Elevational Metacommunity Analysis of Baja Ranch Arthropods

Eric Stiner

This elevational metacommunity analysis integrates biodiversity, spatial, and environmental data to test how arthropod communities shift along elevation and landscape gradients. The workflow merges site-level community data with metadata and environmental layers, normalizing abundances, and linking replicates to ranch identity. Alpha diversity metrics (richness, Shannon, Simpson) are modeled against elevation, while beta diversity partitioning and NMDS ordinations evaluate turnover and nestedness patterns across the gradient. Resistance surfaces and distance-based GDM models quantify how environment, geography, and topographic resistance explain community dissimilarity. Finally, indicator analyses reveal which taxa define each ranch and which sites are compositionally most unique, mirroring the hierarchical multiscale approach of *Noguerales et al. (2023)* but adapted for a smaller metacommunity system.

Noguerales, V., Meramveliotakis, E., Castro-Insua, A., Andujar, C., Arribas, P., Creedy, T., Overcast, I., Morlon, H., Emerson, B., Vogler, A., & Papadopoulou, A. (2021). Community metabarcoding reveals the relative role of environmental filtering and dispersal in metacommunity dynamics of soil microarthropods across a mosaic of montane forests. Wiley. https://doi.org/10.22541/au.162501879.93982613/v1

[notes: packages run in .venv environment locally from Eric Stiner with Genus level taxonomy]

FILE PREP:

INPUT FILE SCRIPT: combine_ALL_Ranchsites_HL_taxonomy_weighted_v2_4.py - Creates input Files.

---

1. Data Assembly

Purpose: link communities to metadata, clean, and check.

Script: - Step1_Build_Community_Metadata_Tables_Pseudoreps_GENUS.py

Actions:

- Join counts with sample metadata.
- Remove negative controls.
- Preserve all replicate Malaise samples per ranch (don't collapse yet).
- Extract elevation from DEM if not included in metadata.

- Create ranch_sites.csv with 4 rows = ranch centroids (lat, lon, elevation).

---

2. Normalization & Transformation

Script: **Step2_Normalize_Transform_FIGS.py** prepare matrices for different diversity analyses.

Actions:

- Hellinger-transform counts → for Bray–Curtis dissimilarity.
- Presence/absence matrix → for Jaccard, Sørensen, Simpson turnover.
- Retain raw counts for richness (α-diversity).

2.5 Find elevation gain between each ranch. This takes an onX Backountry map exported to KML then finds elevation points along the track in Google Earth Pro and export KML with elevation points.

Script: **ranch_segments_canonical_v2_WORKS_Metric.py** - This script automates the segmentation of elevation transects across the Baja ranch landscapes by processing KML or GPX tracks exported from Google Earth Pro or onX Backcountry. It samples evenly spaced points along each route and extracts their corresponding elevation values, generating a continuous profile of the terrain. From these points, it calculates key geomorphometric metrics such as slope, cumulative elevation change, and distance along the transect. Outputs include a CSV file containing geographic coordinates and terrain metrics, and optionally, a KML or GeoJSON file for visual inspection of elevation segments in Google Earth. This step effectively establishes a canonical elevation framework for comparing biodiversity patterns among ranches and along environmental gradients.

Script: **Plot_ranch_elevation_profile.py** -This script visualizes the elevation transects produced by the segmentation workflow, plotting elevation against cumulative distance to generate a clear elevation profile for each ranch route. Using the CSV output from the segmentation script, it constructs a smooth line plot of elevation change and can overlay site boundaries, sample points, or ecological zones along the transect. The resulting figure illustrates the topographic structure of each ranch and highlights major elevational transitions relevant to arthropod community turnover. By providing a visual reference for elevation gradients, the script aids in interpreting biodiversity-environment relationships and in verifying the accuracy of the segmentation process. The final output is a PNG or PDF elevation profile that can be integrated into summary figures or modeling visualizations.

3a. Alpha Diversity

Script: **Step3a_Alpha_Diversity_Average_Merge.py** - Alpha Diversity (with replicate collapse and visible spline fits).

