

Comparación entre el análisis canónico de correspondencias  
y el análisis factorial múltiple  
en tablas de frecuencias-variables continuas

Tesis para optar al título de Maestría en Ciencias Estadística

Universidad Nacional de Colombia

Melba Liliana Vertel Morinsón <sup>1</sup>  
Director: Campo Elías Pardo <sup>2</sup>

29 de enero de 2010

<sup>1</sup>Estudiante Magister en estadística, Universidad Nacional

<sup>2</sup>Profesor Asociado, Universidad Nacional sede Bogotá

Departamento de Estadística  
Facultad de Ciencias  
Universidad Nacional de Colombia

Jurado uno: \_\_\_\_\_

Jurado dos: \_\_\_\_\_

Jurado tres: \_\_\_\_\_

Director: \_\_\_\_\_

# Dedicatoria

A Dios

A mis hijos: Sebastián David y Jesús Manuel

A mi compañero: Jesús Antonio

A mis padres: Melba y Manuel

A mis familiares y amigos

# Agradecimientos

Al Director del proyecto profesor Campo Elías Pardo, por su constante apoyo durante este trabajo.

A los profesores: Liliana Blanco, Hector Mora, Fabio Nieto, Luis Alberto López, Luis Guillermo Díaz, Alberto Vargas, y María Nelcy Rodríguez; por los conocimientos adquiridos en la Maestría, así como también sus consejos y enseñanzas.

A la Universidad Nacional de Colombia y la Universidad del Magdalena, por brindarme esta oportunidad.

A la Universidad de Sucre por brindarme la comisión de estudios.

Finalmente deseo agradecer a mis compañeros: Mario, Ketty, Roberto, Carlos y Victor, por su constante apoyo.

Sincelejo, Sucre  
Diciembre, 2009.

Melba Liliana Vertel Morinsón

# Índice general

<b>Resumen</b>	<b>1</b>
<b>Introducción</b>	<b>2</b>
<b>1. Elementos básicos</b>	<b>4</b>
1.1. Notación . . . . .	4
1.2. Ejemplo . . . . .	5
1.3. $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$ . . . . .	6
1.4. ACS de $\mathbf{T}$ . . . . .	8
1.5. $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$ . . . . .	9
<b>2. ACC de <math>(T, Z)</math></b>	<b>10</b>
2.1. ACC de $(T, Z)$ . . . . .	11
2.2. Gráficas y ayudas para la interpretación . . . . .	11
2.2.1. Biplot . . . . .	11
2.2.2. Circulo de correlaciones . . . . .	11
2.2.3. Gráfico triplot . . . . .	12
2.2.4. Prueba de permutación Monte Carlo . . . . .	12
2.3. Análisis del ejemplo <i>Gorgona</i> con ACC . . . . .	12
<b>3. AFM de <math>(\mathbf{T}, \mathbf{Z})</math></b>	<b>16</b>
3.1. Inercia y valores propios . . . . .	17
3.2. Grupos de variables . . . . .	17
3.3. Gráficas y ayudas a la interpretación . . . . .	17
3.3.1. Gráficas y ayudas a la interpretación de <i>individuos y variables</i> . . . . .	17
3.3.2. Gráfica y ayudas a la interpretación para los grupos de variables . . . . .	18

3.3.3. Gráfica de individuos superpuesta . . . . .	19
3.4. Análisis del ejemplo <i>Gorgona</i> con $AFM(\mathbf{T}, \mathbf{Z})$ . . . . .	19
3.4.1. Análisis separados . . . . .	19
3.4.2. Resultados preliminares para determinar estructura común . . . . .	20
3.4.3. Análisis global . . . . .	21
<b>4. Comparación entre ACC y AFM aplicado a la tabla <math>(\mathbf{T}, \mathbf{Z})</math></b>	<b>23</b>
4.1. Comparación entre ACC y AFM aplicado a la tabla $(\mathbf{T}, \mathbf{Z})$ . . . . .	23
4.1.1. Teoría ACP ponderado . . . . .	23
4.1.2. Peso de los individuos . . . . .	24
4.1.3. Primera etapa común: análisis separados . . . . .	24
4.2. Comparación entre ACC y AFM aplicado a la tabla $(\mathbf{T}, \mathbf{Z})$ . . . . .	25
4.2.1. Objetivos de los métodos . . . . .	25
4.2.2. Ponderación de variables . . . . .	25
4.2.3. $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$ de los métodos . . . . .	25
4.2.4. Inercia y valores propios . . . . .	26
4.2.5. Distancias . . . . .	26
4.2.6. Relaciones de transición . . . . .	27
4.2.7. Gráficas y ayudas a la interpretación . . . . .	27
4.3. Comparación entre ACC y AFM aplicado a la tabla $(\mathbf{T}, \mathbf{Z})$ . . . . .	28
4.4. Comparación entre ACC y AFM aplicado a la tabla $(\mathbf{T}, \mathbf{Z})$ . . . . .	28
<b>5. Ejemplos de aplicación</b>	<b>32</b>
5.1. Primera aplicación . . . . .	32
5.1.1. Datos y objetivos del análisis . . . . .	32
5.1.2. Análisis factorial múltiple (AFM) . . . . .	33
5.1.3. Análisis canónico de correspondencias (ACC) . . . . .	36
5.2. Segunda aplicación . . . . .	37
5.2.1. Datos y objetivo del análisis . . . . .	37
5.2.2. Análisis factorial múltiple (AFM) . . . . .	38
5.2.3. Análisis canónico de correspondencias (ACC) . . . . .	41
5.3. Guía de análisis . . . . .	42
Software . . . . .	44

<b>Conclusiones</b>	<b>45</b>
<b>Recomendaciones</b>	<b>46</b>
<b>Apéndice</b>	<b>50</b>

# Índice de cuadros

1.1. Frecuencias de herperfauna y mediciones en los sitios en el ejemplo Gorgona . . . . .	7
1.2. Fórmulas del $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$ . . . . .	8
4.1. Comparación teórica entre los métodos ACC y AFM . . . . .	26
5.1. Datos de calidad e indicadores socio-educativos en los departamentos colombianos	33
5.2. Coordenadas y ayudas a la interpretación de los grupos activos . . . . .	35
5.3. Resultados del $ACS(\mathbf{T})$ y del $ACC(\mathbf{T}, \mathbf{Z})$ . . . . .	36
5.4. nombre y codificación de causas de mortalidad . . . . .	38
5.5. Resultados del análisis parcial y global en el AFM de la segunda aplicación . . . . .	39

# Índice de figuras

1.1. Tabla [ <b>T</b> <b>Z</b> ] de frecuencias - variables continuas. . . . .	4
1.2. Isla de Gorgona . . . . .	5
2.1. Plano factorial 1-2 del ACS( <b>T</b> ). Secciones y especies . . . . .	13
2.2. Plano factorial 1-2 del ACP( <b>Z</b> ). Secciones y círculo de correlaciones . . . . .	13
2.3. Plano Factorial 1-2 del ACC. Individuos, frecuencias y variables continuas . . . . .	14
2.4. Plano Factorial 1-2 del ACC. Triplot de secciones, especies y variables ambientales . . . . .	15
3.1. Resultados para los análisis separados y global del análisis factorial múltiple (AFM) . . . . .	20
3.2. Plano Factorial 1-2 del AFM. Secciones, especies y círculo de correlaciones . . . . .	21
3.3. Plano Factorial 1-2 del AFM. Individuos: Puntos medios y Puntos parciales . . . . .	22
4.1. Plano Factorial 1-2 del AFM. Columnas e individuos . . . . .	29
4.2. Plano Factorial 1-2 del ACC. Individuos, frecuencias y variables continuas . . . . .	30
5.1. Resultados para el análisis parcial y global del AFM. Calidad Educativa e indicadores socio-educativos	34
5.2. Plano factorial 1-2 en el AFM. Factores parciales . . . . .	35
5.3. Plano Factorial 1-2 del ACC. Individuos, frecuencias y variables continuas . . . . .	37
5.4. Plano factorial 1-2 en el AFM. Algunas comunidades autónomas. Puntos medios y puntos parciales.	40
5.5. Plano Factorial 1-2 del ACC. Individuos, frecuencias y variables continuas . . . . .	41

# Resumen

El análisis canónico de correspondencias (ACC) analiza tablas de frecuencias-variables continuas, es decir, tablas en las cuales las unidades estadísticas están descritas por dos grupos de variables, uno de frecuencias y otro de variables continuas, propuesto por Ter-Braak (1986) para estudiar la influencia de las condiciones del medio ambiente en la distribución de las especies de flora y fauna. Otro método que analiza tablas en las cuales un conjunto de individuos está descrito por varios grupos de variables, es el análisis factorial múltiple (AFM), desarrollado por Escofier & Pagès (1984, 1992). El principal objetivo de este trabajo es comparar metodológicamente el ACC y el AFM aplicado a tablas de frecuencias-variables continuas (Abdessemed & Escofier 1992). La comparación de los dos métodos hace referencia a ponerlos en paralelo, ya que no apuntan exactamente a los mismos objetivos metodológicos; se presenta propiedades, elementos comunes y diferentes de los métodos, y se ilustra con el ejemplo Urbina & Londoño (2003). También, se analizan con los dos métodos dos aplicaciones en otras áreas diferentes a la investigación medioambiental: una en *educación* y la otra en *salud pública*.

