TAREA CONJUNTA - SPIDER

María Pallares Diez, Martin Gramage Casado y Pedro Martí Requena

2025-06-09

Descripción detallada del conjunto de datos `Spider`

El conjunto de datos `Spider.RData` contiene información sobre la **abundancia de 12 especies de arañas** en **28 ubicaciones diferentes**, junto con variables morfológicas propias de cada especie y variables ambientales asociadas a cada sitio. El objetivo del estudio es **analizar qué características morfológicas de las arañas se relacionan con su capacidad de adaptación y proliferación**, en función del entorno en el que habitan.

La variable **respuesta** del estudio es `Abundance`, que representa el número de individuos observados por especie y sitio. Se trata de una variable **cuantitativa discreta**, ya que corresponde a un conteo entero, no negativo y sin cota superior.

Las variables **explicativas** se dividen en tres grupos:

- **Variables morfológicas** (asociadas a cada especie):
  - `length`: longitud corporal, tratada como variable cuantitativa continua (transformada logarítmicamente).
  - `colour`: color predominante de la especie (categórica nominal), con niveles `dark` y `yellow`.
  - `marks`: tipo de marcas corporales (categórica nominal), con niveles `none`, `spots` y `stripes`.
- **Variables ambientales** (asociadas a cada sitio):
  - `soil.dry`: cantidad de materia seca del suelo (cuantitativa continua).
  - `fallen.leaves`: proporción del suelo cubierta por hojas caídas (cuantitativa continua).
  - `moss`: cobertura de musgo, transformada en logaritmo (cuantitativa continua).
  - `herb.layer`: densidad de vegetación herbácea, también log-transformada (cuantitativa continua).
- **Variables estructurales**:
  - `Site`: identificador del sitio de muestreo (categórica nominal, 28 niveles).
  - `Specie`: especie de araña (categórica nominal, 12 niveles).

Descriptiva numérica y distribución

La variable `Abundance` muestra una media de aproximadamente 10, una mediana de 1, y un máximo de 135. Está fuertemente sesgada a la derecha, con una mayoría de valores bajos (incluidos muchos ceros) y pocos valores extremos altos. Este tipo de distribución es común en estudios ecológicos, donde muchas especies son raras o tienen baja densidad local.

Las variables ambientales presentan rangos amplios. `soil.dry` y `herb.layer` tienden a valores medios-altos, mientras que `fallen.leaves` y `moss` muestran una distribución más asimétrica, con muchas observaciones cercanas a cero. La longitud corporal (`length`) oscila entre 1.5 y 2.4 en escala logarítmica, con valores faltantes en una única especie.

Datos faltantes y su imputación

Se identifican **28 valores perdidos** en las variables `length`, `colour` y `marks`. Todos corresponden a la especie **Auloalbi**, que no tenía información morfológica completa en el conjunto original. Para no excluir esta especie del análisis, se imputaron los valores **basándose en fuentes bibliográficas**. Se asignó un valor de longitud correspondiente a `log(4.1)`, y se codificó su morfología como `colour = dark` y `marks = none`.

Esta imputación permite mantener la coherencia del conjunto de datos sin introducir sesgo sistemático, ya que se apoya en información descriptiva fiable sobre la especie. Además, cabe destacar que **las variables morfológicas (`length`, `colour`, `marks`) son constantes dentro de cada especie**: es decir, **no se mide individualmente a cada araña**, sino que **todas las réplicas asociadas a una misma especie comparten el mismo valor morfológico**. Por tanto, imputar estos valores para Auloalbi en base a bibliografía es coherente con la estructura del resto del conjunto de datos y representa adecuadamente su información específica.

Creación de la variable `appearance`

Para representar de forma más realista los rasgos externos de las arañas, se creó una nueva variable categórica denominada `appearance`, que combina las variables `colour` y `marks`. Esto se debe a que algunas combinaciones no existen en los datos: por ejemplo, no hay especies `dark` con `spots`, ni especies `yellow` sin marcas. La variable `appearance` sintetiza solo las combinaciones realmente observadas, evitando que los modelos posteriores intenten estimar efectos inexistentes. Las cuatro combinaciones presentes son: `none_dark`, `spots_yellow`, `stripes_dark` y `stripes_yellow`.Esta recodificación mejora la interpretabilidad biológica y reduce la complejidad innecesaria en los modelos.