This step merges replicate Malaise trap samples and computes alpha diversity metrics (Richness, Shannon, Simpson) at either the trap or ranch level. Replicates are collapsed into a single representative value per ranch using a selectable mean (arithmetic, weighted, median, geometric, or harmonic), ensuring accurate site-level estimates of community composition. The script outputs alpha_diversity.csv, a ranch-level summary, linear regression statistics versus elevation, and publication-ready figures. Each plot includes larger, outlined points with bold ranch labels and displays both a solid linear fit and a dashed orange spline fit. For small datasets (e.g., four ranches), the spline defaults to a quadratic (k = 2, s = 0) for interpretability while retaining the linear fit as the primary statistical model.

3b. Alpha Diversity vs Elevation

Why: checks whether α diversity declines upslope or changes predictably with elevation.

Script: **Step3b_Alpha_vs_Elevation_v13b.py -** test environmental filtering on local diversity.

Actions:

- Calculate richness, Shannon, Simpson per sample.
- Models: richness ~ elevation (linear and GAM).
- Plot diversity vs elevation with ranch labels.

---

4. Beta Diversity Partitioning

Why: whether turnover explains most elevational differentiation.

Script: **Step4_beta_partition_v6.py-** script integrates replicate-level alpha diversity metrics with site-level environmental context to evaluate diversity–environment relationships along the ranch transect. It first builds a unified analysis table by merging replicate alpha diversity values (richness, Shannon, Simpson) with site metadata (ranch identity, geographic coordinates) and a digital elevation model (DEM) to extract standardized elevations. Cumulative geographic distance and elevation gain are then calculated across the ordered ranch transect. The script produces replicate-level and ranch-mean scatterplots and regression models of diversity versus

elevation, distance, and cumulative gain. For each metric × predictor, it outputs both visual summaries (scatterplots with fitted regression lines and annotated slopes, R^2, p-values, and sample sizes) and tabular summaries (CSV files of model statistics and the data used for plotting). Together, these outputs provide publication-ready figures and quantitative results for testing how alpha diversity changes across environmental gradients and geographic structure.

---

5. Ordination and Environmental Fitting

Why: Visualize and test effects of elevation/climate. Ordination shows gradients; PERMANOVA quantifies whether elevation drives compositional differences beyond chance.

OPTIONAL PREP_Script *-* ***Step5_Preflight_FixIDs.R*** *- if files are not oriented correctly auto-reorients matrices to* ***rows = samples****.*

Script: Step5_Ordination_Envfit_v12_orient.R - This script performs ordination and environmental fitting to visualize and quantify patterns of community turnover among ranch sites. It takes the normalized community matrices (Hellinger and presence–absence) and aligned metadata as input, computes a two-dimensional Non-metric Multidimensional Scaling (NMDS) ordination using Bray-Curtis dissimilarities, and then fits elevation as a continuous environmental variable using envfit to test its correlation with community composition. The ordination is rotated so that the elevation vector aligns intuitively from right (lowland sites) to left (upland sites), without altering the underlying data. Ranch-level centroids are calculated to represent the mean community composition for each site, and segments link individual samples to their corresponding centroid to illustrate within-site variation. A PERMDISP test quantifies differences in within-group dispersion, providing a measure of compositional heterogeneity among ranches. The script outputs NMDS coordinates, centroids, stress values, environmental fit statistics, and a high-resolution, thesis-ready ordination plot showing both the community structure and the fitted elevation gradient.

Actions:

- NMDS on Bray–Curtis and Jaccard.
- Fit vectors: elevation, temperature, aridity, NDVI/EVI.
- PERMANOVA: community ~ elevation (with ranch as grouping).
- PERMDISP: test if dispersion differs across ranches.

---

6. Resistance Surface Construction (IBR)

Why: translates elevation and environment into cost surfaces. Model effective distances beyond Euclidean geography.