**Palabras claves:** tabla de frecuencias - variables continuas, análisis en componentes principales ponderado, análisis canónico de correspondencias, análisis factorial múltiple.

## Abstract

Canonical Correspondence Analysis (CCA) analyzes tables of frequencies - continuous variables, this is, tables in which statistical units are described by two groups of variables, one frequency and a continuous variable, proposed by Ter-Braak (1986) to study the influence of environmental conditions on the distribution of species of flora and fauna. Another method that analyzes tables in which a set of individuals is described by several groups of variables, is Multiple Factor Analysis (MFA), developed by Escofier & Pagès (1984, 1992). The main objective of this work is to compare methodologically CCA and the MFA applied to tables of frequencies - continuous variables (Abdessemed & Escofier 1992). Comparing the two methods referred to them in parallel, do not point to exactly the same methodological objectives; properties presents common elements and different methods, and illustrated with the example Urbina & Londoño (2003). Also, we analyze the two methods two different applications in other areas for environmental research, one in education and other public health.

**Key words:** frequency table - continuous variables, Ponderated Principal Component Analysis, Canonical Correspondence Analysis, Multiple Factor Analysis.

# Introducción

Diversas problemáticas conducen a construir tablas de frecuencias-variables continuas, es decir, tablas en las cuales las unidades estadísticas están descritas por dos grupos de variables, uno de frecuencias y otro de variables continuas. Los datos del grupo de frecuencias pueden ser conteos, porcentajes o respuestas binarias (presencia/ausencia). A continuación se mencionan algunos ejemplos de este tipo de tablas encontrados en la literatura:

- En Ecología, se estudia la influencia de las condiciones del medio ambiente (grupo de variables continuas) en la distribución de las especies de flora y fauna (grupo de frecuencias) (Chessel et al. 1987, Lebreton et al. 1988, Lebreton et al. 1991, Doledec & Chessel 1991, Birks & Austin 1994, Villalobos et al. 2000, Pavoine et al. 2003, Urbina & Londoño 2003, Berti et al. 2004).
- En análisis sensorial, se podría estudiar la preferencia de productos alimenticios teniendo en cuenta la frecuencia de consumo semanal (grupo de frecuencias) y sus características sensoriales (grupo de variables continuas). En Díaz (2002) se encuentra un ejemplo donde se analizan las frecuencias y en Pagès (2004) otro con el análisis de las variables continuas.
- En Salud Pública, al determinar factores de riesgo en el desarrollo de enfermedades cardiovasculares a poblaciones específicas, se realiza una encuesta estructurada de hábitos saludables (grupo de frecuencias) y se toma información del perfil lipídico: colesterol total, LDL, HDL y triglicéridos (grupo de variables continuas) (Ulate-Montero & Fernández-Ramírez 2001).

El análisis canónico de correspondencias (ACC) propuesto por Ter-Braak (1986) para estudios medioambientales, es uno de los métodos que permite estudiar la relación entre un grupo de frecuencias y un grupo de variables continuas sobre un mismo conjunto de individuos. El grupo de frecuencias juega el papel de variables de respuesta y el grupo de variables continuas juega el papel de variables explicativas que son de tipo cuantitativo.

El análisis factorial múltiple (AFM) (Escofier & Pagès 1984, 1992) permite tener en cuenta varios grupos de variables como elementos activos en un único análisis factorial, la condición es que las variables dentro de cada grupo sean del mismo tipo (cuantitativo o cualitativo). En el AFM, la información del grupo de variables se toma sobre un mismo conjunto de individuos.

En el presente trabajo se hace una comparación metodológica para poner en paralelo elementos comunes y diferentes entre el análisis canónico de correspondencias (ACC) y el análisis factorial múltiple (AFM) aplicado a tablas de frecuencias-variables continuas (Abdessemed & Escofier 1992),

con una estructura como la que se muestra en la figura 1.1; y se provee de una guía metodológica, primero para decidir cuando aplicar ACC, AFM o ambos y luego para la ejecución práctica de los métodos.

En el capítulo 1 se presenta la notación adoptada, una descripción del ejemplo que ilustra los métodos utilizando los datos del estudio realizado por Urbina & Londoño (2003), y un repaso del análisis en componentes principales (ACP) ponderado, método sobre el cual se construyen los métodos factoriales a comparar. En los capítulos 2 y 3 se repasan el análisis canónico de correspondencias (ACC) propuesto por (Ter-Braak 1986) y el análisis factorial múltiple (AFM) propuesto por (Escofier & Pagès 1984), respectivamente, vistos como ACP ponderados.

En el capítulo 4 se presenta la comparación metodológica entre el análisis canónico de correspondencias (ACC) y análisis factorial múltiple (AFM) aplicado a tablas de frecuencias-variables continuas (Abdessemed & Escofier 1992), con una estructura como la que se muestra en la figura 1.1. La comparación de los dos métodos hace referencia a ponerlos en paralelo, ya que no apuntan exactamente a los mismos objetivos.

En el capítulo 5 se ilustra con dos ejemplos diferentes al área de la ecología, la guía metodológica, primero para decidir cuando aplicar ACC, AFM o ambos y luego para la ejecución práctica de los métodos.

Para ejecutar los métodos se utiliza el lenguaje estadístico R (R Development Core Team 2009): los paquetes: *ade4* (Thioulouse et al. 1997) y *vegan* (Oksanen et al. 2007) para el método ACC y *ade4* para el método AFM.

# Capítulo 1

## Elementos básicos

### 1.1. Notación

La tabla a analizar se nota  $[\mathbf{T} \ \mathbf{Z}]$ , donde  $\mathbf{T}$  es una tabla de frecuencias en la que las celdas se expresan en términos absolutos (conteos, respuestas binarias) o en términos relativos (porcentajes); y  $\mathbf{Z}$  es una tabla de variables continuas en la que las celdas son datos cuantitativos (mediciones, tasa, etc) (figura 1.1). La tabla  $[\mathbf{T} \ \mathbf{Z}]$  de frecuencias - variables continuas tiene en común la información de los individuos en las filas.

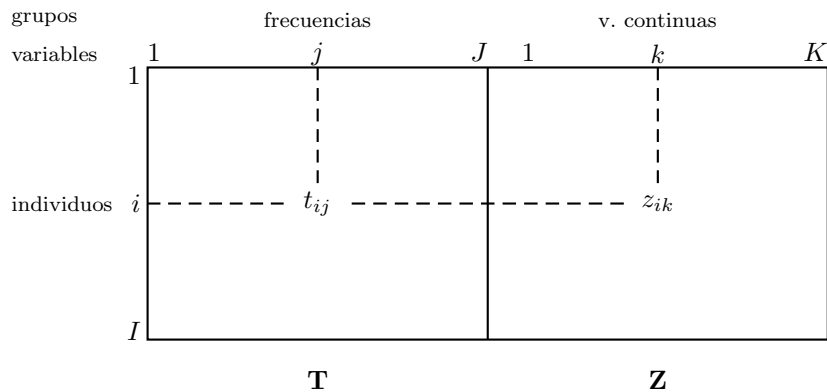


Figura 1.1: Tabla  $[\mathbf{T} \ \mathbf{Z}]$  de frecuencias - variables continuas.

