Empleo de modelos estadísticos multivariantes para el análisis de comportamiento de comunidades ecológicas

Paula Silvar Viladomiu
Máster universitario en Bioinformática y Bioestadística UOC-UB
Estadística y bioinformática 27 aula 1

Fernando Carmona Berraquero
Carles Ventura Royo

02/01/2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Empleo de modelos estadísticos multivariantes para el análisis de comportamiento de comunidades ecológicas*
Nombre del autor:	*Paula Silvar Viladomiu*
Nombre del consultor/a:	*Fernando Carmona Berraquero*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega :	01/01/2018
Titulación::	*Máster universitario en Bioinformática y Bioestadística UOC-UB*
Área del Trabajo Final:	*Estadística y Bioinformática 27 aula 1*
Idioma del trabajo:	*Español*
Palabras clave	*Datos de abundancia de comunidad, modelos estadísticos, análisis multivariante*

Resumen del Trabajo (máximo 250 palabras):

Los análisis de datos de abundancia de especies de una comunidad analizados mediante un modelo estadístico ofrece muchas posibilidades de resolver preguntas clave en ecología. Avances en el estudio de modelos que puedan servir para este tipo de datos y avances tecnológicos han hecho posible el uso de este tipo de metodologías. En este trabajo se presenta un flujo de estudio para la aplicación de los modelos a este tipo de datos y se analizan dos casos prácticos para ilustrar como el uso de modelos para el análisis de datos de abundancia multivalentes ofrece una metodología eficaz y con mucho potencial futuro. En el primer caso práctico se construye un modelo para probar si la técnica de buceo afecta a la comunidad de peces observada. En el segundo ejemplo se construye un modelo con variables latentes y

variables ambientales para cuantificar como afectan a la comunidad.

Abstract (in English, 250 words or less):

In community ecology analysis of the abundance data using a statistical model approach offers multiple possibilities of addressing key questions in ecology. Advances in the research of models suitable for abundance data and in technology made possible the use of these methodologies. In this project a workflow for the analysis is presented and two practical cases are presented in order to illustrate how model approach is powerful and has a lot of future potential for community ecology. In the first example we test the hypothesis of the effect of the diving technique in the observed community. In the second example we incorporate latent and environmental variables in a model to study their relationship.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

Históricamente uno de los retos críticos para los ecólogos es ser capaces de entender cómo las comunidades ecológicas responden a las condiciones ambientales, para comprender el impacto potencial de cambios externos, tales como el cambio climático o modificaciones antropogénicas. Entender e identificar procesos detrás de las respuestas de la comunidad a cambios ambientales, así como clasificar especies en relación con su vulnerabilidad, es necesario para la gestión y conservación de comunidades ecológicas.

Los datos de abundancia de taxones tomados simultáneamente son una herramienta clave para estudios ecológicos a nivel de comunidad. El análisis de este tipo de datos ecológicos multivariantes supone un problema ya que contienen muchos ceros y las bases de datos son de gran dimensionalidad, debido a la existencia de muchos taxones. Además, suelen presentar alta colinealidad entre las variables. El análisis de datos comunidades comenzó como una ciencia descriptiva que resolvía cuestiones especificas directamente relacionadas con la estructura multivariante de las comunidades y su relación con el ambiente (Borcard et al. 2011).

Recientemente se ha extendido la aplicación de modelos estadísticos multivariantes para este tipo de análisis. Los modelos multivariantes ecológicos han avanzado de tal manera que es posible la construcción de modelos apropiados que aborden cuestiones claves de ecología de una manera estadísticamente coherente (Warton et al. 2015b). Se cree que mediante el uso de modelos se puede crear un marco flexible y poderoso para abordar estos retos en ecología (Hui et al. 2015; Warton et al. 2015a, b; Lyons et al. 2016). Además, dado que los ecosistemas marinos están menos estudiados en comparación con los progresos que se han hecho con los ecosistemas terrestres (Klais et al. 2017), se realizará una aplicación práctica de modelos estadísticos multivariantes en un ecosistema marino.

En resumen, en el presente trabajo se desarrollará el uso de modelos aplicados a estudiar el comportamiento de comunidades ecológicas. Adicionalmente, se pretende ejemplificar mediante dos casos prácticos de flujos de trabajo de uso de modelos estadísticos en comunidades marinas utilizando R.

1.2 Objetivos del Trabajo

Objetivos generales

- Adquirir los conocimientos necesarios para el uso adecuado de modelos estadísticos multivariantes ecológicos.
- Aplicar los modelos estadísticos multivariantes adecuados en R a dos conjunto de datos real de una comunidad ecológica marina.

Objetivos específicos

- Realizar una revisión inicial bibliográfica sobre el material publicado en relación con modelos multivariantes aplicados al ámbito ecológico
- Evaluar la idoneidad de los diferentes modelos estadísticos para el análisis de datos ecológicos de comunidades. Identificar fortalezas y debilidades.
- Analizar la relación entre los datos de origen disponibles y el modelo multivariante a aplicar.
- Desarrollar y comentar en profundidad y detalle técnico el análisis de los modelos multivariantes para datos ecológicos en R.
- Presentar un modelo estadístico multivariante real sobre datos ecológicos utilizando R.
- Interpretar y discutir los resultados del modelo

1.3 Enfoque y método seguido

El presente trabajo aborda el análisis de diferentes datos ecológicos utilizando modelos estadísticos multivariantes, con el objeto de comprobar las ventajas de dichos modelos para el análisis de este tipo de comunidades. Los modelos multivariantes ofrecen muchas posibilidades y disponen de características óptimas a la hora de analizar datos ecológicos. Publicaciones relativamente recientes (Wang et al. 2013; Brown et al. 2014; Hui et al. 2015; Warton et al. 2015a, b) sostienen que el análisis multivariante basado en modelos presenta mejores propiedades que los enfoques basados en proximidades y algoritmos, por su flexibilidad, facilidad de interpretación y mejoría en relación al poder predictivo. En la parte práctica del proyecto se explorarán los datos disponibles de la comunidad marina y se examinarán sus propiedades. A continuación, debido a la diversidad de modelos y metodologías existentes, se discutirá la idoneidad del modelo a aplicar, mediante una revisión y comparación de las aproximaciones utilizadas en este ámbito.

Para la realización del análisis estadístico, existen diferentes programas disponibles. En el presente trabajo se utilizará R, en concreto R Studio, debido a su acceso libre y sus múltiples posibilidades para la programación estadística. Mediante scripts en el lenguaje R se construirá un modelo multivariante para datos reales de una comunidad ecológica marina. Se usará R MarkDown para guardar el pipeline utilizado.

1.4 Planificación del Trabajo

1.4.1. Tareas

Desarrollo del trabajo- Fase 1 (17/10/2017- 20/11/2017)

- Revisión bibliográfica sobre modelos multivariantes ecológicos (métodos estadísticos).
- Investigación de los paquetes y pipelines de R para la construcción de modelos multivariantes ecológicos.
- Análisis y comparación de modelos estadísticos multivariantes y metodología asociada aplicados a comunidades ecológicas. Descripción de fortalezas y debilidades asociadas.

- Depuración y preparación de la base de datos en R.
- Exploración de las propiedades de los datos en R y realización de las transformaciones pertinentes.
- Descripción de modelos multivariantes disponibles en R para el análisis de datos de comunidades ecológicas.

Desarrollo del trabajo- Fase 2 (21/11/2017- 18/12/2017)

- Elección y construcción del modelo en R (analizar la distribución de las variables respuesta y la correlación entre las mismas).
- Comprobación y diagnóstico del modelo (gráficos de los residuos, evaluación de asunciones del modelo, etc.).
- Estudio del alcance de los resultados del modelo.

Redacción de la memoria (19/12/2017- 02/01/2018)

- Redacción de la introducción.
- Recopilación de los capítulos, estructuración de la memoria.
- Desarrollo de la discusión y conclusiones asociadas.

Elaboración de la presentación (03/01/2018- 10/01/2018)

- Elaboración de la presentación**.**

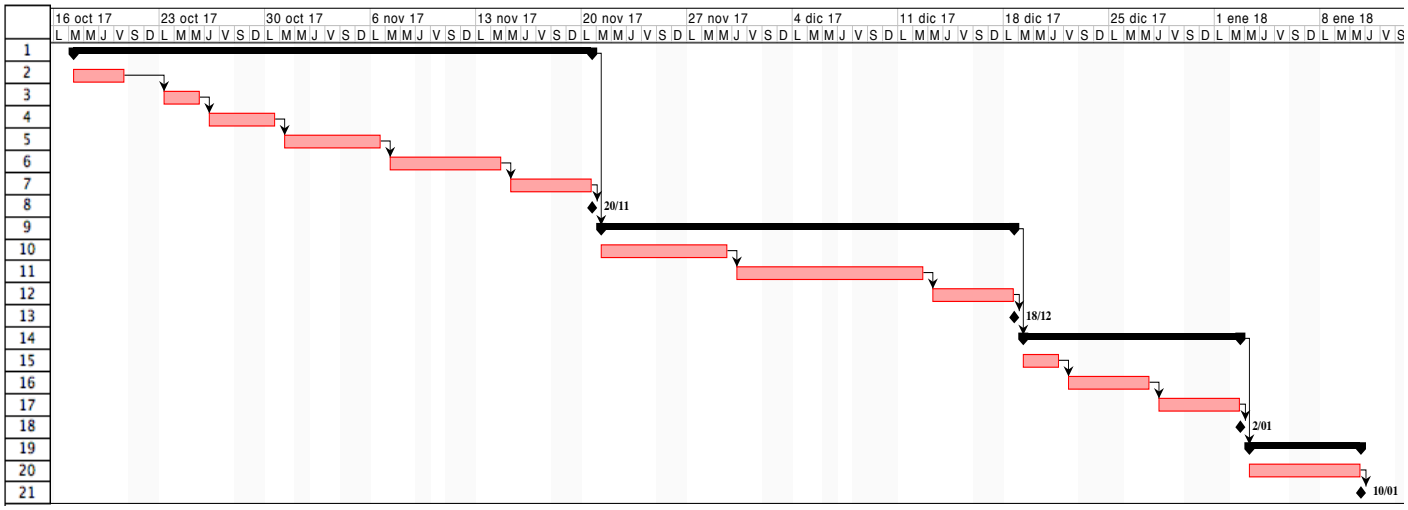
Defensa publica (11/01/2018- 22/01/2018)

- Preparación de la defensa pública

1.4.2. Calendario

A continuación, se presenta la tabla de tareas y el Diagrama de *Gantt* creados con el software Project libre$_{tm}$.

	Nombre	Duracion	Inicio
1	**Desarrollo del trabajo-Fase 1**	**30 days**	**17/10/17 8:00**
2	Revisión bibliográfica sobre modelos multivariantes ecológicos	4 days	17/10/17 8:00
3	Investigación de los paquetes y pipelines de R	3 days	23/10/17 8:00
4	Análisis y comparación de los modelos estadísticos multivariantes	3 days	26/10/17 8:00
5	Depuración y preparación de la base de datos en R	4 days	31/10/17 8:00
6	Exploración de las propiedades de los datos en R y realización de las transformaciones pertinentes	5 days	7/11/17 8:00
7	Descripción de modelos multivariantes disponibles en R para el análisis de datos de comunidades ecológicas	4 days	15/11/17 8:00
8	Entrega del Informe de seguimiento del proyecto 1	0 days	20/11/17 18:00
9	**Desarrollo del trabajo- Fase 2**	**18 days**	**21/11/17 8:00**
10	Elección y construcción del modelo en R	7 days	21/11/17 8:00
11	Comprobación y diagnóstico del modelo (gráficos de los residuos, evaluar asunciones del modelo, etc.)	7 days	30/11/17 8:00
12	Estudio del alcance de los resultados del modelo.	4 days	13/12/17 8:00
13	Entrega del Informe de seguimiento del proyecto 2	0 days	18/12/17 17:00
14	**Redaccion de la memoria**	**9 days**	**19/12/17 8:00**
15	Redacción de la introducción.	3 days	19/12/17 8:00
16	Recopilación de los capítulos, estructuración de la memoria.	3 days	22/12/17 8:00
17	Desarrollo de la discusión y conclusiones asociadas.	3 days	28/12/17 8:00
18	Entrega de la Memoria del TFM	0 days	2/01/18 17:00
19	**Elaboracion de la presentacion**	**6 days**	**3/01/18 8:00**
20	Elaboración de la presentación.	6 days	3/01/18 8:00
21	Defensa publica del TFM	0 days	10/01/18 17:00



1.5 Breve sumario de productos obtenidos

Primero se describen los tipos de datos colectados por ecólogos de comunidades, se describe la metodología de análisis tradicional y se presenta la aplicación de modelos estadísticos para analizar este tipo de datos y el flujo de trabajo a seguir. Posteriormente se presentan dos casos prácticos de cómo aplicar el flujo de trabajo para el análisis de datos multivariantes de comunidades ecológicas marinas (el código completo de R se puede encontrar

en el anexo). Finalmente se discuten las ventajas y las limitaciones de las metodologías.

1.6 Breve descripción de los otros capítulos de la memoria

Análisis de datos de comunidades ecológicas

Este capitulo examina los datos de comunidades ecológicas y su análisis. Se revisan las metodologías mas comúnmente utilizadas y sus fortalezas y debilidades.