Script: **Step6_Construct_Resistance_v3.py** - This takes the raw environmental layers (DEM, slope, roughness, and climate rasters from MODIS and ERA5) and translates them into *resistance surfaces*. The script first reprojects the DEM into a local UTM grid, then derives slope (steeper = more costly) and roughness (rugged = more costly). Each climate raster is scaled two ways: (1) **global scaling**, where high and low values across the whole study area are rescaled to 0–1, and (2) **site-referenced mismatch**, where each pixel is compared to the mean and variability of conditions at the ranch sites to measure how different it is from those conditions (more mismatch = higher resistance). Individual resistance layers are saved for each variable, and two combined layers are built: a **global baseline resistance map** that blends slope, roughness, and climate layers with user-defined weights, and a **climate-only mismatch map** that blends site-referenced climate layers. Both raster sets are written to GeoTIFF with optional quicklook PNGs, and a summary file documents the layers created and the weights used.

Actions:

- From DEM, create resistance rasters:
    1. Slope = steeper = more resistance.
    2. Roughness = more rugged = more resistance.
    3. MODIS and ERA5 = resistance increases with environmental differences..
    4. Combined = average/weighted mean of above.
- Align rasters to same CRS/resolution.

---

7. Effective Distance Matrices

Why: these become predictors to test against community dissimilarities to calculate pairwise distances among site centroids.

Script: **Step7_Build_Distance_Matrices_v1.py - Build core predictor distances for Step 8–9 -** This script generates the three “standard” distance matrices used downstream plus optional least-cost (resistance) distances. It (1) reads your **Hellinger-transformed community matrix** and **samples.csv** (with sample_id, ranch, lon, lat), averages replicate samples to ranch centroids, and writes **D_bray.csv** (Bray–Curtis among ranch means); (2) computes geodesic great-circle distances between ranch centroids to **D_geo_km.csv**; (3) loads **env_table_sites.csv**, z-scores numeric environmental variables, and writes Euclidean

distances in that space to **D_env_scaled.csv**; and (4) if one or more resistance rasters are supplied, computes **least-cost path distances** for each and saves them as **D_cost_<rastername>.csv** (falls back gracefully if least-cost routing can't run). Outputs are square CSV matrices with matching row/column order (ranches), plus an INDEX.txt summary. Use these files directly in Step 8 MRM comparisons and Step 9 variance partitioning.

Actions:

- Compute:
    - D_geo = straight-line geographic distances.
    - D_env = Euclidean distance in scaled environment space (elevation, temp, aridity).
    - D_resX = cost distances from each resistance raster.

---

8. Distance Model Comparisons

Why: Key to show environmental filtering/topographic resistance matter more than geography alone. test whether resistance/environment beat straight-line distance.

Script: Step8_MRM_Distance_Models_v3.R - This script performs multiple regression on distance matrices (MRM) to evaluate how well spatial and environmental predictors explain variation in community dissimilarity among ranches. It uses pairwise matrices of community dissimilarity (e.g., Bray–Curtis or Jaccard) as response variables and compares them against predictor distances representing geography (D_geo), environment (D_env), and resistance (D_res). The script fits permutation-based linear models using ecodist::MRM, testing both single-predictor and multi-predictor combinations to quantify their individual and combined explanatory power. Results include R^2, adjusted R^2, and permutation-based p-values for each model, along with visualizations that show which predictors most strongly shape metacommunity turnover across sites. Outputs include CSV tables summarizing R^2 values and a figure comparing single and combined predictor performance across models.

Script: Step8_Report_MRM_v1.R - This script compiles, summarizes, and visualizes the results of the MRM analyses to produce a clean, publication-ready summary of model performance. It aggregates R^2 and p-values from all MRM models, calculates ΔR^2 (the gain in explained variance) relative to geographic distance alone, and formats the results for inclusion in thesis figures and tables. The script generates clear barplots and summary tables that highlight how environmental and resistance distances improve explanatory power beyond pure geographic distance. The resulting figure complements the primary MRM output, providing a concise overview of the relative contributions of environment, resistance, and geography to community dissimilarity patterns across the four-ranch transect.

Actions:

- MRM (permutation regression):
    - D_comm ~ D_geo
    - D_comm ~ D_resX
    - D_comm ~ D_env
    - D_comm ~ D_geo + D_resX + D_env
- Run for Jaccard and Bray–Curtis.