Se adopta la misma notación de Escofier & Pagès (1992, cap. 7) en donde se utiliza la misma letra para denotar tanto al conjunto como al número de elementos:

- individuos:  $I = \{i : i = 1, \dots, I\}$ ;
- columnas-frecuencias:  $J = \{j : j = 1, \dots, J\}$ ;
- columnas-variables continuas:  $K = \{k : k = 1, \dots, K\}$ .

$\mathbf{T}$  es de dimensión  $I \times J$  y de término general  $t_{ij}$ . La tabla de frecuencias relativas asociada a la tabla  $\mathbf{T}$  se nota  $\mathbf{F}$  y su término general es  $f_{ij}$ . Las marginales fila y columna de la tabla  $\mathbf{F}$  se notan

$f_i$  y  $f_j$ . Se definen las matrices diagonales:  $\mathbf{D}_I = \text{diag}(f_i)$  y  $\mathbf{D}_J = \text{diag}(f_j)$ .  $\mathbf{Z}$  es de dimensión  $I \times K$  y de término general  $z_{ik}$ .

Las  $I$  filas de  $[\mathbf{T} \ \mathbf{Z}]$  conforman la nube  $N_I$  en  $\mathbb{R}^{J \oplus K}$  y las  $(J + K)$  columnas conforman la nube  $N_{(J \cup K)}$  en  $\mathbb{R}^I$ ; las  $I$  filas de la tabla  $\mathbf{T}$  conforman la nube de puntos  $N_I^1$  en  $\mathbb{R}^J$  y las  $J$  columnas conforman la nube de puntos  $N_J$  en  $\mathbb{R}^I$ ; las  $I$  filas de la tabla  $\mathbf{Z}$  conforman la nube de puntos  $N_I^2$  en  $\mathbb{R}^K$  y las  $K$  columnas la nube de puntos  $N_K$  en  $\mathbb{R}^I$ .

## 1.2. Ejemplo

Para ilustrar los métodos ACC, AFM y la comparación entre ellos, se usan los datos del estudio realizado por Urbina & Londoño (2003). El objetivo general es conocer la distribución de la comunidad de herpetofauna (anfibios y reptiles) en la Isla de Gorgona<sup>1</sup>, y determinar la posible relación de algunas especies con la temperatura, la humedad relativa y la cobertura vegetal sobre los microhábitats. Los autores hicieron conteos de especies de anfibios y reptiles en cuatro áreas (cultivos de palma, prisión<sup>2</sup>, bosques primarios, bosques secundarios) con diferente grado de perturbación antrópica en la Isla de Gorgona, durante junio y julio de 2001. La zona estudiada se muestra en la figura 1.2.

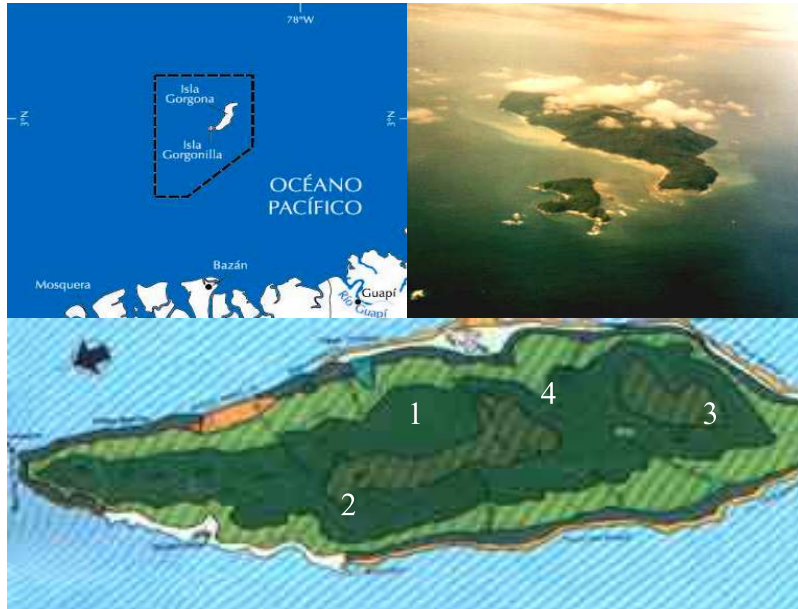


Figura 1.2: Isla de Gorgona. Ubicación de la isla desde la costa pacífica colombiana, vista satelital y mapa con la ubicación de las áreas de muestreo: 1. Cultivos de palma ( $CI-8$ ), 2. Prisión ( $PsI-8$ ), 3. B. Secundarios ( $SeI-8$ ), 4. B. Primarios ( $PrI-8$ ) Fuente: Google Earth

<sup>1</sup>Parque Nacional Natural ubicado en el departamento del Cauca, jurisdicción de Guapi.

<sup>2</sup>Hasta el 7 de agosto de 1985, fue una prisión de máxima seguridad.

El análisis del estudio Urbina & Londoño (2003) está orientado por las siguientes preguntas:

1. La distribución de las especies de anfibios y reptiles está asociada a los sitios?
2. La distribución de las especies de anfibios y reptiles en los diferentes sitios, está relacionada a las características de clima y habitat?

Los datos se muestran en la tabla 1.1: la tabla  $\mathbf{T}$  de frecuencias absolutas cruza 32 filas (secciones ubicadas en las diferentes áreas de la isla Gorgona) y 11 columnas (especies de reptiles y anfibios). La tabla  $\mathbf{Z}$  de variables continuas cruza las mismas filas (32 secciones) y 5 columnas (variables relacionadas a clima y habitat).

### 1.3. El análisis en componentes principales ponderado $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$

El análisis en componentes principales (ACP) recurre a dos representaciones geométricas: una para comparar a los individuos (nube de individuos) y otra para estudiar las relaciones entre las variables (nube de variables). Estas representaciones requieren de transformaciones de la tabla de datos. La transformación más utilizada es la de la estandarización de los datos, es decir restar la media (centrado) y dividir por la desviación estándar (reducido), lo que da origen al ACP normado. En este ACP se utiliza la distancia euclidiana canónica entre puntos.

El ACP ponderado es un ACP de una matriz  $\mathbf{X}$ , que contiene los datos a analizar (transformados); con distancias euclidianas definidas a partir de productos internos dados por matrices simétricas definidas positivas. La matriz  $\mathbf{M}$  define el producto interno en el espacio de las filas ( $\mathbb{R}^K$ ) y  $\mathbf{D}$  el producto interno en el espacio de las columnas ( $\mathbb{R}^J$ ). En la mayoría de los métodos las matrices  $\mathbf{M}$  y  $\mathbf{D}$  son diagonales conformadas por los pesos de las columnas y de las filas, respectivamente.

El ACP ponderado se denota  $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$  donde:

- $\mathbf{X}$  es la matriz a analizar (matriz de datos transformada según el método específico),
- $\mathbf{M}$  la matriz diagonal de pesos de las columnas, y
- $\mathbf{D}$  la matriz diagonal de pesos de las filas.

Las principales fórmulas del  $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$  se resumen en la tabla 1.2, de donde se pueden derivar las de un método particular una vez se han establecido las tres matrices (Escofier & Pagès 1992, capítulo 4).

En el ACP ponderado se busca, cómo en ACP clásico (Lebart et al. 1995, Escofier & Pagès 1992, Dray 2003), representaciones gráficas de la nube de las filas (planos factoriales) caracterizada por las columnas, y representaciones gráficas de la nube de las columnas caracterizada por las filas.