Casos prácticos

El capítulo de casos prácticos detalla dos procesos de análisis de dos conjuntos de datos de abundancias de especies de peces de la comunidad, para ejemplificar la aplicación de modelos estadísticos para el análisis de datos multivariantes de comunidades ecológicas utilizando el software estadístico R. El primer caso analiza datos de abundancia de comunidades de peces costeras en Croacia con un modelo factorial. El segundo caso estudia comunidades de peces del mar de Barents con modelos con variables latentes.

Discusión

El último capitulo contiene la discusión sobre los resultados de los casos prácticos y alcance de los resultados de los modelos. También se realiza un reflexión critica sobre los objetivos y el seguimiento de la planificación y la metodología. Por último se proponen líneas de trabajo futuro que no se han podido abordar en este trabajo.

Conclusión

En el último capitulo se describen las conclusiones del trabajo.

2. Análisis de datos de comunidades ecológicas

2.1. Datos ecológicos

Los datos de abundancia de especies recolectados simultáneamente son una herramienta clave para estudiar cómo cambia la estructura de las comunidades en respuesta al cambio en las condiciones ambientales (Warton et al. 2015b). Los datos de abundancia pueden ser de varios tipos: presencia/ausencia, presencia sólo, conteos, porcentajes de cobertura, biomasa etc. Hay muchas maneras de tomar datos de abundancia directa o indirectamente y hay que tener cuidado ya que las observaciones en estudios de abundancia de especies y su distribución suelen ser a menudo imperfectas (Kellner and Swihart 2014). Los datos de abundancia generalmente están acompañados de datos de variables ambientales que son importantes para explicar la composición de la comunidad.

Las bases de datos de abundancia de especies tienen unas particularidades especiales que hace que el análisis de este tipo de datos ecológicos sea complejo. Contienen una gran proporción de ceros lo que hace que las variables contengan poca información y son bases de datos de gran dimensionalidad, debido a la existencia de muchos taxones. Los datos de abundancia son variables que no responden a distribuciones normales.

El conjunto de variables que contiene datos de abundancia de una comunidad frecuentemente presenta estructuras de dependencia internas, por lo cual pueden presentar alta colinealidad entre las variables. Asimismo hay que tener en cuenta que pueden presentar autocorrelación espacial y/o temporal entre muestras. También se suelen presentar una correlación entre la media y la

varianza muy fuerte, lo cual debe ser adecuadamente explicado para asegurar un análisis correcto.

2.2. Métodos de análisis multivariante tradicionales

El análisis multivariante en ecología tradicionalmente estaba basado en matrices de disimilaridades, generalizaciones de análisis de correspondencia o análisis redundantes. Son métodos descriptivos, más sencillos en su aplicación ya que no requieren hipótesis ni asunciones previas.

Los métodos de análisis multivariante como análisis de componentes principales (PCA), escalamiento multidimensional no métrico (NMDS) y el análisis de correspondencia canónico (CCA), son métodos puramente algorítmicos que se basan en ideas geométricas y sintetizan la información de las variables reduciendo su dimensión. No tienen en cuenta las propiedades estadísticas de los datos, y su conexión con contraste de hipótesis es débil, por lo que se pierden oportunidades de detectar estructuras y procesos ecológicos.

Algunas de sus limitaciones son que no son capaces manejar algunos tipos de datos y tienen dificultad para alcanzar una solución estable única. Las métricas de proximidad y similaridad a menudo fallan en especificar la relación entre la media y la varianza (Warton et al. 2012), y asumen incorrectamente que con una transformación esto se puede estabilizar dicha relación (O'Hara and Kotze 2010). Algunos métodos están limitados por las medidas de fidelidad y homogeneidad, y confían en elecciones equivocadas de transformación y métricas de disimilaridades (Lyons et al. 2016). Además, muestran algunos problemas potencialmente serios en su interpretación y actuación, ya que tienen poco poder estadístico (Warton et al. 2015b), por lo cual ofrecen un entendimiento pobre sobre la comunidad, y además su interpretación puede ser poco clara o engañosa, lo que puede llevar a conclusiones erróneas (Hui et al. 2015).

Tabla1. Debilidades y consecuencias de los métodos tradicionales de análisis multivariante.

Debilidad	Consecuencia
No tienen en cuenta las propiedades de los datos	No se pueden explicar tendencias presentes de media-varianza
No tienen capacidad de conexión con teorías	Pierden oportunidades de detectar estructuras y procesos ecológicos
Poco poder estadístico, no se puede hacer inferencia formal sobre los datos	Un entendimiento pobre de la ecología de la comunidad
Puede haber una interpretación no clara o engañosa	Se sacan conclusiones erróneas

2.3. Uso de modelos en ecología de comunidades

Los modelos estadísticos constituyen una representación matemática explícita que explica la variación sistemática y el ruido en los datos. El enfoque del análisis utilizando modelos permite lidiar con las características de los datos de abundancia de comunidades. El uso de técnicas de remuestreo o el hecho de que ajustamos el modelo a los datos permite que este método sea capaz de manejar la correlación entre los datos y la correlación de la media y la varianza. De esta forma, ofrecen ventajas en relación a su interpretación, flexibilidad y eficiencia (Warton et al. 2015b). Mediante el empleo de modelos se pueden cuantificar directamente las relaciones ecológicas y hacer predicciones importantes. También se puede evaluar y comparar clasificaciones objetiva y transparentemente mediante el diagnóstico del modelo.

El uso de modelos permite establecer un marco de análisis que puede manejar la mayoría de datos de abundancia, responder distintos tipos de preguntas y tiene capacidad de incorporar explícitamente teorías ecológicas en los modelos (Brown et al. 2014). Los modelos son suficientemente flexibles para elegir una especificación y distribución del error que sea apropiada para los datos de

origen. Para explicar la autocorrelación espacial o temporal de las muestras se puede incluir efectos aleatorios en el modelo. Cuando el modelo se ajusta bien a los datos, ofrece poder estadístico para realizar inferencia sólida a nivel de comunidad ecológica, mucho mayor que las metodologías alternativas.

Una de las complicaciones del uso de modelos es su gran variedad, la falta de flujos de trabajo y software con funciones especificas disponibles para ecólogos. No existe un único modelo de aplicación universal , por lo que se necesita perspectiva crítica. Parte del trabajo de ajuste y evaluación del modelo se orienta en realizar un cribado inicial y, de entre todos los modelos adecuados, seleccionar el que explica la mayor proporción de la varianza sujeto a la restricción de todos los parámetros del modelo.

Las aplicaciones de modelos para el análisis de datos multivariantes de comunidades ecológicas son muy extensas, dependiendo de los datos disponibles y el tipo de respuesta estadística que se quiere obtener. Algunos ejemplos de aplicaciones en ecología son:

- Descripción de las comunidades ecológicas
- Extensión del estudio de interacciones entre especies
    - Construcción y ordenación basada en modelos
    - Modelos que expliquen las fuentes de correlaciones entre especies
    - Detección de correlación residual entre los taxones
- Modelos predictivos con capacidad de mejorar los pronósticos
- Inferencia multivariable sobre los predictores
- Análisis confirmatorios
- Explicación de predictores no especificados en el modelo

La mayor desventaja del uso de modelos es que son muy exigentes computacionalmente, debido a las técnicas de ajuste y remuestreo. No obstante, gracias a los avances tecnológicos, esto ya no es una barrera para su uso, aunque tiene impacto en el tiempo computacional del análisis.

2.4. Tipos de modelos

Los **modelos de distribución de especies (SDM)** han sido ampliamente utilizados para explicar y predecir cómo un único taxón responde a la variación ambiental. Sin embargo, en este trabajo se presentan modelos que manejan datos multivariantes de comunidades, es decir, pueden analizar las abundancias del conjunto de muchos grupos taxonómicos, para simultáneamente explorar interacciones entre los taxones y la respuesta de la abundancia a variables ambientales.

Los modelos conjuntos usados para el estudio de datos multivariantes de comunidades ecológicas son modelos jerarquizados, extensiones de los **modelos lineales generalizados (GML)**. Los GML son una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores y varianzas no constantes. Una de las adaptaciones necesarias es la inclusión de efectos aleatorios para capturar la correlación entre la abundancia de las especies. Una opción compleja de incorporar dicha correlación es introducirla directamente con un efecto aleatorio multivariante aplicado a cada muestra, construyendo un **modelo lineal mixto generalizado multivariante (GLMM)**. Estos modelos son especialmente útiles cuando el número de taxones es pequeño en comparación con el número de muestras.

$$g(m_{ij}) = \alpha_i + \beta_{0j} + x'_i\beta_j + u_{ij}$$

El GLMM es una extensión del GLM que especifica efectos aleatorios multivariados u_{ij} para cada muestra que capturar la correlación entre los taxones. m_{ij} es la abundancia media, $g(\,)$ es la función de vinculo, x' es el vector traspuesto de x que son los predictores y, para taxón j , β_{0j} es un intercepto y β_j es un vector de coeficientes de regresión relacionados con los

predictores medidos. El efecto de la muestra α_i es opcional y se ajusta para la abundancia total de cada muestra o riqueza, si se quiere definir el modelo enfocándolo en abundancia relativa o composición en lugar de abundancia absoluta.

Una manera flexible de incorporar esta correlación es usar **modelos con variables latentes (LVM)**, que introducen predictores no observados ("latentes") a cada muestra. El número de estas variables controla la complejidad del modelo (Warton et al. 2015a), y éstas median las interacciones entre taxones, pero esos valores de las medidas no están incluidos en el modelo.

$$g(m_{ij}) = \alpha_i + \beta_{0j} + x'_i\beta_j + u_{ij}$$

LVM es una función de las variables latentes z_i y de los predictores medidos. u_{ij} debe estar linealmente relacionado con el set de variables latentes $u_{ij} = z'_i\lambda_j$.

La función de vínculo, $g(\ )$, hace referencia a la transformación de las variables antes de ajustar el modelo. La función vínculo relaciona la esperanza matemática de la variables dependientes con el predictor lineal (Cayuela et al. 2016). La función de vínculo que se usa habitualmente en modelos de datos ecológicos es de familia exponencial, p.e. *log* para conteos o *logit* para presencia/ausencia, que transforman con logaritmos o logaritmos naturales la variable respuesta respectivamente .

2.5. Construcción de modelos

El paso crítico para la construcción de modelos multivariantes en ecología es la especificación de una **distribución** del error que sea apropiada para los datos de origen. Como se ha mencionado anteriormente, los datos de abundancia

mantienen una estructura no normal. Los modelos permiten especificar distintos tipos de distribuciones de errores (Poisson, binomial, gamma, exponencial). La siguiente tabla muestra las distribuciones más comunes para datos de abundancias y las funciones de vínculo a especificar en el modelo en cada caso.

Para datos de conteos se suele recomendar la distribución de Poisson o la binomial negativa. Muchos ecólogos utilizan la binomial negativa porque es discreta como la Poisson pero su varianza puede ser mayor que su media, por lo cual es más adecuada para describir datos sobre dispersos con distribuciones agrupadas, que no tienen limite superior intrínseco y con mayor varianza que la de Poisson (Bolker 2007).

Tabla 2. Distribuciones de errores y funciones de vinculo para tipos de datos de abundancia.

Tipo de datos	Distribución de errores	Función de vinculo
Presencia/ausencia	Binomial	logit
Conteo	Poisson	log, raíz cuadrada
	Negativa binomial	log
Ordinal	Multinomial	logit; “proportional odds”
Biomasa	Tweedie	log
Porcentaje	Binomial	logit

Al construir el modelo se tienen que estimar los parámetros, para ello se quiere encontrar los parámetros que hacen que el modelo se ajuste mejor a los datos. Para el **ajuste del modelo** los métodos mas comúnmente utilizados en la estimación de los parámetros son:

- **Máxima verosimilitud**

  La función de verosimilitud indica la probabilidad de obtener muestra observada en función de posibles valores de los parámetros poblacionales. Por lo tanto, cuando se maximiza la función de verosimilitud, se determinan los parámetros poblacionales que tienen

mayor probabilidad de producir los datos observados. La mayor dificultad de este método reside en la complejidad de integrar sobre valores no medidos. La técnica más común para estimar GLMM y LVMs es la aproximación de Laplace, pero esta puede ser imprecisa en casos de muestras pequeñas (Warton et al. 2015a). Otra técnica disponible es la cuadratura adaptativa, que mejora la precisión pero con algunos costes computacionales, especialmente cuando se necesitan más de dos variables latentes.

- **Estimación Bayesiana**
  La dificultad principal reside en estimar simultáneamente distribuciones para un gran número de parámetros, y ésto puede llevar a problemas de convergencia o a exigir grandes cadenas de algoritmos. El método principal de implementación son los algoritmos de Cadenas Markov de Monte Carlo (MCMC).

La revisión del **diagnóstico** es un paso critico para el ajuste del modelo ya que nos permite asegurar que el modelo provee concordancia con los datos observados.

2.6. Paquetes de R

En el presente trabajo se utilizará el software estadístico R como marco computacional, al tratarse de un código libre, rico en diversidad y complejidad funcional. Existen muchos paquetes de R que pueden ajustar modelos para usar en ecología. Como se observa en la tabla 3, los paquetes varían en el tipo de modelo que implementan y en el proceso de ajuste que utilizan, el cual suele ser muy exigente computacionalmente. En el presente trabajo se ha realizado una selección de paquetes enfocada a la implementación de GML, GMML y LVM.