---

9. Variance Partitioning

Purpose: quantify unique vs shared contributions of resistance attributes

Script: Step9_Pairwise_VarPart_SAFE.R - This script performs the variance partitioning analysis for the four-ranch Baja metacommunity dataset using robust distance-based redundancy analysis (dbRDA) methods. It quantifies the relative and shared contributions of environmental (E), geographic (G), and resistance (R/CR) distance matrices to explaining community dissimilarity. The script first performs PCoA reduction for each distance block (with user-defined thresholds for variance retention), then calculates adjusted R^2 values for single-block and two-block combinations (E + G, E + R, G + R) using vegan::varpart(). It handles degenerate or incomplete matrices (e.g., NAs in resistance surfaces) gracefully, applying Cailliez correction and optional NA filling to maintain model comparability. Outputs include single-block adjusted R^2 tables, pairwise variance-partition results, a consensus summary table, and a metadata log for reproducibility.

Script: Step9_Figures_pub.py - This Python script generates publication-ready visualizations summarizing the results from the Step 9 variance partitioning analysis. It reads the CSV outputs produced by Step9_Pairwise_VarPart_SAFE.R (single-block R^2, pairwise varpart, and consensus tables) and creates a clean, formatted figure that aligns visually with the Step 8 MRM comparisons. The figure displays total adjusted R^2 values for each two-way combination (E + G, E + R, G + R) along with stacked or labeled contributions from each component, facilitating direct interpretation of unique versus shared variance explained. The resulting plot is designed for integration into manuscripts or reports, maintaining consistent color palettes, typography, and labeling conventions established in earlier pipeline figure.

---

10. Indicator Taxa & LCBD

Why: link patterns to taxa and sites and gives ecological interpretation, not just statistical partitioning.

Script: Step10_Indicator_LCBD.R — This script identifies key taxa and sites driving community structure within the Nogales metacommunity. It takes a community abundance matrix (samples × taxa) and corresponding metadata (with ranch or site information) to calculate both indicator species values (IndVal) and beta-diversity contributions. Using the *indicspecies* package, it detects taxa significantly associated with specific groups (e.g., ranches or elevations), reporting their indicator value statistics, specificity, fidelity, and significance. It also computes Local Contributions to Beta Diversity (LCBD) for each site and Species Contributions to Beta Diversity (SCBD) for each taxon using the *adespatial* package. Outputs include full and top-K indicator tables, LCBD and SCBD summaries, and a run metadata log. Together, these outputs reveal which taxa best characterize each ranch and which sites are compositionally most unique within the metacommunity.

Script: step10_two_clean_figs.py — This script produces two clean, publication-ready figures summarizing the results of Step 10. The first figure (*Step10_LCBD.png/pdf*) shows the **mean Local Contributions to Beta Diversity (LCBD)** for each ranch, allowing quick comparison of which sites contribute most to overall community turnover. The second figure (*Step10_Heatmap.png/pdf*) visualizes the **top indicator taxa (IndVal)**, displaying their strength of association to each ranch as a heatmap, with dots marking statistically significant taxa ($p \leq \alpha$, with an automatic fallback if few pass the threshold). Ranches are color-coded consistently across figures, and the layout is designed for clear readability and easy inclusion in manuscripts. The script automatically detects column names (e.g., sample_id, ranch, lat/lon, s.<ranch>), scales colorbars, and saves outputs in both PNG and PDF formats. Together, these two plots provide an ecological summary of which taxa best define each ranch and which sites are compositionally most unique.

Script: Step10_Genus_Top20_DotTable_pub_fixed.py - generates a concise, publication-style “bubble table” that visualizes the relative strength of indicator genera across ranch sites. The script summarizes community data (e.g., the genus-level Hellinger matrix) and calculates the top 20 most abundant or indicator genera for each ranch. Each genus is plotted as a single row, with circular markers positioned under the corresponding ranch column; marker size represents either relative abundance (% of assemblage) or absolute mean abundance per site. This format provides an intuitive way to compare which genera dominate each ranch’s community and how their relative contributions differ across the landscape. The output includes both a high-resolution PNG/PDF figure and a tidy CSV file containing the plotted data for reproducibility and supplementary tables.