Tabla 1.1: Frecuencias de herperfauna y mediciones en los sitios en el ejemplo Gorgona

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	Z1	Z2	Z3	Z4	Z5
Ps1	1	0	0	1	1	3	117	0	0	1	0	28.4	81.8	30.7	54.3	32.4
Ps2	0	0	0	8	0	5	288	0	0	2	0	26.8	85.8	19.4	23.7	51.2
Ps3	0	0	0	1	3	0	141	0	0	26	0	25.5	85.2	56.2	66.2	60.0
Ps4	1	0	0	5	1	1	109	0	0	3	0	27.3	90.4	43.3	52.1	53.0
Ps5	0	0	0	1	0	0	14	1	0	2	0	24.6	83.8	12.5	22.6	69.0
Ps6	0	0	0	2	0	0	0	2	0	4	0	24.5	83.8	12.5	25.0	66.3
Ps7	0	0	0	0	1	0	10	6	0	10	0	26.0	81.2	32.0	77.0	65.0
Ps8	0	0	0	0	0	0	85	5	0	7	0	25.1	86.6	11.3	45.2	83.1
C1	0	0	0	0	0	0	29	1	0	0	0	27.5	84.5	23.7	41.2	55.7
C2	0	0	0	0	0	0	27	0	0	0	0	26.0	94.0	20.0	80.0	60.0
C3	1	1	0	10	0	0	141	0	0	7	0	25.1	84.3	45.0	20.0	60.8
C4	0	0	0	4	2	4	46	0	0	0	0	25.5	91.5	42.5	76.3	68.8
C5	0	0	0	0	0	0	3	12	0	0	0	30.6	88.6	36.3	43.3	92.0
C6	1	0	0	8	0	0	36	0	0	8	0	25.5	87.1	45.0	45.0	64.1
C7	0	0	0	0	0	0	3	12	0	0	0	24.5	90.5	75.0	77.5	40.0
C8	0	0	0	0	0	0	1	7	0	0	0	25.0	83.6	38.3	56.6	53.3
Se1	0	0	0	0	2	0	9	0	0	27	4	25.4	88.6	67.2	78.3	61.6
Se2	0	0	3	0	4	0	11	0	0	8	0	24.8	90.2	83.0	40.0	66.0
Se3	1	0	1	0	4	2	14	0	2	3	0	25.5	89.9	63.3	76.7	64.3
Se4	1	1	0	0	1	1	155	0	0	9	0	27.3	85.1	74.0	77.4	66.4
Se5	0	0	0	0	3	0	0	11	2	16	2	26.8	91.0	69.0	70.0	88.8
Se6	0	0	0	0	12	0	0	12	1	14	1	26.8	91.0	88.0	69.0	72.0
Se7	0	0	0	0	3	0	6	10	0	15	0	25.3	79.8	64.5	62.3	82.3
Se8	0	0	0	1	3	0	2	13	1	2	0	24.8	87.3	76.4	77.8	79.3
Pr1	0	0	0	0	1	0	0	0	0	12	1	25.3	92.6	80.0	58.3	73.3
Pr2	0	0	1	0	2	0	0	0	0	9	0	25.1	87.8	73.3	43.3	58.3
Pr3	1	0	0	0	2	0	0	1	0	16	0	25.2	88.7	63.7	71.2	45.0
Pr4	0	1	0	0	5	0	5	1	0	24	3	25.0	91.0	71.2	56.2	76.2
Pr5	0	0	0	0	1	0	0	14	0	8	2	22.5	96.2	25.5	10.7	95.0
Pr6	0	0	0	0	0	0	0	9	0	3	0	22.0	87.5	36.0	42.5	90.0
Pr7	0	1	1	0	0	0	1	10	1	6	0	24.0	80.0	68.5	63.3	85.5
Pr8	0	0	0	0	4	0	1	6	0	12	1	24.6	86.2	38.0	44.8	88.0

Nombre y código de frecuencias

Frecuencias	Código	Frecuencias	Código	Frecuencias	Código
t1: <i>B. constrictor</i>	Boa	t2: <i>B. atrox</i>	Mapaná	t3: <i>M. mipartitus</i>	Coral
t4: <i>B. galeritus</i>	Pasarroyo	t5: <i>E. heterolepis</i>	Lagarto	t6: <i>A. bridgessi</i>	Lobo
t7: <i>E. boulengeri</i>	R.venenosa	t8: <i>E. gularis</i>	R.brinconca	t9: <i>E. achatinus</i>	R.loteria
t10: <i>A. elegans</i>	R.arlequin	t11: <i>B. thyponius</i>	Sapo		

Nombres y códigos variables continuas

Variables cuantitativas	Código	Variables cuantitativas	Código
Z1: temperatura(°C)	Temp	Z2: humedad relativa (mm)	Humed
Z3: cobertura arbustiva (%)	Arbust	Z4: cobertura herbacea (%)	Herbac
Z5: cobertura de dosel (%)	Dosel		

Tabla 1.2: Fórmulas del  $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$ 

Nube	$N_I$	$N_K$
Espacio	$\mathbb{R}^K$	$\mathbb{R}^I$
Métrica	$\mathbf{M}$	$\mathbf{D}$
Coordenadas	filas de $\mathbf{X}$	columnas de $\mathbf{X}$
Peso	diagonal de $\mathbf{D}$	diagonal de $\mathbf{M}$
Inercia	$\text{traza}(\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M})$	$\text{traza}(\mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{D})$
Valor propio	$\lambda_s$	$\lambda_s$
Vector propio	$\mathbf{u}_s$	$\mathbf{v}_s$
Coordenadas factoriales	$F_s = \mathbf{X}\mathbf{M}\mathbf{u}_s$	$G_s = \mathbf{X}'\mathbf{D}\mathbf{v}_s$
Fórmulas de transición	$F_s = \frac{1}{\sqrt{\lambda_s}} \mathbf{X}\mathbf{M}\mathbf{G}_s$ $F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k=1}^K x_{ik} m_k G_s(k)$	$G_s = \frac{1}{\sqrt{\lambda_s}} \mathbf{X}'\mathbf{D}\mathbf{F}_s$ $G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^I x_{ik} d_i F_s(i)$

Un plano factorial es una aproximación de la nube de puntos y como tal tendrá puntos bien representados, pero podrá contener puntos con mala calidad de proyección. Se utilizan índices complementarios que ayudan a la interpretación de estos puntos (filas y columnas) (Escofier & Pagès 1992):

- *La contribución absoluta*, que indica los puntos que más aportan a la construcción de cada uno de los ejes.
- *La calidad de la representación*, llamada también contribución relativa, que se mide mediante el coseno al cuadrado entre el vector y su proyección sobre el eje. La suma de los cosenos cuadrados sobre los ejes 1 y 2, corresponde a la calidad de un punto sobre el primer plano factorial.
- *La distancia de un punto al origen*, en el espacio completo, que es igual a la norma del vector.

## 1.4. Análisis de correspondencias simples como un ACP ponderado

El análisis de correspondencias simples de la tabla de frecuencias  $\mathbf{T}$ , es el  $ACP(\mathbf{P}, \mathbf{D}_J, \mathbf{D}_I)$ , con  $\mathbf{D}_J = \text{diag}(f_{.j})$ ,  $\mathbf{D}_I = \text{diag}(f_{i.})$  y  $\mathbf{P} = \mathbf{D}_I^{-1} \mathbf{F} \mathbf{D}_J^{-1} - \mathbf{1}_{IJ}$  ( $\mathbf{1}_{IJ}$  es la matriz de unos, de dimensión  $I \times J$ ) (Doledec & Chessel 1991).

La matriz de frecuencias estandarizadas  $\mathbf{P}$ , tiene como término general:

$$p_{ij} = \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}} \quad (1.1)$$

Las  $I$  filas de la tabla  $\mathbf{T}$  conforman la nube de puntos  $N_I^1$  en  $\mathbb{R}^J$  y las  $J$  columnas la nube de puntos  $N_J$  en  $\mathbb{R}^I$ . La inercia de las dos nubes es igual y su valor es:  $\phi^2 = \sum_{s=1}^S \mu_s$ , donde:  $\{\mu_s, s = 1, \dots, S\}$  son los valores propios asociados al  $ACP(\mathbf{P}, \mathbf{D}_J, \mathbf{D}_I)$ ;  $S = \min\{I, J\} - 1$  y  $\phi^2$ , es el cuadrado medio de contingencia de  $\mathbf{T}$  (Greenacre 2007, pág. 28).