Análisis de componentes principales (PCA) de las variables ambientales

Dado que las variables ambientales estan correlacionadas se aplicó un análisis de componentes principales (PCA) sobre `soil.dry`, `fallen.leaves`, `moss` y `herb.layer`, el cual fue escalado debido a la diferencia en escalas de medida. El objetivo es reducir la dimensionalidad y evitar problemas de colinealidad en el modelado.

Las dos primeras componentes explican el 89% de la varianza total:

- **PC1** (58%): representa un gradiente de sequedad y baja cobertura vegetal. Toma valores altos en suelos secos con hojas caídas, y valores bajos en ambientes más húmedos y con vegetación musgosa.
- **PC2** (31%): refleja un gradiente de densidad herbácea, con valores altos en sitios con abundante vegetación baja.

Estas dos variables, PC1 y PC2, se añadieron al conjunto de datos como predictores ambientales sintetizados. Su uso en los modelos posteriores permite representar la estructura ambiental sin redundancias, facilitando la interpretación de las relaciones de las características morfológicas con el ambiente.

Formulación del modelo

La variable respuesta del estudio es `Abundance`, que representa el número de individuos de arañas observados por especie en cada sitio de muestreo. Se trata de una variable de **conteo** (cuantitativa discreta, no negativa), con muchos ceros y una fuerte asimetría a la derecha. Dada su naturaleza discreta, la distribución de probabilidad más adecuada para modelar `Abundance` es inicialmente la **Poisson**, que es el modelo estándar para datos de conteo. Esta distribución asume que la media y la varianza son iguales.

Se propone utilizar un **modelo lineal generalizado mixto** (GLMM), que permite:

- Incorporar efectos fijos como las variables `appearance`, `length`, `PC1` y `PC2`, asociadas a características morfológicas y ambientales.
- Añadir efectos aleatorios para tener en cuenta la **estructura jerárquica** de los datos: las observaciones están agrupadas tanto por `Site` como por `Specie`.

Tanto `Site` como `Specie` se modelan como efectos aleatorios para capturar la variabilidad no explicada entre sitios y entre especies, permitiendo generalizar los resultados más allá de las unidades observadas.

El parámetro de interés es la media esperada de abundancia, denotada por μ_i, para cada combinación especie–sitio. Este valor representa el número medio esperado de arañas como función de las covariables ecológicas y morfológicas consideradas.

El modelo general se formula como:

$$\text{Abundance}_i \sim \text{Poisson}(\mu_i)$$

$$\log(\mu_i) = \beta_0 + \beta_1 \cdot \text{appearance}_i + \beta_2 \cdot \text{length}_i + \beta_3 \cdot \text{PC1}_i + \beta_4 \cdot \text{PC2}_i + b_{\text{Site}[i]} + b_{\text{Specie}[i]}$$

Donde:

- $g(\cdot) = \log(\cdot)$ es la función de **enlace logarítmica**, adecuada para asegurar que $\mu_i > 0$.
- `appearance` es una variable categórica que codifica el tipo de marcas y coloración morfológica.
- `length` es una variable continua que representa la longitud corporal de la araña.
- `PC1` y `PC2` son las dos primeras componentes principales derivadas del análisis PCA ambiental.
- $b_{\text{Site}[i]} \sim \mathcal{N}(0, \sigma^2_{\text{Site}})$ es el efecto aleatorio del sitio de muestreo.
- $b_{\text{Specie}[i]} \sim \mathcal{N}(0, \sigma^2_{\text{Specie}})$ es el efecto aleatorio específico de cada especie.

Este modelo permite predecir la abundancia media de arañas en función de factores ambientales y morfológicos, controlando por la variabilidad espacial (`Site`) e interespecífica (`Specie`).