El paquete ***mvabund*** (Wang et al. 2013) ofrece un set de herramientas para modelar y analizar los datos multivariantes de abundancia en ecología de comunidades mediante modelos lineales generalizados. La función clave para el ajuste es *manyglm()* que ajusta un GLM separado para cada especie, utilizando variables predictoras. Existe una enfatización en inferencias basadas en el diseño del experimento, y en herramientas de presentación de diagnóstico para la comprobación de asunciones, especialmente mediante gráficos.

El paquete ***lme4*** (Bates et al. 2016) puede ajustar modelos lineales generalizados con efectos mixtos mediante máxima verosimilitud o verosimilitud máxima restringida (función *glmer()*).

El paquete ***ltm*** (Rizopoulos 2006) contiene funciones para realizar análisis de datos multivariantes dicotómicos y policotómicos usando modelos de rasgos latentes.

El paquete ***boral*** (Hui 2016) se centra en el análisis de datos ecológicos multivariantes mediante un enfoque Bayesiano. La estimación de los parámetros del modelo se realiza mediante métodos MCMC vía JAGs.

El paquete ***HMSC*** (Ovaskainen et al. 2017) ofrece un marco de análisis denominado “Hierarchical Modelling of Species Communities”, basado en modelos jerárquicos Bayesianos para comunidades de especies, para explicar condiciones ambientales, rasgos de especies y filogenia. Este marco también incluye variables latentes auto-correlacionadas espacial o temporalmente.

El paquete ***misnet*** (Harris 2015) incluye funciones para realizar predicción estructurada con redes neuronales estocásticas en R. Tiene utilizad para definir densidades de probabilidad sobre posibles grupos de especies.

Tabla 3. Paquetes de R para ajuste de modelos en ecología. GLM (Modelos lineales generalizados); GLMM (Modelos lineales mixtos generalizados); LVM

(Modelos con variables latentes); MLE (Máxima verosimilitud); MCMC (Cadenas de Markov Monte Carlo); * aceptable, ** buena, *** excelente

Paquete de R	Modelo	Ajuste/Estimación	Datos	Velocidad
`mvabund`	GLM	MLE	Conteos	***
`lme4`	GLMM y LVM	MLE o verosimilitud restringida	Conteos	**
`ltm`	LVM	MLE restringida	Ordinal	***
`boral`	LVM	Método bayesiano MCMC	Conteos, ordinal	*
`HMSC`	LVM	Método bayesiano MCMC	Conteos	**
`mistnet`	LVM (redes neuronales)	MLE	Conteos, ordinal	**

2.7. Fases de análisis

Exploración de los datos

La exploración de los datos es muy importante ya que, antes de ajustar el modelo, se deben identificar algunas propiedades clave de los mismos para evitar errores estadísticos. Es interesante explorar la relación entre la media y la varianza de las muestras, para elegir la estructura de los errores y función de vinculo, aunque esta elección puede ser mas difícil de lo esperado. Las variables respuesta son las variables de abundancia de taxones en la comunidad. En el caso de que las variables predictoras estén sesgadas requerirán transformaciones previas a la construcción del modelo.

Elección y ajuste del modelo

La cuestión más importante que define la elección del modelo es la pregunta que se quiere resolver sobre los datos. Además, es importante la elección de la distribución para modelar la respuesta de las especies. Cuando se trata de conteos se recomienda la distribución de Poisson o la binomial negativa. Algunos ecólogos eligen la distribución binomial negativa porque admite que su varianza puede ser mayor que su media (p.e. cuando hay sobredispersión).

También es una buena elección cuando se esta describiendo una distribución dispersa y desigual, sin límite superior intrínseco y que tiene más varianza que la Poisson (Bolker 2007).

Evaluación del modelo

Es fundamental comprobar las asunciones del modelo. Se analizan los residuos del modelo para constatar la idoneidad de la distribución de errores elegida. Puede ser recomendable comparar modelos con distintas funciones de vínculo para ver cuál se ajusta mejor a nuestros datos. En cualquier caso es conveniente analizar los siguientes gráficos:

- Histograma de residuos
- Gráfico de residuos frente a valores estimados
- Gráfico probabilístico de normalidad (qq-plot)

Para comprobar el ajuste del modelo a los datos debemos prestar particular atención a los test de significación para los estimadores del modelo, así como a la proporción de varianza explicada por el mismo.

Interpretación

La inferencia estadística consiste en derivar conclusiones estadísticas y biológicas de los datos examinando los estimadores y sus intervalos de confianza, testando hipótesis y seleccionando el mejor modelo o modelos (modelos anidados). Los resultados de los modelos nos pueden permitir entender tanto la respuesta de la comunidad a los cambios ambientales como las respuestas individuales de los taxones a dichas modificaciones

3. Casos prácticos

En este trabajo se presentan dos casos para ilustrar cómo el uso de modelos es una herramienta poderosa para obtener conocimientos sobre problemas ecológicos no alcanzables anteriormente con las metodologías de análisis multivariantes. El código del flujo de trabajo que se ha utilizado para el análisis de los dos casos prácticos se puede encontrar en anexo.

3.1. Comunidades de peces costeras en Croacia

Objetivo

El primer ejemplo ilustra el uso de modelos en un estudio con un diseño experimental donde se utilizan 2 metodologías de muestreo en 4 lugares diferentes. Se pretende estudiar si el empleo de diferentes técnicas de muestreo afecta a los datos de abundancia obtenidos y si en esas diferencias influye la localidad en la que se muestreo. El código de R se puede encontrar en el anexo 1 y la tabla de equivalencias de las abreviaciones con el nombre científico de las especies se puede encontrar en el anexo 3.

Datos

La base de datos llamada Hr Fish fue cedida por 2000 Miljia, una ONG croata que se dedica a la investigación marina. Son datos de censos visuales de comunidades de peces costeras de Croacia. Los datos se recolectaron en 2017 en 4 localidades diferentes de Croacia, y se usaron dos técnicas de buceo diferentes, “rebreather” y “open-circuit”, cuya diferencia principal es que la primera técnica de buceo no libera burbujas. La información de conteos de peces por especies en cada muestra está almacenada en la base de datos *hr_fish_ab* y la información sobre la localidad y la técnica de buceo está registrada en la base de datos *hr_fac*. Las filas de las bases de datos son muestras representando transectos de censos visuales.

En primer lugar eliminamos de la base de datos las especies de peces que se pretendían estudiar que no fueron observadas en ningún transecto. Se consideran especies raras aquellas que aparecen en una sola muestra, se identifican como raras *Labrus viridis, Pagrus pagrus, Phycis physis, Sarda sarda, Scorpaena scrofa, Spicara smaris, Sphyraena sphyraena, Anthias antias, Symphodus cinereus.* Estas especies son excluidas del análisis ya que son variables que no aportan casi información y pueden afectar a los resultados reduciendo su sensibilidad.

Para visualizar la relación entre la media y la varianza de los datos, se construye un gráfico con la varianza de las muestras contra la media de las muestras (fig. 1). Se observa una gran tendencia creciente y una varianza, mucho mayor que la media, lo que sugiere datos sobre-dispersos. Dado que los datos son conteos y el gráfico muestra una relación media-varianza de las muestras que puede ser cuadrática, consideramos una distribución binomial negativa para el modelo.

Modelo

Se construye un modelo según el diseño del muestreo para probar si el método de muestreo afecta a las comunidades encontradas. Se sigue la metodología de análisis por bloques espaciales o localidades (Wang et al. 2013). Se usa el paquete ***mvabund*** el cual fue diseñado para pruebas multivariantes de hipótesis, ya que es un marco flexible y poderoso para el análisis de datos de abundancia. Queremos testar la hipótesis de que la comunidad de peces captada en las muestras se ve afectada según el tipo de técnica de buceo y si esta interacciona con la localidad en la que se tomaron las muestras.

Una de las asunciones claves del modelo es la independencia. En este caso se ha garantizado que los conteos sean independientes entre las localidades. La independencia de las localidades es una asunción muy importante en análisis multivariante y sólo puede asegurarse mediante un diseño apropiado. También se asume implícitamente que hay independencia entre las especies ya que se

usa el estimador de máxima verosimilitud para cada especie. Esta última asunción se relaja en el test de hipótesis.

Se ajustar el modelo predictivo usando la función *manyglm()*, la cual ajusta un modelo generalizado lineal (GLM) para cada una de las especies, utilizando una función logaritmica (log) como función vínculo. Se especifica un modelo con dos factores ortogonales con la siguiente formula: ~localidad*técnica de buceo. El argumento *family* especifica la distribución asumida de los datos, en este caso se especifica "negative.binomial" .

Evaluación del modelo

El ajuste con la función *manyglm()* asume la relación entre la media y la varianza especificada con la elección de la distribución y una función vínculo logarítmica. Para evaluar si modelo se ajusta bien a los datos se analizan los residuos mediante gráficos construidos con la función *plot.manyglm()*. Para ejemplificar un mal ajuste del modelo debido a la elección de una distribución no apropiada ajustamos un segundo modelo con distribución Poisson. Los gráficos de los residuos de Dunn-Smith contra los valores ajustados para el modelo con distribución binomial negativa y el modelo con distribución Poisson (fig.2 y 3) muestran que la distribución Poisson no es adecuada para modelar la relación entre la media y la varianza ya que en el gráfico se ve una clara forma de embudo. También se puede construir un gráfico Q-Q, utilizando el parámetro *which=2* (fig. 4) que muestra los residuos del modelo Dunn-Smyth contra los cuantiles teóricos para ver si se ajustan a la distribución normal. En general observamos que el ajuste del modelo asumiendo distribución binomial negativa de los datos de abundancia puede ser verosímil.

Interpretación

Las hipótesis multivariantes se pueden comprobar con la función *anova(),* que usa pruebas basadas en el remuestreo para hacer inferencia sobre que factores están asociados con las variables de abundancia, y devuelve una tabla con los test de significancia para cada término del modelo. El argumento *p.uni* nos permite ver los resultados de los modelos univariantes

especie por especie. El argumento test especifica el test estadístico usado, que puede ser “LR” para máxima versatilidad, “wald” o “score”. El ANOVA (Paso 4-Interpretación del anexo 1) nos devuelve tanto los resultados test de significación general como el individual para cada especie. Tanto las diferencias entre localidades como entre técnicas de buceo obtienen un pvalor muy bajo por lo que se concluye que son significativas, sugiriendo una buena evidencia del efecto de la técnica de buceo en el conteo de especies de la comunidad de peces y entre las localidades. La interacción entre localidades y técnica de buceo también es significativa, lo cual muestra que el efecto multiplicativo en la abundancia media de la técnica de buceo usada no es consistente entre localidades.

La tabla del ANOVA muestra un test estadístico calculado asumiendo independencia de las variables de abundancia de especies, pero los pvalores están ajustados para controlar la tasa de error por familia entre las especies. Para las especies que mostraron pvalores ajustados bajos se evidencia diferencias de abundancia significativas en las diferentes localidades son *Dentex dentex*, *Diplodus sargus*, *Labrus merula*, *Mullus surmuletus*, *Spondyliosoma cantharus*, *Coris julis*, *Symphodus ocellatus*, *Symphodus tinca*, *Thalassoma pavo*. Las especies *Dentex dentex*, *Diplodus vulgaris* y *Serranus scriba* muestran un efecto significativo de la técnica de buceo en la abundancia de estas especies. Poniendo de manifiesto que estas especies son las más afectadas por respectivamente la localidad y la técnica de buceo.

También se ha realizado el ANOVA de modelos anidados con modelos ajustados con el paquete *mvabund*. Se realiza uno con el modelo anterior y un modelo sin el factor localidad que obtiene resultados de pvalor muy bajo (<0.05), lo cual muestra que esta reducción del modelo no es apropiada.

Figuras y tablas

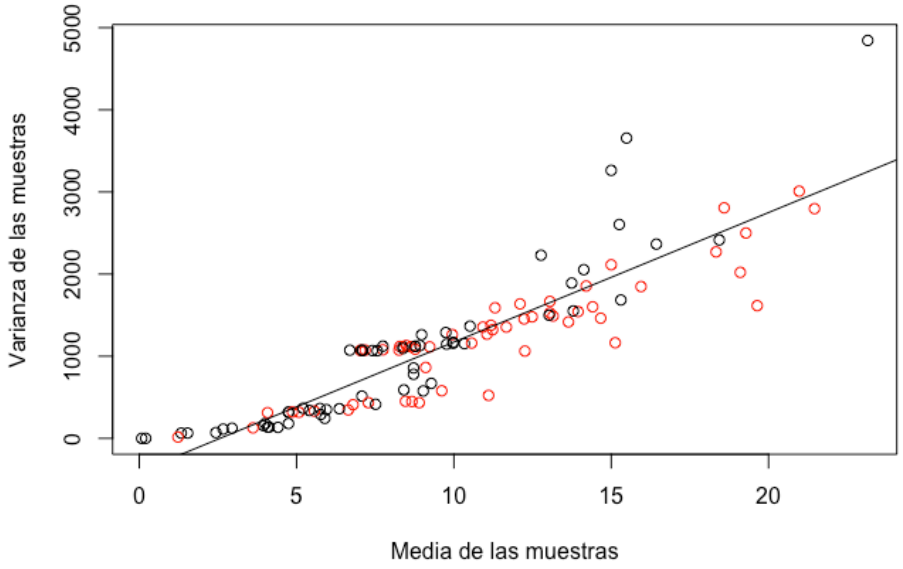


Figura 1. Gráfico de la relación entre la media y la varianza de las muestras. Cada punto representa la media de la muestra y la varianza. El color representa la técnica de buceo utilizada para recolectar la muestra (negro: OC, rojo: R). Se ve que la media aumenta con la varianza de manera lineal si la técnica de buceo es “rebreather” (R) y no lineal si es “open-circuit” (OC).