## 1.5. Análisis en componentes principales normado ponderado por las marginales fila de $\mathbf{F}$

En los métodos ACC y AFM, en una primera fase se realiza un ACP de la tabla de variables continuas  $\mathbf{Z}$ , utilizando como pesos de las filas  $\{f_i : i = 1, \dots, I\}$ , que son los mismos utilizados en el análisis de correspondencias simples de la tabla de frecuencias  $\mathbf{T}$  (sección 1.4). Estos pesos intervienen en el cálculo de la media ( $m_k$ ) y la varianza ( $s_k^2$ ) para la estandarización de las variables continuas (Ter-Braak 1986, Chessel et al. 1987, Abdessemed & Escofier 1992):

$$m_k = \sum_i f_i \cdot z_{ik} \quad \text{y} \quad s_k^2 = \sum_i f_i \cdot (z_{ik} - m_k)^2$$

El ACP ponderado de  $\mathbf{Z}$ , es el  $ACP(\mathbf{Z}_o, \mathbf{I}_K, \mathbf{D}_I)$ , donde  $\mathbf{I}_K$  es la matriz identidad de tamaño  $K$  y  $\mathbf{D}_I = \text{diag}(f_i)$ . La matriz de variables continuas estandarizadas  $\mathbf{Z}_o$ , tiene como término general:

$$z_{o_{ik}} = \frac{z_{ik} - m_k}{s_k} \quad (1.2)$$

Las  $I$  filas de la tabla  $\mathbf{Z}_o$  conforman la nube de puntos  $N_I^2$  en  $\mathbb{R}^K$  y las  $K$  columnas la nube de puntos  $N_K$  en  $\mathbb{R}^I$ . La inercia de las dos nubes es igual y su valor es  $K$ . El primer valor propio <sup>3</sup> asociado al  $ACP(\mathbf{Z}_o, \mathbf{I}_K, \mathbf{D}_I)$  se nota  $\nu_1$ .

---

<sup>3</sup>El primer valor propio es el valor propio más grande.

## Capítulo 2

# Análisis canónico de correspondencias (ACC)

El ACC (Ter-Braak 1986), es un método que permite analizar simultáneamente un grupo de frecuencias (conteos, respuestas binarias o porcentajes) y un grupo de variables (cuantitativas, cualitativas o ambas) sobre el mismo conjunto de individuos.

El ACC sólo toma en cuenta la parte de la estructura asociada a la tabla de frecuencias que se puede explicar por las variables continuas.

El ACC de la tabla  $[\mathbf{T} \ \mathbf{Z}]$  (figura 1.1) se hace de la siguiente manera:

1. El grupo de *frecuencias*  $\mathbf{T}$  juega el papel de *variables de respuesta o dependientes* y el grupo de *variables continuas* juega el papel de *variables independientes o explicativas*. Para el análisis, el grupo de variables continuas se estandariza (la tabla estandarizada se nota por  $\mathbf{Z}_o$ , ver sección 1.5). A partir de la tabla  $\mathbf{T}$  (grupo de frecuencias) se obtiene  $\mathbf{Y} = \mathbf{D}_I^{-1} \mathbf{F} \mathbf{D}_J^{-1}$ , con término general,  $y_{ij} = \frac{f_{ij}}{f_{i.} f_{.j}}$ .

2. Los valores de  $\mathbf{Y}$  se estiman mediante una regresión múltiple ponderada sobre el grupo de variables continuas estandarizadas  $\mathbf{Z}_o$ ,  $\hat{\mathbf{Y}} = \mathbf{Z}_o \hat{\mathbf{B}}$ , con  $\hat{\mathbf{B}} = (\mathbf{Z}_o' \mathbf{D}_I \mathbf{Z}_o)^{-1} \mathbf{Z}_o' \mathbf{F} \mathbf{D}_J^{-1}$ .

$\hat{\mathbf{Y}}$  es la proyección de  $\mathbf{Y}$  sobre el subespacio generado por  $\mathbf{Z}_o$ , es decir:  $\hat{\mathbf{Y}} = \mathbf{P}_{\mathbf{z}_o} \mathbf{D}_I^{-1} \mathbf{F} \mathbf{D}_J^{-1}$ , donde:  $\mathbf{P}_{\mathbf{z}_o} = \mathbf{Z}_o (\mathbf{Z}_o' \mathbf{D}_I \mathbf{Z}_o)^{-1} \mathbf{Z}_o' \mathbf{D}_I$ , es el  $\mathbf{D}_I$ -proyector (Sabatier et al. 1989).

A partir de  $\hat{\mathbf{Y}}$  se puede calcular la tabla de *frecuencias predichas*  $\hat{\mathbf{T}}$ :  $\hat{\mathbf{T}} = \mathbf{D}_I \hat{\mathbf{Y}} \mathbf{D}_J$

3. Finalmente se realiza el análisis de correspondencias (AC) de la tabla de frecuencias estimadas  $\hat{\mathbf{T}}$ , que es el análisis en componentes principales ponderado de  $\hat{\mathbf{Y}}$ ; con métricas para filas y columnas  $\mathbf{D}_J = \text{diag}(f_{.j})$  y  $\mathbf{D}_I = \text{diag}(f_{i.})$ , que son las mismas utilizadas en el análisis de correspondencias simples (ACS) de la tabla de frecuencias  $\mathbf{T}$  (sección 1.4) (Greenacre 2007).

En resumen, el ACC de la tabla  $[\mathbf{T} \ \mathbf{Z}]$ , notado  $ACC(\mathbf{T}, \mathbf{Z})$ , es el  $ACP(\hat{\mathbf{Y}}, \mathbf{D}_J, \mathbf{D}_I)$ . Todas las fórmulas se pueden derivar de las fórmulas correspondientes del  $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$  (ver tabla 1.2, pág.8).

## 2.1. Inercia y valores propios

La *inercia* total asociada con el análisis de correspondencias simples de la tabla de frecuencias  $\mathbf{T}$ , se divide en dos partes: la primera explicada por el  $ACC(\mathbf{T}, \mathbf{Z})$ , que es la parte relacionada linealmente al grupo de frecuencias por el grupo de variables continuas, llamado *inercia en el espacio restringido* (Greenacre 2007).

La inercia total asociada al análisis canónico de correspondencias es:

$$Inercia(ACC) = \sum_{i=1}^I \sum_{j=1}^J f_{i.j} (\hat{y}_{ij})^2 = \sum_{s=1}^S \lambda_s, \quad \text{donde } S = \min\{I-1, J-1, K\} \quad (2.1)$$

A las  $I$  filas de  $\mathbf{T}$  está asociada la nube  $N_I^1$  en el espacio restringido  $\mathbb{R}^{J^*}$  y a las  $J$  columnas esta asociada la nube  $N_J$  en el espacio  $\mathbb{R}^I$ . Los valores propios asociados al  $ACC(\mathbf{T}, \mathbf{Z})$  se notan  $\lambda_s$ .

Para los ejes principales del  $ACC(\mathbf{T}, \mathbf{Z})$  se definen:

- La proporción de inercia en cada eje  $s$  asociada al  $ACC(\mathbf{T}, \mathbf{Z})$  con respecto a la inercia asociada al mismo eje en el  $ACS(\mathbf{T})$ :  $\lambda_s/\nu_s$ , que es la proporción de *inercia* asociada al  $ACS(\mathbf{T})$  explicada por la relación lineal entre frecuencias y variables continuas.
- La proporción de inercia proyectada en cada eje  $s$  con respecto a la inercia total de las nubes en el ACC:  $\lambda_s / \sum_{s=1}^S \lambda_s$ , es decir la proporción de *inercia* explicada por la relación lineal entre las frecuencias y variables continuas que se retiene en el eje  $s$  del  $ACC(\mathbf{T}, \mathbf{Z})$ .

## 2.2. Gráficas y ayudas para la interpretación

El análisis canónico de correspondencias de  $[\mathbf{T} \ \mathbf{Z}]$  se interpreta como una aplicación regular de análisis de correspondencias simples ( $ACS(\hat{\mathbf{T}})$ ), por tanto, las ayudas a la interpretación (contribuciones, calidad de representación y distancias al origen) son aplicables.

### 2.2.1. Biplot

La gráfica para individuos y frecuencias se realiza con las coordenadas principales estandarizadas de individuos y las coordenadas factoriales de las frecuencias sobre los ejes del  $ACC(\mathbf{T}, \mathbf{Z})$ , este gráfico se denomina un biplot (Grabiel 1971) con escalamiento tipo 2. Las coordenadas factoriales estandarizadas de individuos son por construcción combinaciones lineales de las variables continuas estandarizadas y definen los ejes sobre los que se pueden proyectar las frecuencias.