Modelos GLM

Modelo Poisson

El modelo (anteriormente presentado) Poisson fue el punto de partida. Aunque es adecuado para datos de conteo, los residuos simulados mostraron claramente la presencia de sobredispersión. El test de sobredispersión de DHARMa también fue significativo, indicando un mal ajuste de este modelo.

Modelo binomial negativo (nbinom2)

Se ajustó un modelo con distribución binomial negativa tipo 2, el cual introduce un parámetro que controla la sobredispersión, que permite que la varianza crezca cuadráticamente con la media. Aunque mejoró el ajuste respecto al Poisson, los residuos aún mostraban cierta sobredispersión, confirmada también por los tests diagnósticos DHARMa.

Modelo binomial negativa tipo 1 (nbinom1)

Finalmente, se probó un modelo con distribución binomial negativa tipo 1, que en este caso permite que la varianza crezca linealmente con la media. Este modelo resolvió el problema de sobredispersión, con residuos bien comportados y un p-valor alto en el test de DHARMa.

Selección del mejor modelo GLM

Tras comprobar que la distribución nbinom1 era adecuada, se procedió a simplificar el modelo mediante eliminación manual de términos no significativos, ya que las funciones automáticas de selección como `step()` no están disponibles para modelos mixtos con efectos aleatorios. El modelo final retenido fue:

Abundance ~ PC2 + appearance:PC1 + appearance:PC2 + (1 | Site) + (1 | Specie)

Este modelo obtuvo el **AIC más bajo (1594)** entre todos los probados y presentó un excelente comportamiento predictivo, con un **RMSE de 13** en validación cruzada leave-one-out (LOOCV), frente a valores superiores a 15 en modelos anteriores.

Modelos GAM

Para la realización del análisis mediante modelos aditivos generalizados (GAM), seguimos una estrategia de selección manual "stepwise" de tipo forward, tal y como se detalla en el anexo de código. Al igual que en los modelos GLM, se utilizó una distribución binomial negativa para modelar la variable de conteo, dado que fue la que mejor resolvió la sobredispersión en los análisis previos.

Se evaluaron un total de 14 modelos diferentes. Como parte del proceso, probamos también aplicar suavizadores sobre las variables del **mejor modelo GLM** (es decir, un GAM con estructura equivalente: `Abundance ~ PC2 + appearance:PC1 + appearance:PC2 + s(Site, bs = "re") + s(Specie, bs = "re")`). Sin embargo, este modelo resultó tener un **AIC más alto** y una **deviance explicada en torno al 67 %**, muy inferior a otros modelos GAM probados.

El modelo con **mejor AIC** y **mayor deviance explicada (91 %)** fue el siguiente:

$$Y_i \sim \text{NegBin}(\mu_i, \theta), \quad i = 1, 2, \ldots, 336$$

$$\log(\mu_i) = \sum_{j=2}^{4} \beta_j \times \text{appearance}_i + (\text{Specie}_i, \text{PC1}_i) + s_2^{fs}(\text{Specie}_i, \text{PC2}_i) + s_3^{re}(\text{Site}_i) + s_4(\text{length}_i)$$

Este modelo incorpora términos suavizados que capturan interacciones no lineales complejas entre especie y los componentes principales ambientales (`PC1`, `PC2`), así como efectos aleatorios suaves de `Site` y un término de suavizado para la longitud corporal (`length`).

No obstante, pese a su buen rendimiento en términos de ajuste (AIC más bajo y mayor porcentaje de deviance explicada), **el modelo no resulta fácilmente interpretable desde el punto de vista ecológico**, principalmente porque **la variable `Specie` no se incluye como efecto aleatorio**, sino únicamente **interaccionando de forma no lineal con las componentes principales ambientales** (`PC1` y `PC2`). Esto impide capturar adecuadamente la variabilidad estructural entre especies y dificulta la generalización de los resultados, ya que las diferencias interespecíficas no están modeladas de forma explícita.