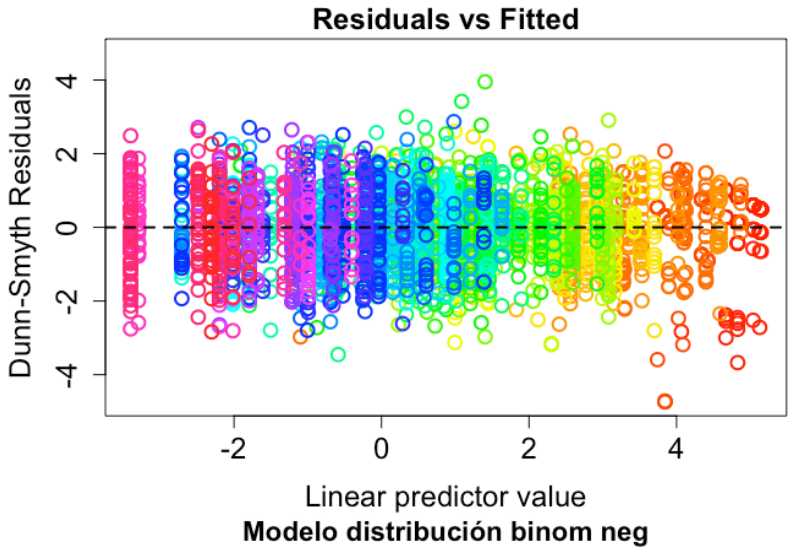


Figura 2. Gráfico de los residuos Dunn-Smyth del modelo ajustado asumiendo distribución binomial negativa contra los predictores lineales, para comprobar las asunciones del modelo. Cada color representa una especie diferente.

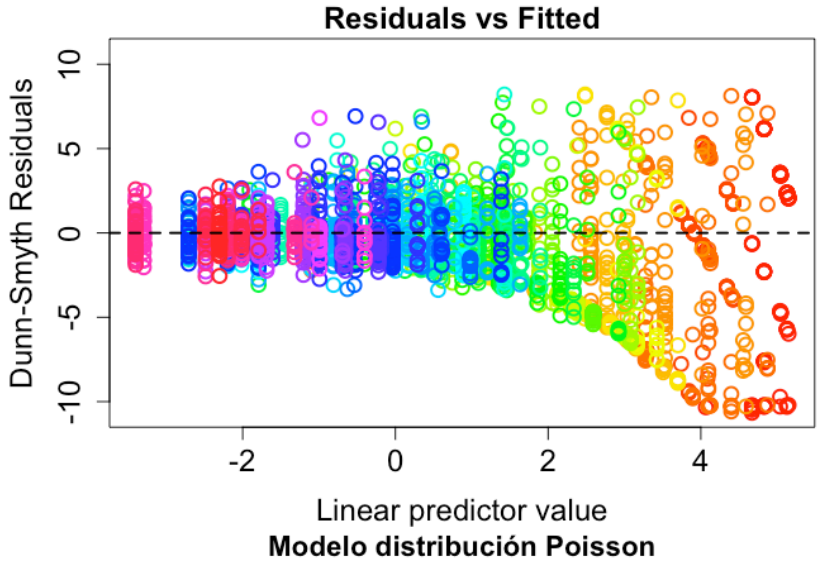


Figura 3. Gráfico de los residuos Dunn-Smyth del modelo ajustado asumiendo distribución binomial negativa contra los predictores lineales, para comprobar las asunciones del modelo. Cada color representa una especie diferente.

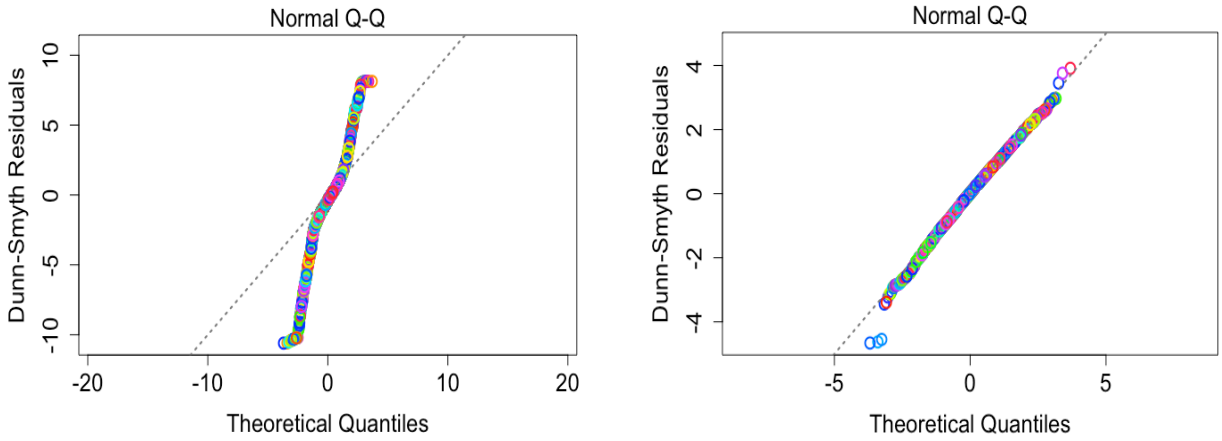


Figura 4. Gráfico de los residuos Dunn-Smyth del modelo contra los cuantiles teóricos, derecha modelo ajustado asumiendo distribución binomial negativa e

izquierda modelo ajustado asumiendo distribución Poisson, para comprobar las asunciones del modelo. Cada color representa una especie diferente.

3.2. Comunidades de peces del mar de Barents

Objetivo

En el segundo ejemplo, incorporamos al modelo variables ambientales con el fin de explorar la naturaleza de la relación entre variables ambientales y variables de abundancia mediante un modelo con variables latentes. El código de R se puede encontrar en el anexo 2 y la tabla de equivalencias de las abreviaciones con el nombre científico de las especies se puede encontrar en el anexo 4.

Datos

Consideramos una base de datos conocida con conteos de peces en el mar Barent. La base de datos Barent Fish proviene del libro de texto “Multivariate Analysis of Ecological Data” escrito por M. Greenacre y R. Primicerio. Los datos fueron muestreados con pesca de arrastre de agosto a septiembre de 2004 a 2009. Se tienen 2 bases de datos, una con los datos de abundancia de peces (*br_fish_ab*) y otra con las variables ambientales (*br_env*).

Las filas son las muestras y las columnas especies de peces. Evitamos transformar las variables respuesta de abundancia de especies, pero sí se transforman las variables ambientales para que estén en la misma escala, y sea más fácil interpretar el modelo. Para la transformación utilizamos la función *scale().*

Se construye el gráfico del logaritmo de la media contra el logaritmo de la varianza de las muestras, para evaluar la distribución a elegir para construir el modelo (fig. 7). El gráfico muestra que la relación entre la media y la varianza en escala logarítmica es lineal lo cual sugiere que la distribución de Poisson o la distribución binomial negativa podría ser adecuada para construir el modelo.

Modelo

Para este caso práctico se utiliza el paquete *boral,* el cual permite ajustar un modelo jerárquico con métodos de ajuste bayesiano (Hui 2016). Las variables latentes en el modelo explican la correlación residual debido a variables no incluidas en el modelo. Las funciones del paquete nos permiten ajustar un modelo puro con variables latentes (ordenación sin restricciones) y un modelo con variables latentes que incorpore covariables ambientales.

El ajuste *boral* usa el programa JAGs para realizar el muestreo por MCMC. Por lo tanto se debe descargar e instalar el programa JAGs por separado de R, antes de instalar *boral*. Se ajustan dos modelos: uno puro LVM y otro LVM con covariables ambientales utilizando la función *boral()*. El modelo construido con *boral* utiliza logaritmo (log) como función vinculo. El parámetro *family* en la función permite escoger la distribución que asumimos para los errores. Ajustaremos LVM con distribución Poisson y con distribución binomial negativa para estudiar cual es la más adecuada para nuestros datos. Elegimos dos variables latentes para el modelo, ya que luego nos servirán de ejes de ordenación en el gráfico de las variables latentes. Incluimos en el modelo el efecto aleatorio de las muestras como un efecto fijo, mediante el parámetro *row.eff = "fixed".* Una vez ajustado, el modelo se puede obtener un resumen con la función *summary().*

Con la función *boral()* también se ajusta un segundo modelo incluyendo covariables ambientales. Este ajusta un GLMs para cada especie con las covariables ambientales disponibles y usa las variables latentes para explicar la correlación residual.

Evaluación del modelo

Para comprobar las asunciones del modelo se analizan los residuos mediante cuatro gráficos creados con la función plot() (fig. 6 ,7 y 8). En las figuras el gráfico de arriba a la izquierda muestra de los residuos Dunn-Smyth contra los predictores lineales y es conveniente que la relación no muestre ningún patrón.

Si esto ocurriera puede indicar que la distribución elegida para el modelo no es la apropiada como observamos claramente en el gráfico de la figura 8 del modelo ajustado con distribución Poisson. La distribución binomial negativa es la que muestra un mejor ajuste del modelo a los datos, ya que la nube de puntos no muestra ningún patron en los datos.

Interpretación

Los gráficos de ordenación con la función *lvsplot()* ponen de manifiesto como se ordenan las muestras de comunidades y las especies. El modelo LVM muestra la ordenación sin restricciones de los datos de abundancia (fig. 9) y el modelo LVM con covariables ambientales muestra la ordenación residual (fig. 10). Estas dos figuras pueden dar una indicación visual de la reducción de la cantidad de covariación después de introducir las variables ambientales. En la figura 9 podemos observar como las muestras se dividen en dos grandes grupos a lo largo del eje de la variable latente 1. La figura 10 no visualiza ninguna agrupación clara de las muestras y la mayoría de especies están en el centro del gráfico. Las especies en la misma dirección y lejos del origen están muy correlacionadas, como en la figura 10 los taxones *Gadus mohua* y Melanogrammus aeglefinus.

La función *corplot()* con las correlaciones del modelo muestra la correlación entre especies. Se construye este tipo de gráfico para correlaciones significativas entre especies consecuencia de la respuesta a las covariables ambientales (fig. 11), el cual muestra correlaciones positivas y negativas entre algunas especies. También se construye un grafico de las correlaciones residuales significativas entre especies (fig. 12) y se ve que la mayoría de las correlaciones entre especies son positivas, es decir que esas especies están correlacionadas de forma directa a variables diferentes a las variables ambientales del modelos, tendiendo a ocurrencia simultanea.

Con la función *get.residual.cor()* podemos obtener las correlaciones residuales del modelo. Al comparar las correlaciones residuales de los modelos LVM puro y LVM con covariables, se ve que éstas se reducen considerablemente en el

segundo modelo (323 y 253.8), lo que implica que las covariables explican un gran porcentaje de la variabilidad entre especies.

Figuras y tablas

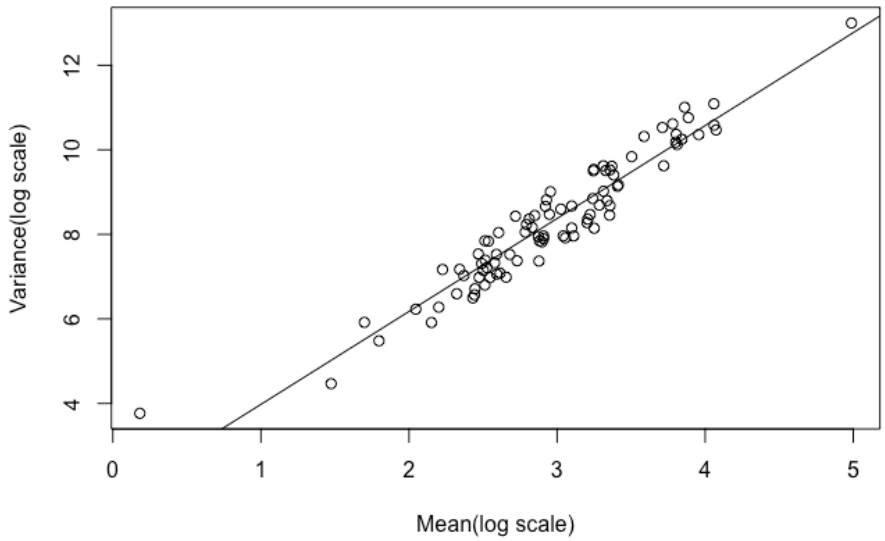


Figura 5. Gráfico de la relación entre la media y la varianza de las muestras en escala logarítmica. Se ve que la varianza aumenta linealmente con la media en escala logarítmica.