### 2.2.2. Circulo de correlaciones

El círculo de correlaciones en el ACC para las variables continuas se construye buscando en cada eje, las correlaciones entre las variables continuas estandarizadas y las componentes principales estandarizadas de las filas del  $ACC(\mathbf{T}, \mathbf{Z})$ . La contribución a la formación de los ejes es nula. La

calidad de la representación en el plano se observa visualmente al dibujar el círculo de radio uno en el plano factorial.

Los coeficientes canónicos de las variables continuas (componentes principales del ACC en términos de las variables continuas) son los coeficientes de la regresión múltiple ponderada de las coordenadas factoriales estándar de las filas del ACC sobre las variables continuas estandarizadas.

### 2.2.3. Gráfico triplot

Es el gráfico donde aparecen los tres objetos de estudio relacionados: individuos, frecuencias y variables continuas. Los elementos (individuos, frecuencias) que participan en el  $ACC(\mathbf{T}, \mathbf{Z})$  se denominan activos, son representados en el gráfico por puntos, al igual que en el biplot con escalamiento tipo 2 (sección 2.2.1); las variables continuas se proyectan como elementos suplementarios. La coordenada de la proyección de una variable continua suplementaria en el  $ACC(\mathbf{T}, \mathbf{Z})$  equivale a su correlación con el eje y se representa por flechas desde el centro del gráfico por los coeficientes canónicos.

### 2.2.4. Prueba de permutación Monte Carlo

Es una prueba de hipótesis para determinar relación lineal entre frecuencias y variables continuas (Ter-Braak & Smilauer 2002). La hipótesis a contrastar es

$H_0$ : Las **columnas-frecuencias** no están relacionadas linealmente con las **columnas-variables continuas**.

Para esta prueba la estadística que se usa es la estadística *pseudo-F*:

$$pseudo-F = \frac{Inercia(ACC)/S}{\phi^2 - Inercia(ACC)}, \quad \text{donde } S = \min\{I - 1, J - 1, K\} \quad (2.2)$$

Si el *p-value* es significativo ( $p - value < \alpha$ ), las *columnas-frecuencias* están relacionadas linealmente a las *columnas-variables continuas*.

## 2.3. Análisis del ejemplo *Gorgona* con ACC

En el ejemplo *Gorgona* (Urbina & Londoño 2003), la inercia total asociada al ACS de la tabla de frecuencias  $\mathbf{T}$  es 1.308, los dos primeros valores propios se destacan sobre los demás y explican el 75 % de esta inercia (figura 3.1 a. y c.). El primer eje (figura 2.1) separa la especie *R.venenosa* (presentes en las áreas de prisión y cultivos) de las especies *R.arlequin*, *R.brincona* (presentes en las áreas de bosques primarios y secundarios).

En el ACP ponderado de las variables de clima y habitat, los dos primeros ejes explican el 69 % (figura 3.1, pág. 20) de la inercia total, el primer eje separa secciones con alta cobertura arbustiva y de dosel con secciones de baja cobertura arbustiva y de dosel, mientras que el segundo eje separa secciones con alta cobertura herbácea y alta temperatura con secciones con baja cobertura herbácea y baja temperatura (figura 2.2).

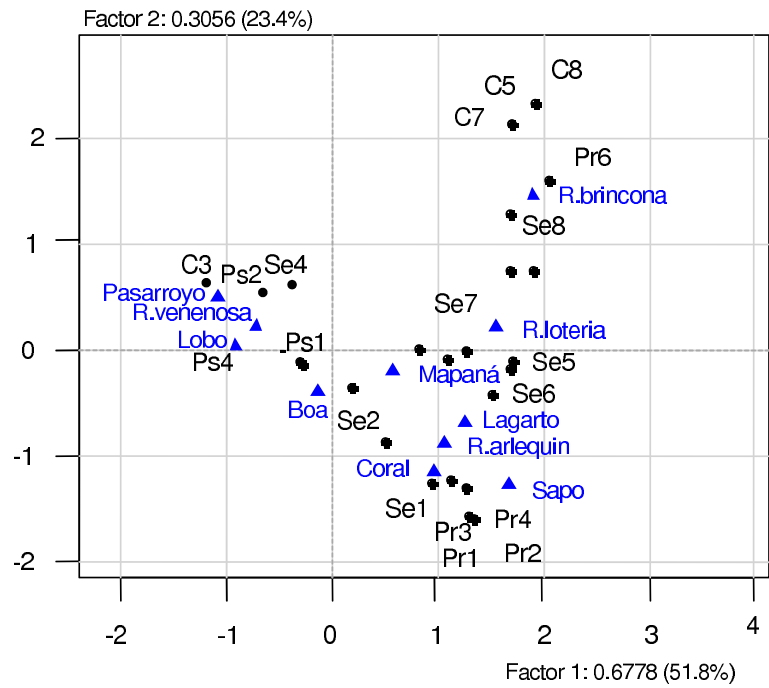


Figura 2.1: Plano factorial 1-2 del ACS(T). Secciones y especies

Las especies y las variables de clima y habitat tienen una relación lineal significativa (estadística  $Pseudo - F = 0.865$ ;  $P_{valor} = 0.005$ ).

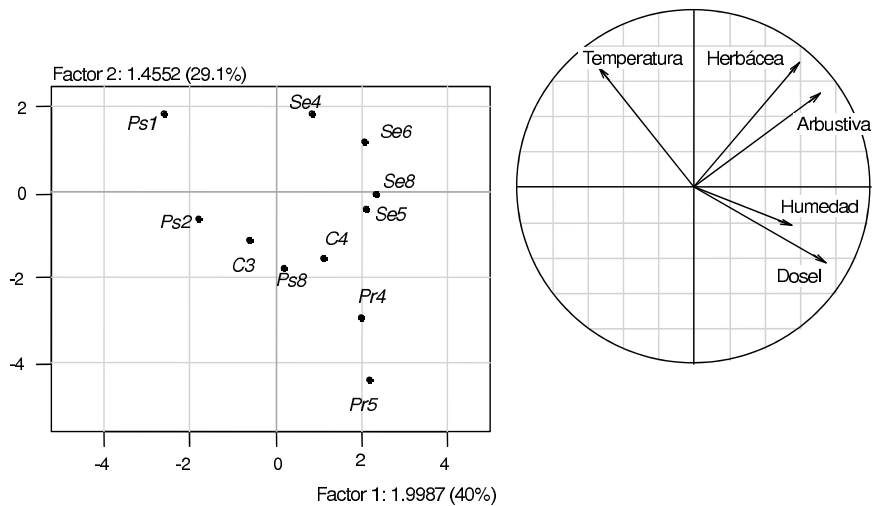


Figura 2.2: Plano factorial 1-2 del ACP(Z). Secciones y círculo de correlaciones

El 30% ( $Inercia(ACC)/Inercia(ACS) = 0.395/1.308$ ) de la inercia total asociada al análisis de correspondencias simples de la tabla de frecuencias es explicada por las variables continuas. Los valores propios asociados al  $ACC(\mathbf{T}, \mathbf{Z})$  muestran que los factores recuperados son bajos en comparación con los obtenidos en estudios que han empleado el  $ACC(\mathbf{T}, \mathbf{Z})$  (Ter-Braak 1986, Chessel et al. 1987, Lebreton et al. 1988). El primer eje del  $ACC(\mathbf{T}, \mathbf{Z})$  representa el 47.2% ( $\lambda_1/\mu_1=0.32/0.678$ ) de la inercia proyectada por el mismo eje del  $ACS(\mathbf{T})$ , indicando que las variables de clima y habitat relacionadas con este factor no explican las especies tan satisfactoriamente. Los restantes ejes canónicos  $ACC(\mathbf{T}, \mathbf{Z})$  no llegaron a representar más del 20% de los equivalentes en el  $ACS(\mathbf{T})$ , por lo que las variables de clima y habitat seleccionadas no explican tan satisfactoriamente estos ejes como el primero (Eje 2:  $\lambda_2/\mu_2 = 0.04/0.31 = 13.7\%$ ; Eje 3:  $\lambda_2/\mu_2 = 0.016/0.106 = 15.1\%$ ).