Además, el modelo introduce interacciones complejas mediante suavizadores, lo que impide extraer estimaciones marginales o efectos directos claramente atribuibles a las variables morfológicas. Las contribuciones individuales de estas variables (`appearance`, `length`) no resultan significativas, y su eliminación no modifica sustancialmente los indicadores de ajuste (AIC, deviance explicada), lo que sugiere que **no aportan información explicativa relevante en este marco aditivo**. Desde el punto de vista predictivo, **el RMSE en validación cruzada del mejor modelo GAM fue de 16**, valor que **no mejora el RMSE de 13 obtenido con el modelo GLM seleccionado**, a pesar de su mayor complejidad.

Modelos Bayesianos

En esta parte del análisis, hemos trasladado al contexto bayesiano el modelo que previamente había mostrado mejor rendimiento dentro de los modelos lineales generalizados (GLM). El modelo seleccionado fue:

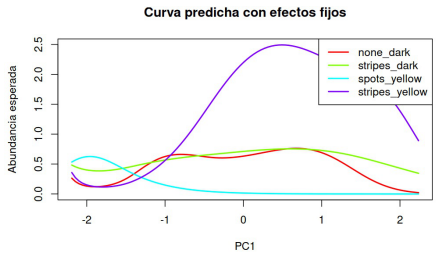
Abundance ~ PC2 + appearance:PC1 + appearance:PC2 + (1 | Site) + (1 | Specie)

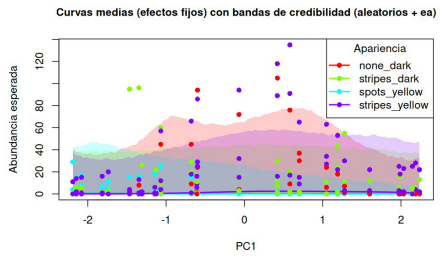
Este modelo incluye tanto efectos fijos como interacciones entre variables ambientales (`PC1`, `PC2`) y la apariencia (`appearance`), así como efectos aleatorios para capturar la variabilidad debida al sitio de muestreo y a la especie. Además, se parametrizó la varianza linealmente con el predictor lineal como en el modelo negativo binomial 1 visto en GLM. Suponíamos que la abundancia de arañas podía depender no solo de los valores individuales de las componentes principales, sino también de cómo se combinan con el tipo de hábitat. La implementación bayesiana se llevó a cabo en **JAGS**, utilizando distribuciones previas no informativas para los parámetros del modelo. La convergencia fue adecuada, sin señales de autocorrelación preocupante ni cadenas que no alcanzaran el equilibrio. El **Deviance Information Criterion (DIC)** asociado a este modelo fue de **1194**, lo cual refleja un ajuste razonable al conjunto de datos observado.

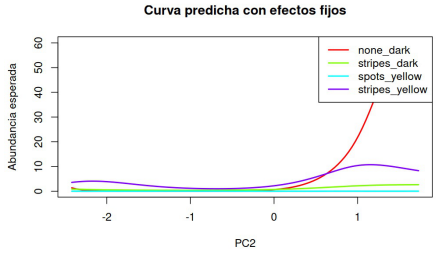
Sin embargo, en los análisis anteriores realizados con modelos aditivos generalizados (GAM), observamos que la inclusión de funciones de suavizado permitía capturar mejor la posible no linealidad en los efectos de las variables explicativas, obteniendo así mejores valores de AIC. Por esta razón, decidimos probar un modelo bayesiano más flexible, incorporando **splines** de interacción sobre las componentes ambientales implicadas en las interacciones (es decir, sobre `PC1` y `PC2`, condicionadas a los niveles de `appearance`). Esta decisión responde a la hipótesis de que la relación entre la abundancia de arañas y los gradientes ambientales puede no ser estrictamente lineal, especialmente cuando se modula por la apariencia. El nuevo modelo, al permitir formas funcionales más complejas, se adapta mejor a este tipo de patrón.