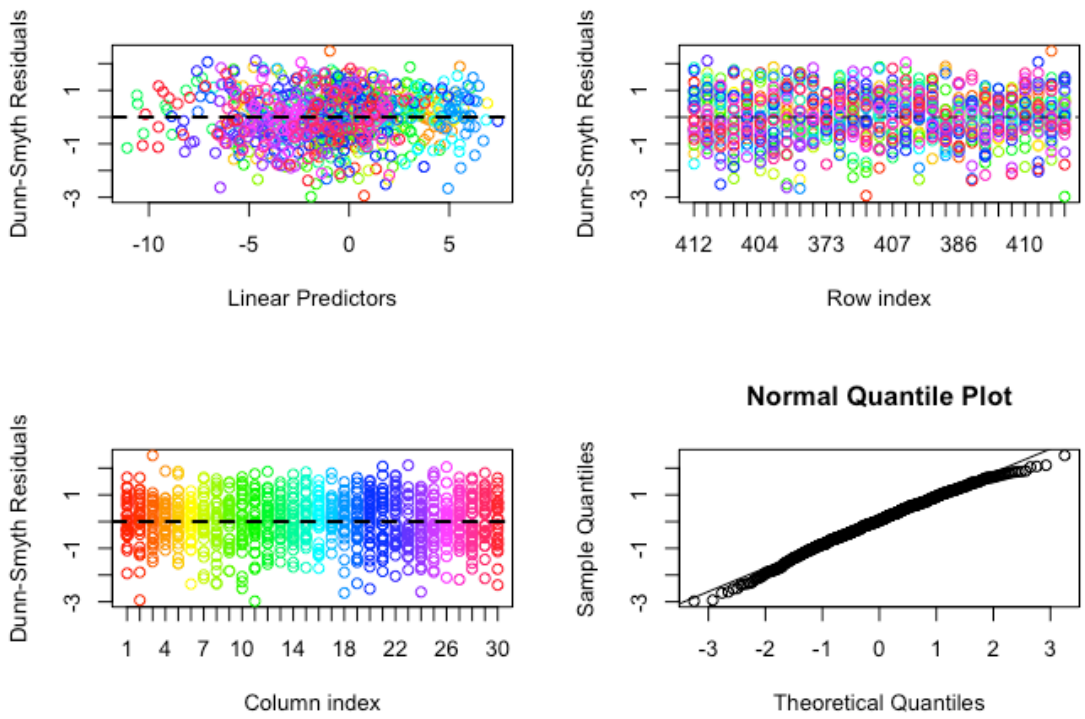


Figura 6. Gráficos de análisis de los residuos Dunn-Smyth del modelo LVM puro ajustado asumiendo distribución binomial negativa, para comprobar las asunciones del modelo. Cada color representa una especie diferente.

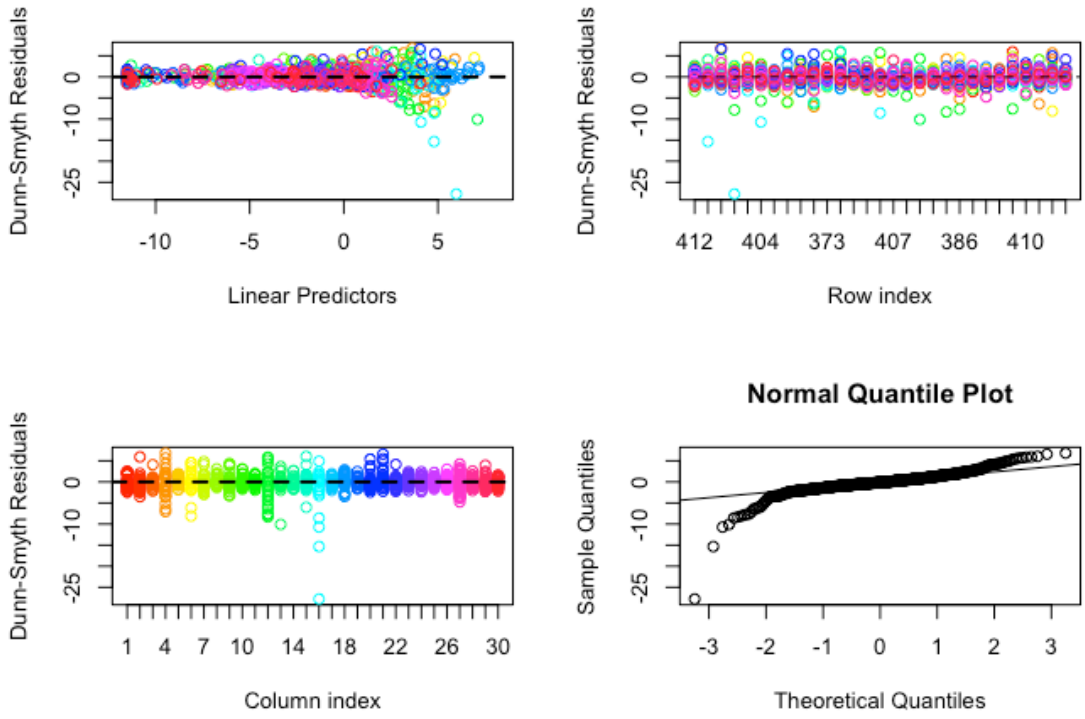


Figura 7. Gráficos de análisis de los residuos Dunn-Smyth del modelo LVM puro ajustado asumiendo distribución Poisson, para comprobar las asunciones del modelo. Cada color representa una especie diferente.

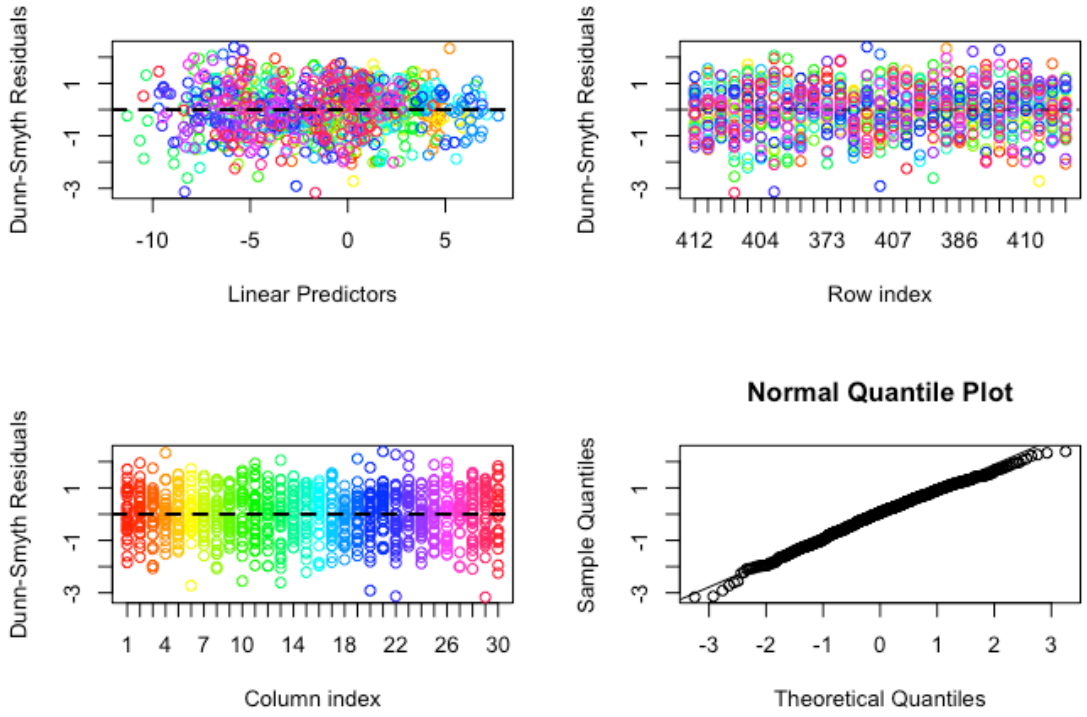


Figura 8. Gráficos de análisis de los residuos Dunn-Smyth del modelo LVM con covariables ambientales ajustado asumiendo distribución binomial negativa, para comprobar las asunciones del modelo. Cada color representa una especie diferente.

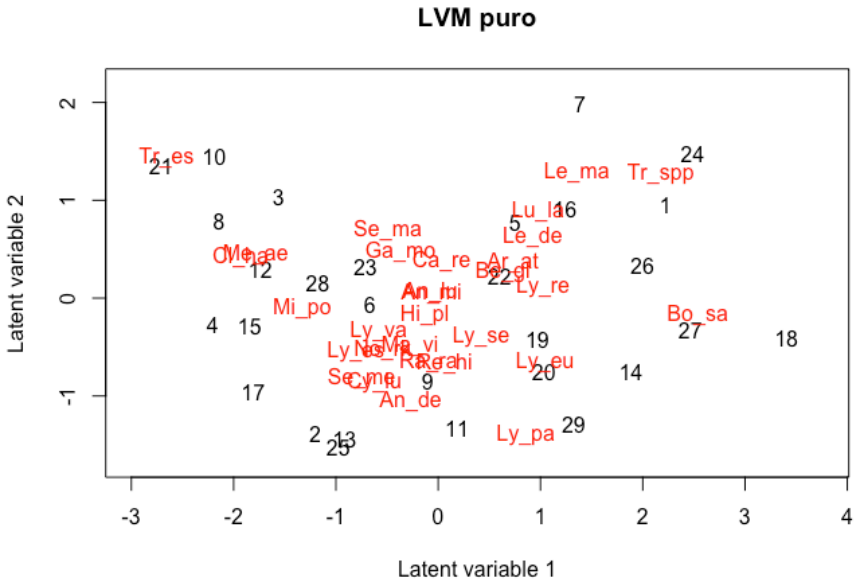


Figura 9. Gráfico de ordenación con las variables latentes del modelo puro LVM de ejes, representando la ordenación sin restricciones. Las muestras se

representan con el número de la fila y las especies, en rojo, se representan con su nombre abreviado.

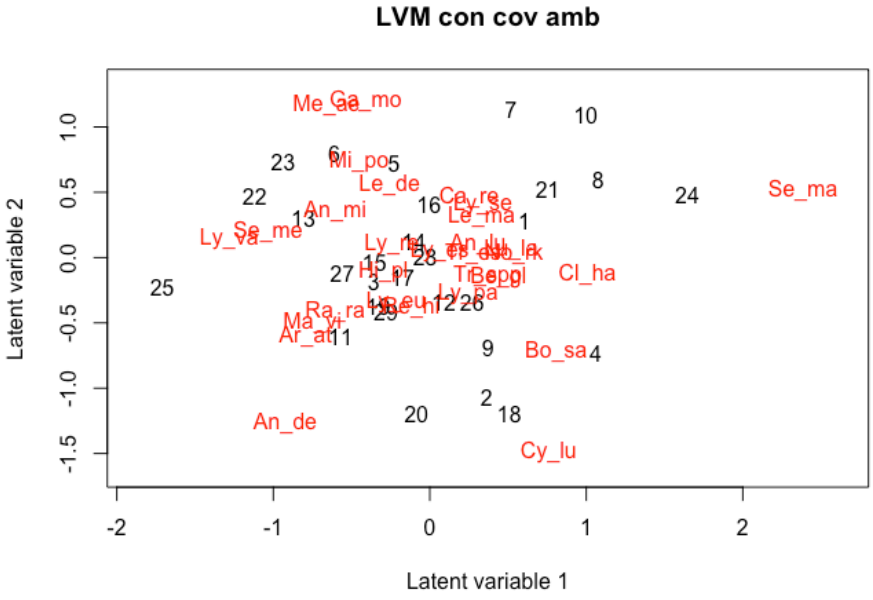


Figura 10. Gráfico de ordenación con las variables latentes del modelo puro LVM de ejes, representando la ordenación residual. Las muestras se representan con el número de la fila y las especies, en rojo, se representan con su nombre abreviado.

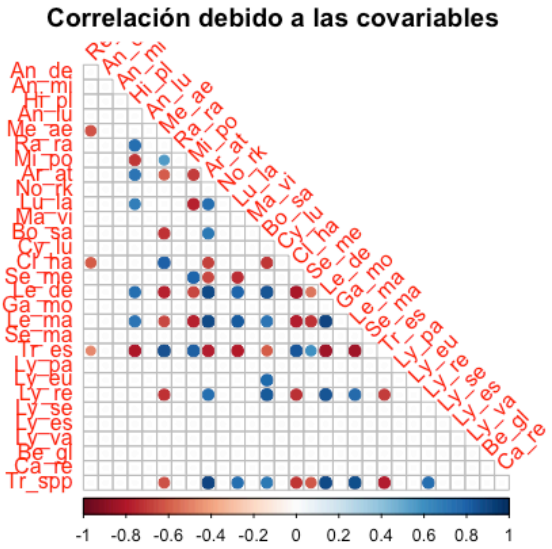


Figura 11. Gráfico de las correlaciones entre especies debido a la respuesta a las variables ambientales. Sólo se representan las correlaciones significativas de los intervalos creíble al 95% excluyendo el cero. El color representa el signo de la correlación (rojo: negativo y azul: positivo) y el tamaño del círculo representa la intensidad de la correlación.