La inercia acumulada de la relación entre *variables de clima-habitat* y las *especies* en el primer eje del  $ACC(\mathbf{T}, \mathbf{Z})$  es del 81.0%, indica que las variables continuas explican satisfactoriamente este factor, el primer plano factorial recoge un 91.2% de la inercia total del  $ACC(\mathbf{T}, \mathbf{Z})$  suficiente para resumir la información de la relación entre *variables de clima-habitat* y las *especies*.

Del *biplot* y del círculo de correlaciones (figura 2.3) se destacan los resultados siguientes:

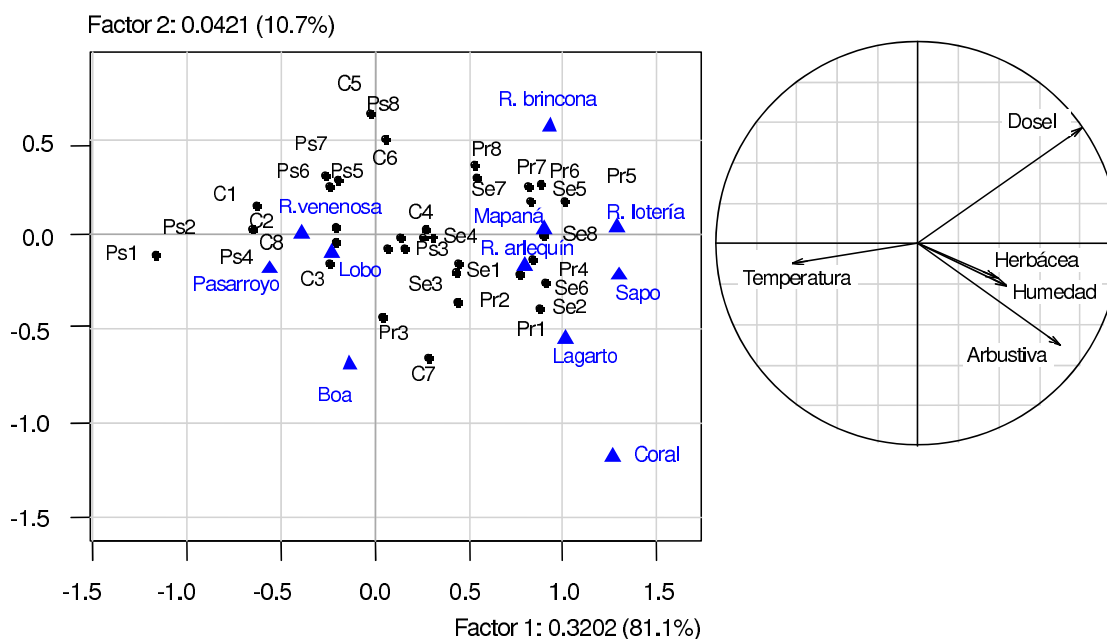


Figura 2.3: Plano Factorial 1-2 del ACC. Individuos, frecuencias y variables continuas

**Variables continuas:** las variables continuas que presentaron mayor asociación con respecto a la distribución de la comunidad de herpetofauna fueron: cobertura herbácea, cobertura de dosel y temperatura.

**Secciones:** el primer eje, se interpreta principalmente como la contraposición de las secciones de prisión  $Ps1$  y  $Ps2$  con las de bosques primarios ( $Pr1-Pr8$ ). Las secciones de prisión se

encuentran asociadas principalmente a ambientes con altas temperaturas, mientras que las áreas boscosas (bosque primario y secundario) aparecen asociadas a la cobertura de dosel y cobertura herbácea (agrupación de puntos *Se1-Se8* y *Pr1-Pr8*) lo que promueve mayor humedad y menores temperaturas en los microhabitats, generando un microclima similar en estas áreas.

**Especies:** se identificaron algunas especies afines a las áreas abiertas (prisión y cultivos) como: Boa, Lobo, Pasarroyo y R.venenosa; R.brincona, R.loteria y Mapaná se encuentran asociadas a las áreas boscosas; R.arlequin, Sapo, Coral y Lagarto están asociada al bosque primario y secundario.

Las relaciones entre especies y variables de clima-habitat sobre las secciones se puede leer en el *triplot* (figura 2.4):

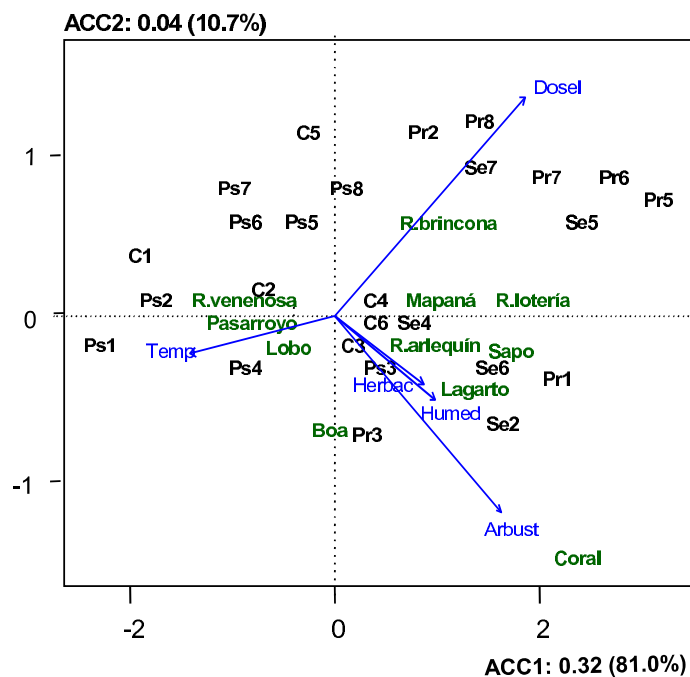


Figura 2.4: Plano Factorial 1-2 del ACC. Triplot de secciones, especies y variables ambientales

- La riqueza de especies es mayor en el bosque secundario.
- Las especies de reptiles Boa, Pasarroyo, Lobo y de anfibios R.venenosa se encontraron asociadas a áreas abiertas y su distribución estuvo fuertemente determinada por la temperatura del hábitat.
- Las especies asociadas a áreas boscosas, R.brincona, R.loteria y Mapaná se encontraron muy influenciadas por la cobertura de dosel sobre los microhábitat; mientras que la distribución de R.arlequin, Sapo, Coral y Lagarto estuvo fuertemente influenciada por la cobertura arbustiva.

## Capítulo 3

# Análisis factorial múltiple (AFM) aplicado a tablas de frecuencias y variables continuas

El análisis factorial múltiple (AFM) desarrollado por Escofier & Pagès (1984, 1992), es un método factorial adaptado al tratamiento de tablas de datos en las que un mismo conjunto de individuos se describe a través de varios grupos de variables. En cada grupo las variables deben ser de la misma naturaleza (cuantitativa o cualitativa).

Un AFM de la tabla  $[\mathbf{T} \ \mathbf{Z}]$  (figura 1.1) comparable con el ACC de la misma tabla, se realiza mediante las etapas siguientes:

***Etapla 1. Análisis parcial.*** Se realiza un ACP ponderado de cada uno de los grupos: un análisis de correspondencias simples para el grupo de frecuencias (sección 1.4, pág. 8) y un análisis en componentes principales normado para el grupo de variables continuas, utilizando como pesos de las filas las mismas del ACS( $\mathbf{T}$ ) (sección 1.5, pág. 9). Se nota  $\mu_1$  el primer valor propio asociado al  $ACP(\mathbf{P}, \mathbf{D}_J, \mathbf{D}_I)$  y  $\nu_1$  el primer valor propio asociado al  $ACP(\mathbf{Z}_o, \mathbf{I}_K, \mathbf{D}_I)$ .