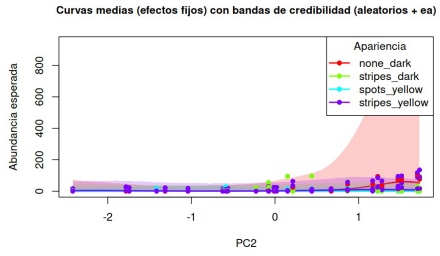
El modelo con splines fue también implementado en JAGS, utilizando una parametrización de tipo **B-spline** con un número moderado de nodos para evitar sobreajuste. Al igual que el anterior, este modelo mostró buena convergencia y estabilidad en las cadenas MCMC, sin señales de problemas con las distribuciones a priori. El **DIC obtenido fue de 1187**, lo que representa una mejora respecto al modelo lineal. Esta reducción en el DIC sugiere que la mayor flexibilidad ofrecida por los splines contribuye efectivamente a mejorar el ajuste sin penalizar en exceso la complejidad del modelo. Además, desde el punto de vista predictivo, este modelo con splines obtuvo un **RMSE (Root Mean Square Error) de aproximadamente 4**, lo cual es **sustancialmente menor que los RMSE obtenidos con los modelos GAM (16) y GLM (13)**. Este valor indica que las predicciones del modelo bayesiano con splines están, en promedio, a una distancia de solo 4 unidades de los valores reales de abundancia, lo que implica una **precisión mucho mayor** en la predicción.

En conjunto, estos resultados apoyan la idea de que la abundancia de arañas responde a **efectos no lineales** de las variables ambientales, modulados por el tipo de hábitat, y que la incorporación de splines dentro del marco bayesiano es una estrategia **potente y robusta** para capturar esa complejidad ecológica, ofreciendo además una **predicción más precisa que los enfoques clásicos**.









Los gráficos muestran las curvas medias del modelo según PC1 y PC2 para cada nivel de appearance, basadas únicamente en los efectos fijos. Estas curvas se quedan por debajo de los datos observados, indicando que los efectos fijos no capturan toda la variabilidad real. Las bandas de credibilidad incorporan la variabilidad de los efectos aleatorios y la dispersión modelada mediante ea, que dependen de la media. Estas fuentes de variabilidad ensanchan los intervalos alrededor de la media, aumentando la flexibilidad del modelo y mejorando el ajuste a los datos, pero también evidencian el problema de pocas observaciones y gran dispersión en nuestros datos.

Interpretación del modelo

Aunque el modelo bayesiano implementado en **JAGS** ofrece un mejor ajuste y mayor flexibilidad al incorporar términos no lineales mediante splines, su interpretación resulta compleja debido a la **gran cantidad de coeficientes generados**, especialmente al modelar interacciones con funciones de suavizado específicas por grupo. Por este motivo, en esta sección se opta por **interpretar el modelo GLM final**, que presenta una estructura más sencilla y fácilmente interpretable. Este modelo permite analizar de forma clara los efectos principales y las interacciones significativas entre los predictores ambientales y morfológicos sobre la abundancia de arañas.

Parámetros del modelo (efectos fijos):

Parámetro	Estimación	Error estándar	z value	p-valor	Significación
(Intercept)	1.243	0.303	4.097	< 0.001	***
PC2	1.776	0.213	8.342	< 0.001	***
appearance: none_dark × PC1	0.003	0.158	0.020	0.984	
appearance: stripes_dark × PC1	0.041	0.102	0.405	0.686	
appearance: spots_yellow × PC1	-1.041	0.223	-4.677	< 0.001	***
appearance: stripes_yellow × PC1	0.305	0.084	3.636	< 0.001	***
PC2 × appearance:stripes_dark	-1.246	0.227	-5.489	< 0.001	***
PC2 × appearance:spots_yellow	-1.654	0.287	-5.755	< 0.001	***
PC2 × appearance:stripes_yellow	-0.953	0.217	-4.384	< 0.001	***

El modelo ajustado tiene como fórmula:

Abundance ~ PC2 + appearance:PC1 + appearance:PC2 + (1 | Site) + (1 | Specie)