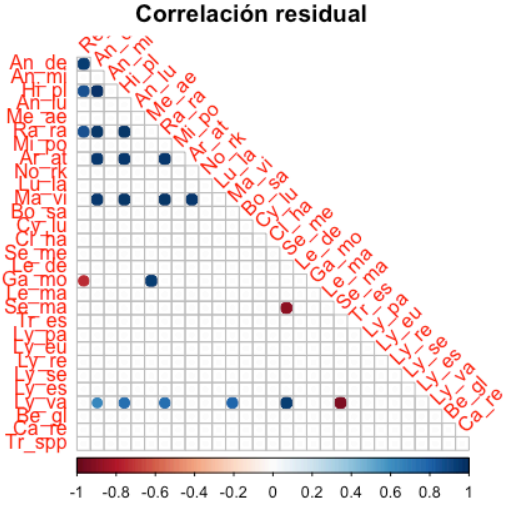


Figura 12. Gráfico de las correlaciones entre especies debido a la correlación residual. Sólo se representan las correlaciones significativas de los intervalos creíbles al 95% excluyendo el cero. El color representa el signo de la

correlación (rojo: negativo y azul: positivo) y el tamaño del círculo representa la intensidad de la correlación.

4. Discusión

En este trabajo se han estudiado metodologías para el análisis de datos de abundancia multivariables basados en modelos. Como habían indicado algunos autores anteriormente los métodos de análisis multivariantes tradicionales utilizados en ecología tienen limitaciones (Hui et al. 2015; Warton et al. 2015b; Lyons et al. 2016). Los métodos de análisis basados en modelos ofrecen muchas ventajas como es la capacidad de seleccionar el modelo, la capacidad de evaluar las asunciones del modelo, la capacidad de realizar inferencia sobre los datos. Como se ve en los casos prácticos el uso de modelos para el análisis de datos de abundancia multivariables es una metodología con mucho potencial para estudios a nivel de comunidad en ecología. Además el uso del programa estadístico R como plataforma para el análisis ha demostrado ser óptimo para el análisis, debido a su acceso libre y sus múltiples posibilidades que han permitido explorar diferentes paquetes para realizar el análisis.

El primer caso práctico ejemplifica el uso de modelos en datos de comunidades para hacer inferencia multivariante sobre el efecto de las variables predictoras. Se utiliza el paquete *mvabund* el cual ofrece funciones eficaces para el análisis de datos de abundancia multivariantes mediante modelos y para la creación de gráficos para visualizar los datos.

El segundo caso práctico ilustra como el uso de LVM ofrece una forma parsimoniosa de estimar correlación entre especies, realizar análisis de ordenación y hacer inferencia sobre los datos, utilizando funciones del paquete *boral*. Los LVM nos permiten inferir sobre las asociaciones entre los datos de abundancia y contabilizar como esa asociación se relaciona con condiciones ambientales. Además más allá del ejemplo ofrecido los LVM también ofrecen la capacidad de incorporar información sobre características de las especies para explicar interacciones entre especies (Hui 2016; Ovaskainen et al. 2017). El tiempo computacional que toma la función boral para ajustar el LVM es moderado ya que el ajuste bayesiano es mas difícil de calcular que el MLE.

Una de las desventajas de estas metodologías es el tiempo computacional, ya que ralentiza mucho el análisis. Además los paquetes para implementarlas son relativamente nuevos y, en algunos casos, están en desarrollo. Ello comporta que la instalación pueda dar problemas y haya incompatibilidades con algunos sistemas operativos y/o versiones.

Una ventaja que tienen las metodologías que usan estos modelos jerárquicos es que tienen potencial para progresar, es decir, es posible su extensión. Por ejemplo, algunos autores están estudiando incorporar al marco de los modelos explicación para observaciones inexactas de abundancias de las especies, ya que los errores de mediciones o de detección de las especies de la comunidad, falsas detecciones o falsos positivos pueden generar problemas a la hora de construir modelos útiles (Guillera-Arroita 2017). También se estudian modelos que combinen varios tipos de datos de abundancia, para poder ser analizados conjuntamente. Otra incorporación interesante a futuros modelos de comunidades es datos genéticos, debido a la importancia de conectar la ecología de comunidades y la biología de la evolución (Johnson and Stinchcombe 2007).

El uso de modelos estadísticos para el análisis de datos de abundancia multivariantes es muy flexible y amplio. En futuros estudios puede ser interesante comparar con los mismos datos la actuación de otros paquetes existentes, por ejemplo *HMSC*.

5. Conclusiones

Hasta hace muy poco el análisis multivariante en ecología para datos de abundancia a nivel de comunidad se limitaba a una mera descripción y presentaba una metodología con múltiples limitaciones.

El uso de modelos para el análisis de datos de abundancia de especies conjuntas es reciente pero ofrece un gran potencial para resolver de forma eficiente infinidad de cuestiones de ecología claves.

Los métodos de análisis de datos de abundancia multivariantes mediante modelos estadísticos se han manifestado como una herramienta fiable y su potencial va a permitir implementarlos en manejo para conservación ecológica.

Los paquetes para ejecutar estas metodologías son herramientas muy potentes pero aún son jóvenes y se están desarrollando. Este desarrollo conllevara a la implantación de nuevos paquetes, o mejoras en los paquetes actuales, cuyo resultado será la obtención de modelos y predicciones más fiables.

6. **Glosario**

Abundancia. Medida referente a una especie presente en una muestra, medida en conteos, biomasa, porcentaje de cobertura, presencia/ausencia.

Análisis multivariable. Conjunto de métodos estadísticos y mate' máticos, destinados a describir e interpretar los datos que provienen de la observación de varias variables estadísticas, estudiadas conjuntamente (Cuadras 2014).

Función vinculo. La función de vínculo en los GLM hace referencia a la transformación de las variables antes de ajustar el modelo.

Intervalo de confianza creíble. Un rango de valores con el 95% de probabilidad para un parámetro (es la versión bayesiana de un intervalo de confianza)

Modelo estadístico. Es una ecuación matemática que reproduce fenomentos que observamos de la forma más exacta posible.

Muestra. Es un subconjuntos de casos

GLM. Modelo Lineal Generalizado (Generalized Lineal Model), modelo de regresión para predecir la respuesta de una variable asumiendo que sigue una distribución de la familia exponencial y asumiendo que una transformación de la respuesta de la media es una función lineal de las variables predictoras.

GLMM. Modelo Lineal Mixto Generalizado (Generalized Lineal Mixed Model), GLM con efectos aleatorios.

HMSC. “Hierarchical Modelling of Species Communities” es un marco de análisis basado en modelos jerárquicos Bayesianos para comunidades de

especies, para explicar condiciones ambientales, rasgos de especies y filogenia.

JAGs. “Just Another Gibbs Sampler” es un programa para análisis Bayesiano de modelos jerárquicos usando MCMC.

LVM. Modelo con variables Latentes (Lantent Variables Model), un modelo de regresión para datos multivariables que incluye algunos predictores latentes no observados o latentes, introducidos para modelar la correlación o para explicar los predictores no especificados por el modelo.

MCMC. “Markov Chain Monte Carlo” es una simulación de distribución de probabilidad.

MLE. “Maximun Likelihood Estimaton “ Máxima Verosimilitud, que es un método de estimación de parámetros de un modelo, en el que se calcula la probabilidad de que ocurra un determinado suceso.

OC. “Open circuit” circuito abierto es la técnica de buceo más popular en el que ninguno de los gases es recirculado y se liberan en forma de burbujas.

R. “Rebreather” circuito cerrado es una técnica de buceo en la que hay una recirculación total del gas que se le suministra al submanirista.

7. **Bibliografía**

Bates D, Maechler Martin, Walker S (2016) Package “lme4.” In: CRAN Repos. https://cran.r-project.org/web/packages/lme4/lme4.pdf

Bolker BM (2007) Ecological Models and Data in R. Ecology 408 . doi: 10.1111/j.1442-9993.2010.02210.x

Borcard D, Gillet F, Legendre, Legendre P (2011) Numerical Ecology with R

Brown AM, Warton DI, Andrew NR, et al (2014) The fourth-corner solution - using predictive models to understand how species traits interact with the environment. Methods Ecol Evol 5:344–352 . doi: 10.1111/2041-210X.12163

Cayuela L, Guillen M, Bolancé C (2016) introducción GLMs- Función vínculo. Univ Val 1–24

Cuadras CM (2014) Nuevos Métodos De Análisis Multivariante. C Ed 304 . doi: 10.1017/CBO9781107415324.004

Guillera-Arroita G (2017) Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. Ecography (Cop) 40: . doi: 10.1111/ecog.02445

Harris DJ (2015) Generating realistic assemblages with a joint species distribution model. Methods Ecol Evol 6:465–473 . doi: 10.1111/2041-210X.12332

Hui FKC (2016) boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in r. Methods Ecol Evol. doi: 10.1111/2041-210X.12514

Hui FKC, Taskinen S, Pledger S, et al (2015) Model-based approaches to unconstrained ordination. Methods Ecol Evol 6:399–411 . doi: 10.1111/2041-210X.12236

Johnson MTJ, Stinchcombe JR (2007) An emerging synthesis between community ecology and evolutionary biology. Trends Ecol Evol 22:250–257 . doi: 10.1016/j.tree.2007.01.014

Kellner KF, Swihart RK (2014) Accounting for imperfect detection in ecology: A quantitative review. PLoS One 9

Klais R, Norros V, Lehtinen S, et al (2017) Community assembly and drivers of phytoplankton functional structure. Funct Ecol 31:760–767 . doi: 10.1111/1365-2435.12784

Lyons MB, Keith DA, Warton DI, et al (2016) Model-based assessment of ecological community classifications. J Veg Sci 27:704–715 . doi: 10.1111/jvs.12400

O'Hara RB, Kotze DJ (2010) Do not log-transform count data. Methods Ecol Evol 1:118–122 . doi: 10.1111/j.2041-210X.2010.00021.x

Ovaskainen O, Tikhonov G, Norberg A, et al (2017) How to make more out of community data? A conceptual framework and its implementation as models and software. Ecol Lett 20:561–576 . doi: 10.1111/ele.12757

Rizopoulos D (2006) An R Package for Latent Variable Modeling and Item Response Theory Analysis. J Stat Softw 17:

Wang Y, Naumann U, Wright ST, Warton DI (2013) Package " mvabund ": Statistical methods for analysing multivariate abundance data. Packag " mvabund " 1–69

Warton DI, Blanchet FG, O'Hara RB, et al (2015a) So Many Variables: Joint Modeling in Community Ecology. Trends Ecol. Evol. 30:766–779

Warton DI, Foster SD, De'ath G, et al (2015b) Model-based thinking for community ecology. Plant Ecol 216:669–682 . doi: 10.1007/s11258-014-0366-3

Warton DI, Wright ST, Wang Y (2012) Distance-based multivariate analyses confound location and dispersion effects. Methods Ecol Evol 3:89–101 . doi: 10.1111/j.2041-210X.2011.00127.x

8. **Anexos**

Anexo 1. Código del análisis de una comunidad de peces a lo largo de la costa Croata

Paso 1- Obtención de los datos

```
# Base de datos de abundancia
hr_fish_ab<-read.table("hrfab.txt")
# Base de datos con los factores
hr_fact<-read.table("hrff.txt")
# Dimensiones de la base de datos
dim(hr_fish_ab)
```

```
[1] 114  48
```

```
hr_fish_ab[1:6,1:10]
```

	Bo_bo <int>	De_de <int>	Di_la <int>	Di_an <int>	Di_pu <int>	Di_sa <int>	Di_vu <int>	Ep_ma <int>
1	0	0	0	1	3	0	38	0
2	0	0	0	0	0	0	32	0
3	0	0	0	0	2	1	26	0
4	0	0	0	0	0	2	58	0
5	0	0	0	4	2	4	229	0
6	0	0	0	0	0	13	60	0

6 rows | 1-9 of 11 columns

```
# Eliminamos las columnas de especies que no se encontraron en ninguna de las muestras
keep<-which(colSums(hr_fish_ab[,1:ncol(hr_fish_ab)])>0);hr_fish_ab<-hr_fish_ab[,keep]
# Especies solo encontradas en una muestra
names(which(colSums(hr_fish_ab>0)<=1))
```

```
[1] "La_vi" "Pa_pa" "Ph_ph" "Sa_sa" "Sc_sf" "Sp_sm" "Sp_sp" "An_an" "Sy_ci"
```

```
sel.spp<-colSums(hr_fish_ab>0)>1 hr_fish_ab<-hr_fish_ab[,sel.spp]
# Nuevas dimensiones de la base de datos
dim(hr_fish_ab)
```

```
[1] 114  39
```

```
# Gráfico  de relación entre la media y la varianza de los datos
```

```
plot(x<-apply(hr_fish_ab,1, mean),y<-apply(hr_fish_ab,1, var), ylab = "Varianza de las muestras", xlab="Media de las muestras",      col=hr_fact$DS_factor)
abline(lm(y~x))
```

