dependent environmental effects (local vs. regional processes)

6. Explore **model averaging** techniques, such as Akaike weights or Bayesian Model Averaging, to account for model uncertainty. Model averaging yields more believable predictions and insights into the relative importance of different environmental variables.

6 Reconciling Ecological and Statistical Knowledge

Achieving a model with optimal explanatory power requires you to reconcile your ecological knowledge with the statistical techniques at your disposal:

1. **Synthesise** theoretical insights on environmental gradients and biological responses with statistical techniques to build defensible models.

2. Develop and **test hypotheses** using multiple regression models and consider both ecological relevance and statistical fit. Remain open to unexpected results that may challenge existing ideas.

3. Extend analysis to **multivariate methods** (for example constrained ordinations for gradient detection and attribution, and clustering for group identification) to uncover more complex ecological patterns.

New Concepts

4. Consider modern techniques like **Joint Species Distribution Models (JSDMs)** to simultaneously model multiple species and environmental factors.

Joint Species Distribution Models (JSDMs)

Joint Species Distribution Models would be a great addition to the seaweed community-level analysis. JSDMs simultaneously fit distributions for multiple species and account for residual correlations among species after environmental effects are removed. For the seaweed dataset, JSDMs should be able to offer the following:

- Separate environmental filtering from potential biotic interactions
- Improve predictions for rare species by borrowing strength from common species
- Identify species associations that persist across environmental gradients
- Provide more defensible estimates of environmental effects by accounting for community-level patterns

The **Hmsc** R package provides hierarchical JSDMs that can handle various response types and spatial/temporal structures.

5. Begin identifying the most influential species using approaches such as **multivariate abundance analysis with Generalised Linear Models (GLMs)**, which offers a different approach to “Model-based Multivariate Analyses”.

7 Model Validation

7.1 Cross-Validation

The purpose of model building is to develop models that can generalise well to new data (*i.e.*, they must be able to **predict**), and to do so, we need to carefully validate our model’s performance.

Traditional random cross-validation often inadequately represents the predictive challenges in ecological datasets. **Spatial blocking cross-validation** could be used when dealing with autocorrelated data. This involves partitioning data into spatially contiguous blocks rather than random subsets to better mimic real-world prediction scenarios where models must extrapolate to new locations or time periods.

For the seaweed data, something I might try is to implement **leave-one-site-out cross-validation** or **geographic blocking** that respects the spatial structure of the South African coastline. This should provide realistic estimates of model performance when predicting to unsampled coastal areas and help identify models that really capture generalisable ecological patterns versus those that exploit spatial autocorrelation.

7.2 Model Averaging and Ensemble Methods

Model averaging techniques takes our modelling beyond simply selecting single “best” models. One may use **Akaike weight-based averaging** or **Bayesian Model Averaging** to account for model uncertainty. These methods combine insights from multiple (different) models if they offer similar support from the data; averaging predictions across these models often provides more defensible forecasts than relying on any single model.

For seaweed community analysis, ensemble approaches combining multiple statistical techniques (GLMs, GAMs, machine learning methods) can capture different aspects of species-environment relationships. **Boosted Regression Trees (BRTs)** are particularly effective for ecological modelling as they handle non-linear relationships, variable interactions, and mixed data types while providing interpretable variable importance measures.

8 Practical Steps for Model Selection

1. Use ecological knowledge to select relevant environmental predictors, and consider both direct and indirect effects on seaweed physiology and ecology.
2. Evaluate the strength and pattern of spatial gradients using regression, mapping techniques, and spatial statistics.
3. Refine your models by address multicollinearity and other data issues using appropriate statistical methods. Consider interaction terms and non-linear relationships where ecologically justified.
4. Validate models using both ecological and statistical criteria to ensure parsimony and ecological meaning. Employ techniques like cross-validation or bootstrapping to assess model robustness.
5. Present findings accessibly to both ecologists and statisticians. This requires that you emphasise the biological significance of your models alongside their statistical performance.

9 Some Advanced Ideas

(To be developed)

9.1 Hierarchical and Multi-Scale Modeling

Hierarchical Bayesian models may be used with ecological data with multiple sources of variation. For seaweed communities, implement hierarchical structures that account for:

- Site-level random effects (unmeasured local conditions)
- Region-level variation (biogeographic differences)
- Species-level random effects (unmeasured species traits)
- Temporal variation (if data spans multiple years)

It allows borrowing strength across hierarchical levels while maintaining appropriate uncertainty estimates.

9.2 Multi-Scale Environmental Predictors

One can build scale-explicit models that incorporate environmental predictors measured at multiple spatial and temporal scales. For coastal systems, this might include:

- Fine-scale habitat characteristics (substrate, local topography)
- Intermediate-scale oceanographic features (upwelling cells, current boundaries)
- Large-scale climate patterns (ENSO, Indian Ocean Dipole)

9.3 Advanced Methods and Approaches

Null Models and Randomisation Tests

Null models provide baselines for interpreting ecological patterns. Implement randomisation tests to determine whether observed species associations, environmental relationships, or spatial patterns exceed those expected by chance alone. This is useful when dealing with sparse ecological datasets where spurious patterns can easily emerge.

Variable Transformation and Standardisation

Develop structures and principled methods around data transformation that maintain ecological interpretability. For species composition data, consider:

- Hellinger transformation for abundance data to reduce the influence of dominant species
- Chord transformation for presence-absence data
- Species-specific standardisation that accounts for different detection probabilities

Model Interpretation and Communication

Uncertainty Quantification

Stop relying on point estimates but favour comprehensive uncertainty assessment. For example, implement:

- Bootstrap confidence intervals for parameter estimates
- Prediction intervals for spatial forecasts
- Model selection uncertainty through information criteria weights
- Structural uncertainty through ensemble approaches

Residual Analysis for Model Diagnostics

Develop full residual analysis workflows specific to ecological data. For multivariate community data, examine:

- Spatial autocorrelation in residuals using Moran's I
- Temporal autocorrelation for time series data
- Heteroscedasticity across environmental gradients
- Outlier detection and influence analysis

Software

There are many R packages that can do almost any conceivable analysis> At the very least, know what they are and what they can do.

Reproducible Workflows

As always, emphasise reproducible research practices, especially as one starts to build on more advanced statistical methods involving larger teams of people. This includes:

- Version control such as GitHub for code and data
- Documented model selection procedures
- Apply standardised validation protocols

- Clear reporting of model assumptions and limitations

Bibliography