El nivel de referencia para la variable `appearance` es **none_dark**, por lo que tanto el intercepto como el efecto principal de `PC2` se interpretan en relación a este grupo. Las interacciones indican cómo cambia el efecto de los gradientes ambientales `PC1` y `PC2` en función del tipo de apariencia de las arañas. Por tanto, en condiciones promedio, se espera observar aproximadamente exp(1.43)= **3.5 individuos** por combinación especie–sitio para arañas con morfología `none_dark`. El intercepto (1.243) representa el logaritmo de la abundancia media esperada para arañas de apariencia `none_dark`, en condiciones ambientales promedio (`PC1 = 0`, `PC2 = 0`), controlando por sitio y especie.

El coeficiente positivo y altamente significativo de `PC2` (1.776) indica que, para las arañas `none_dark`, un aumento en la densidad de vegetación herbácea se asocia con un incremento notable en la abundancia. Este efecto es multiplicativo: por cada unidad de incremento en `PC2`, la abundancia esperada se multiplica por exp(1.776), es decir, casi por 6.

Respecto a las interacciones entre `appearance` y `PC1` (gradiente de sequedad y cobertura superficial), se observa que:

- Las arañas `spots_yellow` muestran una fuerte disminución en la abundancia con valores altos de `PC1` (coef. = -1.041, $p < 0.001$), lo que sugiere que este grupo es sensible a suelos más secos y con menos cobertura vegetal.
- Las arañas `stripes_yellow` muestran un efecto ligeramente positivo (coef. = 0.305, $p < 0.001$), indicando cierta afinidad por ambientes más secos o abiertos.
- En los grupos `none_dark` y `stripes_dark`, el efecto de `PC1` no es significativo.

En cuanto a las interacciones con `PC2`:

- Todos los grupos distintos a `none_dark` muestran efectos negativos significativos.
- En particular, las arañas `spots_yellow` presentan el mayor descenso en abundancia con el aumento de vegetación herbácea (coef. = -1.654).
- También se observan disminuciones importantes en `stripes_dark` (coef. = -1.246) y `stripes_yellow` (coef. = -0.953), todas con alta significación estadística.

En conjunto, estos resultados indican que el efecto del ambiente sobre la abundancia de arañas está fuertemente modulado por su apariencia. Mientras que las arañas `none_dark` se benefician de una mayor cobertura herbácea, otros grupos muestran respuestas negativas, especialmente en sitios con vegetación densa o condiciones más secas. Esto sugiere que la morfología externa puede estar relacionada con la capacidad de adaptación a diferentes microhábitats.

Conclusión

A lo largo del análisis se ha evaluado el efecto de características morfológicas y ambientales sobre la abundancia de arañas en distintos sitios de muestreo. El modelo GLMM seleccionado, basado en una distribución binomial negativa tipo 1, mostró buen ajuste e interpretabilidad, permitiendo identificar interacciones relevantes entre la apariencia de las arañas y los gradientes ambientales sintetizados mediante PCA.

Aunque los modelos GAM y bayesianos ofrecieron mejores métricas de ajuste (con un RMSE de 4 en el modelo bayesiano con splines frente a 13 en el GLMM), su complejidad estructural dificultaba una interpretación clara. Por ello, se optó por interpretar el GLMM final, que refleja que las respuestas a gradientes de sequedad y vegetación herbácea están fuertemente moduladas por la apariencia morfológica. Además, el análisis se ha visto condicionado por la **limitación en el tamaño del conjunto de datos**, con únicamente 336 observaciones y solo 12 especies. Esta escasez de datos hace que el problema sea especialmente complejo desde el punto de vista estadístico, ya que restringe la capacidad de estimar efectos finos y aumenta la sensibilidad a la sobreparametrización en modelos más flexibles.

En conjunto, estos resultados sugieren que la morfología externa puede estar relacionada con la capacidad de adaptación ecológica, reforzando la idea de que ciertos tipos de apariencia ofrecen ventajas o limitaciones bajo condiciones ambientales específicas, pero siendo dificil establecer los efectos de forma más concreta debido a la escasez de observaciones y dispersión de los datos.