Paso 2 -Construcción del modelo

```
library(mvabund)

# Ajuste del modelo hr_abundance<-mvabund(hr_fish_ab) hr_modelo<-manyglm(hr_abundance~hr_fact$Location_factor*hr_fact$DS_factor, family="negative.binomial")

summary(hr_modelo)
```

```
 Test statistics:                                          wald value Pr(>wald)
(Intercept)                                               33.594  0.000999
hr_fact$Location_factorKamenjak                           7.187  0.000999
hr_fact$Location_factorSilba                              6.990  0.000999
hr_fact$Location_factorTelascica                          6.603  0.000999
hr_fact$DS_factorR                                        5.381  0.039960
hr_fact$Location_factorKamenjak:hr_fact$DS_factorR        5.341  0.006993
hr_fact$Location_factorSilba:hr_fact$DS_factorR           5.549  0.016983
hr_fact$Location_factorTelascica:hr_fact$DS_factorR       2.712  0.810190

(Intercept)                                          ***

hr_fact$Location_factorKamenjak                      ***

hr_fact$Location_factorSilba                         ***

hr_fact$Location_factorTelascica                     ***

hr_fact$DS_factorR                                   *

hr_fact$Location_factorKamenjak:hr_fact$DS_factorR   **

hr_fact$Location_factorSilba:hr_fact$DS_factorR      *

hr_fact$Location_factorTelascica:hr_fact$DS_factorR

---  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Test statistic:  23.63, p-value: 0.000999  Arguments:  Test statistics calculated assuming response assumed to be uncorrelated   P-value calculated using 1000 resampling iterations via pit.trap resampling (to account for correlation in testing).
```

Paso 3- Evaluación del modelo

```
# Ajuste del modelo con distribución Poisson

hr_modeloP<-manyglm(hr_abundance~hr_fact$Location_factor*hr_fact$DS_factor, family="poisson")

# Gráficos de análisis de residuos de modelo

plot.manyglm(hr_modelo, sub="Modelo distribución binom neg")
```

```
plot.manyglm(hr_modeloP, sub="Modelo distribución Poisson")

# gráfico Q-Q

plot.manyglm(hr_modelo, which = 2, sub="Modelo distribución binom neg")

# gráfico Q-Q modelo con distribución Poisson plot.manyglm(hr_modeloP, which = 2, sub="Modelo distribución binom neg")
```

Paso 4- Interpretación

```
# ANOVA

hr_anova<-anova.manyglm(hr_modelo, p.uni="adjusted")
```

```
Time elapsed: 0 hr 3 min 42 sec
```

```
hr_anova
```

```
Analysis of Deviance Table  Model: manyglm(formula = hr_abundance ~
hr_fact$Location_factor * hr_fact$DS_factor,  Model:     family = "negative.binomial")
Multivariate test:                                Res.Df Df.diff   Dev Pr(>Dev)

(Intercept)                                          113
hr_fact$Location_factor                              110       3 451.5    0.001
hr_fact$DS_factor                                    109       1 160.3    0.001
hr_fact$Location_factor:hr_fact$DS_factor            106       3 136.8    0.004
(Intercept)

hr_fact$Location_factor                         ***

hr_fact$DS_factor                               ***

hr_fact$Location_factor:hr_fact$DS_factor       **

--- Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Univariate Tests:                                        Bo_bo          De_de

                                        Dev Pr(>Dev)   Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                    4.27    0.859 33.59    0.001
hr_fact$DS_factor                         0.341    1.000 11.72    0.029
hr_fact$Location_factor:hr_fact$DS_factor 2.021    0.997 8.407    0.675

                                                 Di_la          Di_an

                                       Dev Pr(>Dev)       Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                       7.212    0.574 6.093    0.665
hr_fact$DS_factor                             6.946    0.245 6.503    0.268
hr_fact$Location_factor:hr_fact$DS_factor     0.001    1.000 7.788    0.676

                                                 Di_pu          Di_sa

                                        Dev Pr(>Dev)    Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                    10.506    0.237 39.141    0.001
hr_fact$DS_factor                           2.164    0.950    5.3    0.430
hr_fact$Location_factor:hr_fact$DS_factor   0.383    1.000 12.363    0.216

                                                 Di_vu          Ep_ma

                                        Dev Pr(>Dev)    Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                    10.815    0.237 10.042    0.264
hr_fact$DS_factor                          27.886    0.001  0.244    1.000
hr_fact$Location_factor:hr_fact$DS_factor    5.43    0.903  1.405    1.000

                                                 La_me             Mu

                                        Dev Pr(>Dev)   Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                    28.875    0.001 12.07    0.172
hr_fact$DS_factor                           0.304    1.000 1.182    0.996
hr_fact$Location_factor:hr_fact$DS_factor   9.292    0.554 2.626    0.990

                                                 Mu_su          Ob_me

                                        Dev Pr(>Dev)    Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                    16.735    0.026  7.374    0.557
hr_fact$DS_factor                           0.017    1.000  1.599    0.978
hr_fact$Location_factor:hr_fact$DS_factor   3.703    0.965 12.787    0.184

                                                 Pa_er          Sa_sal

                                        Dev Pr(>Dev)    Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                    6.669    0.604  1.038    0.997
hr_fact$DS_factor                          0.031    1.000  5.694    0.368
hr_fact$Location_factor:hr_fact$DS_factor 1.161    1.000 11.726    0.263
```

```
                                          Sc_um          Sc_po

                                    Dev Pr(>Dev)   Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                  11.81    0.182 1.255    0.997
hr_fact$DS_factor                        7.201    0.228 0.211    1.000
hr_fact$Location_factor:hr_fact$DS_factor 2.108    0.997 4.281    0.945

                                          Se_du           Se_ca

                                    Dev Pr(>Dev)    Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                  7.236    0.574 11.514    0.184
hr_fact$DS_factor                        3.623    0.790  0.103    1.000
hr_fact$Location_factor:hr_fact$DS_factor 1.094    1.000  1.392    1.000

                                          Se_sc          Sp_cr

                                    Dev Pr(>Dev)   Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                  11.763    0.184 8.961    0.353
hr_fact$DS_factor                        14.674    0.008 5.433    0.417
hr_fact$Location_factor:hr_fact$DS_factor  8.008    0.675 0.001    1.000

                                          Sp_au          Sp_ma

                                    Dev Pr(>Dev)   Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                  0.296    0.997 4.242    0.859
hr_fact$DS_factor                          6.8    0.249 0.153    1.000
hr_fact$Location_factor:hr_fact$DS_factor 8.352    0.675 5.454    0.903

                                          Sp_ca          Po_sa

                                    Dev Pr(>Dev)   Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                  18.611    0.026 1.826    0.989
hr_fact$DS_factor                         1.392    0.986 3.015    0.876
hr_fact$Location_factor:hr_fact$DS_factor   4.85    0.925 0.001    1.000

                                          Ap_im             At

                                    Dev Pr(>Dev)   Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                  11.471    0.185 4.286    0.859
hr_fact$DS_factor                         3.075    0.876 0.386    1.000
hr_fact$Location_factor:hr_fact$DS_factor  0.421    1.000     0    1.000

                                          Ch_ch           Co_ju

                                    Dev Pr(>Dev)    Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                  2.477    0.961 17.627    0.026
hr_fact$DS_factor                        2.431    0.941  4.311    0.658
hr_fact$Location_factor:hr_fact$DS_factor 1.351    1.000  0.715    1.000

                                          Mu_he          Sc_no

                                    Dev Pr(>Dev)   Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                  3.422    0.912  2.82    0.953
hr_fact$DS_factor                        2.336    0.948 2.653    0.921
hr_fact$Location_factor:hr_fact$DS_factor 0.554    1.000 0.001    1.000

                                          Se_he          Sy_do

                                    Dev Pr(>Dev)   Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                  9.552    0.292 6.208    0.665
hr_fact$DS_factor                        0.154    1.000 3.803    0.777
hr_fact$Location_factor:hr_fact$DS_factor  1.73    1.000 2.425    0.993

                                          Sy_med          Sy_mel

                                    Dev Pr(>Dev)    Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                  8.465    0.422 10.292    0.240
hr_fact$DS_factor                        9.756    0.072  0.594    1.000
hr_fact$Location_factor:hr_fact$DS_factor  1.068    1.000  4.567    0.934

                                          Sy_oc          Sy_roi

                                    Dev Pr(>Dev)    Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                  17.875    0.026  4.381    0.859
hr_fact$DS_factor                         0.053    1.000  1.086    0.996
hr_fact$Location_factor:hr_fact$DS_factor  5.474    0.903  0.367    1.000

                                          Sy_ros           Sy_ti
```

```
                                        Dev Pr(>Dev)    Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                  10.796    0.237 31.972    0.001
hr_fact$DS_factor                         7.452    0.220  8.243    0.156
hr_fact$Location_factor:hr_fact$DS_factor  1.724    1.000  1.682    1.000

                                             Th_pa

                                        Dev Pr(>Dev) (Intercept)
hr_fact$Location_factor                  37.881    0.001

hr_fact$DS_factor                         1.383    0.987
hr_fact$Location_factor:hr_fact$DS_factor   0.07    1.000

Arguments:  Test statistics calculated assuming uncorrelated response (for faster
computation)  P-value calculated using 999 resampling iterations via PIT-trap resampling
(to account for correlation in testing.
```

```
# Diferencias significativas según localidades
which(hr_anova$uni.p[2,]<0.05)
```

```
De_de Di_sa La_me Mu_su Sp_ca Co_ju Sy_oc Sy_ti Th_pa

  2     6     9    11    23    28    35    38    39
```

```
# Diferencias significativas según la técnica de buceo
which(hr_anova$uni.p[3,]<0.05)
```

```
De_de Di_vu Se_sc

   2     7    19
```

```
# ANOVA a modelos anidados

hr_modeloR<-manyglm(hr_abundance~hr_fact$DS_factor, family="negative.binomial")
anova(hr_modelo,hr_modeloR)
```

```
Time elapsed: 0 hr 1 min 13 sec

Analysis of Deviance Table  hr_modeloR: hr_abundance ~ hr_fact$DS_factor hr_modelo:
hr_abundance ~ hr_fact$Location_factor * hr_fact$DS_factor

Multivariate test:     Res.Df Df.diff   Dev Pr(>Dev)

          hr_modeloR     112

          hr_modelo      106        6 618.9    0.001 ***

 --- Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Arguments:  Test
statistics calculated assuming uncorrelated response (for faster computation)

P-value calculated using 999 resampling iterations via PIT-trap resampling (to account
for correlation in testing).
```

Anexo 2. Código del análisis de una comunidad de peces a lo largo de la costa Croata

Paso 1- Obtención de los datos

```
br_fish_ab<-read.table("brfab.txt") br_env<-read.table("bre.txt") br_trait<-
read.table("brt.txt")

dim(br_fish_ab)
```

```
[1] 89 30
```

```
br_fish_ab[1:6,1:10]
```

	Re_hi <int>	An_de <int>	An_mi <int>	Hi_pl <dbl>	An_lu <int>	Me_ae <dbl>	Ra_ra <dbl>	Mi_po <int>	Ar_at <int>
356	0	0	0	31	0	108	0	325	0
357	0	0	0	4	0	110	0	349	0
358	0	0	0	27	0	788	0	6	0
359	0	0	1	13	0	295	0	2	0
363	0	0	0	23	0	13	2	240	0
364	1	0	0	20	0	97	0	0	0

6 rows | 1-10 of 11 columns

```
br_env<-scale(br_env)

# Relación entre la media y la varianza de los datos

plot(x<-log(apply(br_fish_ab,1, mean)),y<-log(apply(br_fish_ab,1, var)),  ylab =
"Variance(log scale)", xlab="Mean(log scale)") abline(lm(y~x)
```