Script: Step10_fetch_genus_barcodes_from_ncbi.R - pulls short barcodes from NCBI by genus using only CRAN packages **[not used but cool]**

Script: Step10_SCBD_Figures.R - This step summarizes and visualizes species contributions to beta diversity (SCBD) from the metacommunity analysis outputs. Starting with the SCBD_HIGHER_TAXA.csv file (which contains class, order, family, genus, species, and SCBD values), the script Step10_SCBD_Figures.R aggregates contributions by taxonomic rank and produces two publication-ready figures. Figure 1 displays *total SCBD by family*, while Figure 2 highlights *the top genera within the most influential families*, with each genus label dynamically repelled to prevent overlap. The current version incorporates improved label geometry—moving long genus names up and to the viewer's left with connecting leader lines—so dense Diptera and Hymenoptera panels remain legible even for large datasets.

Actions:

- IndVal to find taxa strongly associated with each ranch/elevation.
- LCBD/SCBD to see which ranch contributes most to turnover and which taxa drive it.

---

11. Sensitivity & Robustness

Purpose: confirm stability with replication and resistance models.

Script: Step11_extract_env_from_geotiffs.R - This script extracts environmental values (e.g., elevation, slope, vegetation indices, and temperature) from a set of aligned GeoTIFF raster layers for each sample in the community dataset. Using sample coordinates, it queries every raster and compiles the values into a metadata table that links biodiversity samples to their local environmental context. This provides the foundation for downstream ecological modeling by ensuring that each sample has complete and spatially consistent predictor data.

Script: Step11a_Prepare_HiRes_Rasters_and_Metadata.R - This script refines environmental data to a common high-resolution spatial grid (e.g., 30 m), aligning DEM-derived topographic predictors (elevation, slope, roughness, and local relief) with optional vegetation and temperature rasters. It then extracts high-resolution environmental values for each biodiversity sample, producing an updated samples_env_HIRES.csv. The output provides spatially harmonized environmental predictors suitable for fine-scale modeling of community turnover and resistance. This high-resolution dataset forms the analytical input for the Generalized Dissimilarity Model (GDM) stage.

Script: Step11b_Aggregate_HIRES_to_Ranch.R - Joins samples_env_HIRES.csv with the ranch labels in samples.csv and collapses trap-level predictors to a single row per ranch using

unweighted means. Outputs ranch_env_HIRES.csv, the ranch-level metadata used for GDM.

Script: Step11_GLM_Ranch_AllModels.R - performs ranch-level exploratory ecological modeling by integrating community composition, alpha diversity, and high-resolution environmental predictors. It takes the ranch-aggregated taxa matrix and the ranch-scale environmental summary file, aligns them by site, and calculates richness, Shannon diversity, and evenness. It then runs a Bray–Curtis NMDS to summarize multivariate community structure and fits environmental vectors (envfit) to identify which predictors align with major compositional gradients. The script then merges NMDS axes, alpha metrics, and environmental variables into a single data table and fits a series of simple GLMs (elevation or EVI area predicting richness, Shannon, and NMDS1) to test whether major environmental gradients explain variation in community structure. If available, it also runs a ManyGLM (mvabund) to model multivariate abundance across environmental gradients. Finally, it performs a Mantel test comparing Bray–Curtis dissimilarity to multivariate environmental distance.

Script: Step11_GDM_Master.R - This script performs Generalized Dissimilarity Modeling (GDM) to quantify how ecological dissimilarity between communities changes along environmental and geographic gradients. It fits a multivariate nonlinear model using the high-resolution predictors, evaluates deviance explained, and extracts I-spline functions representing the contribution of each predictor to beta diversity. The script includes a bootstrap procedure (typically 500–1000 replicates) to estimate confidence intervals and produce robust importance metrics. Outputs include model diagnostics, bootstrap summaries, and publication-ready figures of I-spline response curves and predictor importance—providing direct evidence of which landscape and environmental variables structure community turnover among ranch sites.

Actions:

- **Bootstrap resampling of site replicates (B = 1000)**
- **Generalized Dissimilarity Modeling (GDM) with nonlinear spline fitting**
- **Comparative β-diversity analysis using alternative community transformations**