Paso 2- Construcción del modelo

```
library(boral)
```

```
Loading required package: coda

If you recently updated boral, please check news(package = "boral") for the updates in
the latest version.
```

```
# Seleccion de una muestra aleatoria de nuestros datos

 y<-br_fish_ab set.seed(989) p<-1/3 muestra<-sample(nrow(y), floor(nrow(y)*p)) y<-
y[muestra,]

# Ajuste del LVM puro distribucion binomial negativa

br_modelo1<-boral(y=y, family = "negative.binomial", num.lv = 2, row.eff =
"fixed", save.model = TRUE)
```

```
row.ids assumed to be a matrix with one column and elements 1,2,...nrow(y) i.e., a row-
specific intercept.

module glm loaded

Compiling model graph    Resolving undeclared variables    Allocating nodes Graph
information:    Observed stochastic nodes: 870    Unobserved stochastic nodes: 1076
Total graph size: 6422  Initializing model

Calculating Information criteria...
```

```
summary(br_modelo1)
```

```
$call boral.default(y = y, family = "negative.binomial", num.lv = 2,      row.eff =
"fixed", save.model = TRUE)

 $coefficients

coefficients

cols      beta0 theta1 theta2 Dispersion

Re_hi    0.902  1.507  0.000      0.753

An_de    0.434  1.792  0.952      0.958

An_mi   -0.264  0.268 -0.412      1.205

Hi_pl    4.605  0.550 -0.091      0.127

An_lu   -4.430  0.251 -0.392     18.549
```

```
Me_ae   2.077 -1.918  2.467      1.113
Ra_ra   1.400  1.347  0.282      0.432
Mi_po  -1.614 -0.650  2.102      3.282
Ar_at   1.473  0.527 -2.178      1.065
No_rk  -0.654  0.726  1.004     13.266
Lu_la  -1.411 -0.078 -3.113      4.041
Ma_vi   3.311  0.917  0.471      1.076
Bo_sa  -2.270  3.020 -5.091      6.200
Cy_lu  -1.236  1.165  1.454     13.972
Cl_ha  -1.860 -2.017  2.763     14.001
Se_me   4.453  0.993  1.649      4.076
Le_de  -0.237  0.280 -2.786      0.252
Ga_mo   5.934 -0.670 -0.220      1.138
Le_ma  -1.973 -0.342 -4.178      0.513
Se_ma   0.553 -1.134 -0.187     10.685
Tr_es  -1.582 -4.260  3.215      7.035
Ly_pa  -2.213  3.362 -0.851     11.614
Ly_eu  -2.477  2.385 -1.867      5.122
Ly_re  -1.856  1.166 -2.517      2.195
Ly_se  -2.973  1.418 -0.916     18.200
Ly_es  -3.440  0.522  1.515     17.251
Ly_va   1.508  0.405  0.911      1.477
Be_gl  -2.760  0.588 -1.907     19.195
Ca_re   0.048 -0.147 -0.889      4.355
Tr_spp -2.991  0.434 -5.708      1.798
$row.coefficients
$row.coefficients[[1]]
     1      2      3      4      5      6      7      8      9     10
 0.026 -0.893  0.208 -0.740 -0.185 -0.214 -0.484 -1.123 -0.720 -0.767
 11     12     13     14     15     16     17     18     19     20
-0.313 -0.223 -0.939 -0.531 -0.351 -0.513 -0.702 -0.020 -0.327  0.655
    21     22     23     24     25     26     27     28     29
 -0.425  0.336  0.316 -1.046  0.395  0.089 -0.478 -0.297 -0.556
$est [1]
 "median"
$calc.ics [1]
 TRUE
```

```
$trial.size
[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
$num.ord.levels [1] 0
$prior.control
$prior.control$ssvs.index [1] -1
```

```
$ics                          Conditional DIC

                                      2334.708

                                       WAIC

                                       2987.425

                                      EAIC

                                      3257.836

                                      EBIC

                                      4240.145

                               AIC at post. median

                                      3191.714

                              BIC at post. median

                                      3897.451

                              Marginal log-likelihood at post. median

                                      -1447.857


attr(,"class") [1] "summary.boral"
```

```
# Ajuste del LVM puro distribucion Poisson

br_modeloP<-boral(y=y, family = "poisson", num.lv = 2, row.eff = "fixed", save.model
= TRUE)
```

```
row.ids assumed to be a matrix with one column and elements 1,2,...nrow(y) i.e., a row-
specific intercept.

Compiling model graph    Resolving undeclared variables    Allocating nodes Graph
information:    Observed stochastic nodes: 870    Unobserved stochastic nodes: 176
Total graph size: 4621  Initializing model

Calculating Information criteria...
```

```
# Ajuste de LVM con covariables ambientales

X<-br_env[muestra,]

br_modelo2<-boral(y=y, X=X, family = "negative.binomial", num.lv = 2, save.model =
TRUE)
```

```
Compiling model graph    Resolving undeclared variables    Allocating nodes Graph
information:    Observed stochastic nodes: 870    Unobserved stochastic nodes: 1167
Total graph size: 7528  Initializing model

Calculating Information criteria...
```

```
summary(br_modelo2)
```

```
$call boral.default(y = y, X = X, family = "negative.binomial", num.lv = 2,
save.model = TRUE)

$coefficients

 coefficients

cols      beta0 theta1 theta2 Dispersion

Re_hi    0.206  0.978  0.000      0.620

An_de    0.016  2.578  0.860      1.167

An_mi   -1.074  0.112  0.996      2.446

Hi_pl    4.238  0.658  0.365      0.226

An_lu   -4.508  0.109 -0.451     21.541

Me_ae    1.789 -1.037  1.380      1.045
```

```
Ra_ra   0.662  1.217  0.708      0.315
Mi_po  -2.197 -0.492  0.912      2.909
Ar_at   0.315  1.562  0.918      0.871
No_rk  -2.165  0.121 -0.817      6.053
Lu_la  -3.241  0.090 -0.781      6.162
Ma_vi   3.036  1.393  0.891      1.199
Bo_sa  -3.390  1.090 -1.497      6.773
Cy_lu  -2.270  2.223 -1.715     12.120
Cl_ha  -3.237  0.165 -1.573     10.772
Se_me   3.573  0.510  1.571      2.079
Le_de  -1.414 -0.324  0.559      0.426
Ga_mo   4.898 -1.181  1.021      0.735
Le_ma  -3.576 -0.198 -0.405      1.228
Se_ma  -0.725 -1.316 -3.370      5.200
Tr_es  -1.996  0.220 -0.429      5.193
Ly_pa  -3.510  0.688 -0.500     11.305
Ly_eu  -2.983  0.957  0.158      9.169
Ly_re  -2.932  0.328  0.366      3.842
Ly_se  -3.768 -0.344 -0.382     20.284
Ly_es  -3.848  0.276 -0.107     19.360
Ly_va   0.831  0.682  1.919      0.830
Be_gl  -3.275  0.425 -0.738     21.818
Ca_re   0.020 -0.393 -0.229      6.206
Tr_spp -2.911  0.426 -0.630      4.264
$X.coefficients
 X
cols     Latitude Longitude  Depth Temperature
Re_hi     0.359     1.366  1.512      -0.066
An_de     0.542     0.429  0.527       0.869
An_mi     0.331     1.314  0.581      -0.036
Hi_pl     0.069     0.098  0.003      -0.266
An_lu     0.430    -0.848  0.727       0.033
Me_ae     0.126     0.613 -1.136       2.589
Ra_ra     0.091     0.904  0.457      -0.188
Mi_po     0.347    -1.026  0.871       2.424
Ar_at     1.380    -1.141 -2.060      -1.270
No_rk    -0.711     0.662  3.342      -0.689
Lu_la     1.768     1.485 -1.019      -1.401
Ma_vi     0.105    -0.078 -0.055       0.008
Bo_sa     5.107    -0.743 -0.034      -0.210
Cy_lu     0.943    -0.489 -0.875       2.101
Cl_ha    -1.399    -0.978 -0.974       2.250
```

```
Se_me    -0.434    -0.799  1.396       0.476
Le_de     1.992    -0.377 -0.816      -1.183
Ga_mo     0.419     1.174 -0.817       0.786
Le_ma     2.287     0.135 -1.827      -1.515
Se_ma     0.979     0.428  1.488       0.862
Tr_es     0.166     0.290  0.262       5.277
Ly_pa     2.126    -0.249  2.462      -0.484
Ly_eu     2.547    -1.311  0.461      -0.596
Ly_re     3.095    -0.278 -0.351      -0.470
Ly_se     1.093    -0.345  1.864      -0.528
Ly_es     0.427    -1.936  0.636       1.250
Ly_va    -0.543    -0.651 -0.202       0.080
Be_gl     2.258    -0.018 -0.077       0.047
Ca_re     0.584     0.026  0.278      -0.414
Tr_spp    2.393    -1.181 -2.402      -1.721
$est [1]
 "median"
 $calc.ics [1]
TRUE
$trial.size
[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 $num.ord.levels [
1] 0
$prior.control
$prior.control$ssvs.index
[1] -1 -1 -1 -1
$ics                           Conditional DIC
                                2265.955
                               WAIC
                               2990.246
                               EAIC
                               3405.416
                                EBIC
                                4821.659
                      AIC at post. median
                               3253.045
                     BIC at post. median
                               4392.714
Marginal log-likelihood at post. median
                               -1387.522
attr(,"class") [1] "summary.boral"
```

Paso 3- Evaluación del modelo

```
# Gráficos de análisis de residuos del LVM puro distribución binomial negativa

par(mfrow=c(2,2))

plot(br_modelo1, ask=FALSE)

# Gráficos de análisis de residuos del LVM puro distribución Poisson
par(mfrow=c(2,2))

 plot(br_modeloP, ask=FALSE)

# LVM con covariables ambientales

par(mfrow=c(2,2))

plot(br_modelo2, ask=FALSE)
```

Paso 4- Interpretación

```
# Gráficos de dos dimensiones de las variables latentes LVM puro
lvsplot(br_modelo1 , alpha=0.55, main="LVM puro")
```

```
All latent variable coefficients included in biplot.
```

```
# Gráficos de dos dimensiones de las variables latentes LVM con covariables ambientales

lvsplot(br_modelo2, main="LVM con cov amb")
```

```
All latent variable coefficients included in biplot.
```

```
# Correlaciones entre las especies debido a las similaridades ambientales

envcors<-get.enviro.cor(br_modelo2)

# Correlacion residual explicadas por las variables latentes

rescors1<-get.residual.cor(br_modelo1);rescors1$trace
```

```
[1] 340.9071
```

```
rescors2<-get.residual.cor(br_modelo2);rescors2$trace
```

```
[1] 253.8304
```

```
# Gráfico con las correlaciones significativas en los intervalos al 95%

library(corrplot)
```

```
corrplot 0.84 loaded
```

```
corrplot(envcors$sig.cor, type="lower", diag=F, title="Correlación debido a las covariables", mar=c(3,0.5, 2,1), tl.srt=45)

corrplot(rescors2$sig.cor, type="lower", diag=F, title="Correlación residual",mar=c(3,0.5, 2,1), tl.srt=45)
```

Anexo 3. Tabla de equivalencia de las abreviaciones al nombre científico de la especie para los peces de *Hr fish*.

Abreviación	Nombre científico
Bo_bo	*Boops boops*
De_de	*Dentex dentex*
Di_la	*Dicentrarchus labrax*
Di_an	*Diplodus annularis*
Di_pu	*Diplodus puntazzo*
Di_sa	*Diplodus sargus*
Di_vu	*Diplodus vulgaris*
Ep_ca	*Epinephelus caninus*
Ep_co	*Epinephelus costae*
Ep_ma	*Epinephelus marginatus*
La_me	*Labrus merula*
La_mi	*Labrus mixtus*
La_vi	*Labrus viridis*
Li_mo	*Lithognathus mormirus*
Mu	*Mugilidae*
Mu_su	*Mullus surmuletus*
Ob_me	*Oblada melanura*
Pa_er	*Pagellus erythrinus*
Ph_ph	*Phycis phycis*
Sa_sa	*Sarda sarda*
Sa_sal	*Sarpa salpa*
Sc_um	*Sciaena umbra*
Sc_sc	*Scomber scombrus*
Sc_po	*Scorpaena porcus*
Sc_sf	*Scorpaena scrofa*
Se_du	*Seriola dumerili*
Se_ca	*Serranus cabrilla*
Se_sc	*Serranus scriba*
Sp_cr	*Sparisoma cretense*
Sp_au	*Sparus aurata*
Sp_ma	*Spicara maena*
Sp_sm	*Spicara smaris*
Sp_ca	*Spondyliosoma cantharus*
Sp_sp	*Sphyraena sphyraena*
An_an	*Anthias anthias*
Ap_im	*Apogon imberbis*
At	*Atherina spp.*

Ch_ch	*Chromis chromis*
Co_co	*Conger conger*
Co_ju	*Coris julis*
Mu_he	*Muraena helena*
Sc_no	*Scorpaena notata*
Se_he	*Serranus hepatus*
Sy_ci	*Symphodus cinereus*
Sy_do	*Symphodus doderleini*
Sy_med	*Symphodus mediterraneus*
Sy_mel	*Symphodus melanocercus*
Sy_me	*Symphodus melops*
Sy_oc	*Symphodus ocellatus*
Sy_roi	*Symphodus roissali*
Sy_ros	*Symphodus rostratus*
Sy_ti	*Symphodus tinca*
Sy_sa	*Synodus saurus*
Th_pa	*Thalassoma pavo*
Th_th	*Thunnus thynnus*
Ps_de	*Pseudocarnax dentex*
My_aq	*Myliobatis aquila*

Anexo 4. Tabla de equivalencia de las abreviaciones a el nombre científico de la especie para los peces de *BarentFish*.

Abreviación	Nombre científico
An_de	Anarhichas denticulatus
An_lu	Anarhichas lupus
An_mi	Anarhichas minor
Ar_at	Artediellus atlanticus
Be_gl	Benthosema glaciale
Bo_sa	Boreogadus saida
Ca_re	Careproctus sp.
Cl_ha	Clupea harengus
Cy_lu	Cyclopterus lumpus
Ga_mo	Gadus morhua
Hi_pl	Hippoglossoides platessoides
Le_de	Leptagonus decagonus
Le_ma	Leptoclinus maculatus
Lu_la	Lumpenus lampretaeformis
Ly_es	Lycodes esmarkii
Ly_eu	Lycodes eudipleurostictus

Ly_pa	Lycodes pallidus
Ly_re	Lycodes reticulatus
Ly_se	Lycodes seminudus
Ly_va	Lycodonus flagellicauda
Ma_vi	Mallotus villosus
Me_ae	Melanogrammus aeglefinus
Mi_po	Micromesistius poutassou
No_rk	Molva molva
Ra_ra	Rajella fyllae
Re_hi	Reinhardtius hippoglossoides
Se_ma	Sebastes mentella
Se_me	Sebastes norvegicus
Tr_es	Triglops murrayi
Tr_spp	Trisopterus esmarkii