***Etapla 2. Análisis global.*** El análisis factorial múltiple de  $[\mathbf{T} \ \mathbf{Z}]$ , notado  $AFM(\mathbf{T}, \mathbf{Z})$ , realiza un análisis en componentes principales ponderado de la tabla global  $[\mathbf{P} \ \mathbf{Z}_o]$  donde:  $\mathbf{P}$  es la tabla de frecuencias estandarizadas, y  $\mathbf{Z}_o$  es la tabla de variables continuas estandarizada; en este análisis cada tabla individual es ponderada por el inverso del primer valor propio obtenido en el ACP separado (Abdessemed & Escofier 1992).

En resumen, el  $AFM(\mathbf{T}, \mathbf{Z})$  como un ACP ponderado es el  $ACP([\mathbf{P} \ \mathbf{Z}_o], \mathbf{M}, \mathbf{D}_I)$ , donde:

- $\mathbf{M} = \text{diag} \left( \frac{1}{\mu_1} \mathbf{D}_J, \frac{1}{\nu_1} \mathbf{I}_K \right)$
- $\mathbf{D}_I = \text{diag}(f_i.)$

Las fórmulas se pueden derivar de las fórmulas correspondientes del  $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$  (tabla 1.2).

### 3.1. Inercia y valores propios

Las inercias de las dos nubes en el  $AFM(\mathbf{T}, \mathbf{Z})$  es igual y su valor es: la suma de la inercia de cada grupo dividido por su primer valor propio. La inercia del ACS( $\mathbf{T}$ ) es  $\phi^2$  (Greenacre 2007) y la inercia del ACP ponderado normado del grupo de variables continuas es igual a  $K$  (número de variables continuas).

$$Inercia(AF\!M) = \frac{\phi^2}{\mu_1} + \frac{K}{\nu_1} \quad (3.1)$$

A las  $I$  filas de la tabla  $[\mathbf{P} \ \mathbf{Z}_o]$  esta asociada la nube  $N_I$  en el espacio  $\mathbb{R}^{J \oplus K}$  y a las  $(J + K)$  columnas esta asociada la nube  $N_{(J \cup K)}$  en el espacio  $\mathbb{R}^I$ .

Si el primer valor propio del  $AFM(\mathbf{T}, \mathbf{Z})$  es cercano al máximo ( $2 =$  número de grupos en este caso), indica una estructura común a los grupos. Los valores propios asociados al  $AFM(\mathbf{T}, \mathbf{Z})$  se notan  $\gamma_s$ .

### 3.2. Grupos de variables

Los dos grupos se representan en  $\mathbb{R}^{I^2}$  por su coordenada sobre el eje factorial del  $AFM(\mathbf{T}, \mathbf{Z})$ , que es la inercia de la proyección de cada grupo sobre el factor principal del  $AFM(\mathbf{T}, \mathbf{Z})$  (Pagès 2004).

Si la coordenada de cada grupo de variables en cada eje factorial es cercana al máximo ( $= 1$ ), se puede decir que la estructura del grupo es más fuerte y su influencia será determinante en la construcción del primer factor del  $AFM(\mathbf{T}, \mathbf{Z})$  (Escofier & Pagès 1992).

Dos grupos son próximos si la distancia  $d^2$  entre *filas* es pequeña, para los grupos en estudio se representa esta distancia de la siguiente manera:

$$d^2(i, i') = \sum_{j=1}^J \frac{f_{.j}}{\mu_1} \left( \frac{f_{ij}}{f_{i.f.j}} - \frac{f_{i'j}}{f_{i'.f.j}} \right)^2 + \sum_{k=1}^K \frac{1}{\nu_1} \left( \frac{z_{ik} - z_{i'k}}{s_k} \right)^2 = \frac{d^2(i^1, i'^1)}{\mu_1} + \frac{d^2(i^2, i'^2)}{\nu_1} \quad (3.2)$$

Aquellos individuos cuyos puntos parciales (puntos que representan a cada individuo desde los diferentes grupos) se sitúen próximos ilustran la estructura común de los dos grupos analizados.

### 3.3. Gráficas y ayudas a la interpretación

En el análisis factorial múltiple de una tabla de frecuencias - variables continuas, se analizan tres tipos de objetos: *individuos*, *variables* y *grupos de variables*.

#### 3.3.1. Gráficas y ayudas a la interpretación de *individuos y variables*

La interpretación de la proyección de la nube de *columnas-variables continuas* se hace de manera análoga a la del ACP sobre *el círculo de correlaciones*, la coordenada de una variable  $k$  sobre un factor del  $AFM(\mathbf{T}, \mathbf{Z})$  representa la correlación entre está variable y el factor, de la misma

manera que en el ACP clásico. La contribución de cada variable a un eje  $s$  sirve para seleccionar las *columnas-variables continuas* que dan más significado al eje.

Para interpretar la relación entre una *columna-frecuencia* y una *columna-variable continua* medida por la covarianza entre el perfil de la *columna-frecuencia* y la *columna-variable continua* se hace igual que en un ACP clásico.

### 3.3.2. Gráfica y ayudas a la interpretación para los grupos de variables

Una representación gráfica de una nube de dos puntos que representa los dos grupos sobre los ejes factoriales de las nubes de individuos y de variables (poco útil en el caso de dos grupos solamente). Las coordenadas de los grupos toman valores entre 0 y 1, la representación de los grupos muestra cuales son similares (o diferentes) según el punto de vista de los factores del análisis global, la suma de las coordenadas de los grupos en cada eje es igual al valor propio en el  $AFM(\mathbf{T}, \mathbf{Z})$ , la contribución de cada grupo al eje es igual a la coordenada del grupo dividida por la suma de las coordenadas. El estudio para los grupos se completa con la *calidad de representación*<sup>1</sup> de cada grupo ubicada sobre el primer cuadrante del plano factorial 1-2.

El parecido entre las dos nubes parciales se puede evaluar globalmente mediante las siguientes ayudas adicionales:

**El coeficiente  $R_V$  de Escoufier.** Es un coeficiente que se obtiene a partir de los coeficientes de correlación lineal entre dos variables cualesquiera (Escoufier & Pagès 1992). Su valor está comprendido entre 0 y 1. Para los grupos en estudio, es:

$$R_V = \frac{\text{Traza}(\mathbf{P}\mathbf{D}_J\mathbf{P}'\mathbf{D}_I\mathbf{Z}_o\mathbf{Z}'_o\mathbf{D}_I)}{\sqrt{\sum_s(\mu_s^j)^2}\sqrt{\sum_s(\nu_s^k)^2}} \quad (3.3)$$

**El coeficiente  $L_g$ .** Mide la dimensionalidad de cada grupo (número de factores considerados). Este coeficiente toma valor cero (0) cuando no existe relación entre los grupos y no tiene límite superior.

$$L_g = \text{Traza}\left[\frac{1}{(\mu_1)^2}\mathbf{P}\mathbf{D}_J\mathbf{P}'\mathbf{D}_I\frac{1}{(\nu_1)^2}\mathbf{Z}_o\mathbf{Z}'_o\mathbf{D}_I\right] \quad (3.4)$$

**El coeficiente de correlación entre grupos y factores del AFM.** Mide la correlación entre las variables canónicas (proyección de los factores parciales obtenidos en el análisis individual de cada grupo sobre los ejes del análisis global) y los factores del análisis global del  $AFM(\mathbf{T}, \mathbf{Z})$ .

Los factores de los análisis separados se representan mediante su correlación con los factores del  $AFM(\mathbf{T}, \mathbf{Z})$ . Así, para comparar las componentes principales de los grupos, es suficiente introducirlos como elementos suplementarios en el análisis de la tabla completa.

---

<sup>1</sup>Los cosenos al cuadrado calculados en  $\mathbb{R}^{I^2}$

### 3.3.3. Gráfica de individuos superpuesta

Es la representación gráfica en un mismo espacio, de los individuos caracterizados por todas las variables (nube global media) y por cada uno de los grupos (nubes parciales).

A las  $I$  filas de la tabla  $[\mathbf{P}, \mathbf{Z}_o]$  esta asociada la nube  $N_I$  en el espacio  $\mathbb{R}^{J \oplus K}$  lo que permite situar las nubes de los grupos (frecuencias:  $N_I^1$  y variables continuas:  $N_I^2$ ) en el mismo espacio, representando los puntos relativos al mismo individuo tan próximos como sea posible. Aquí la distribución de los individuos para cada uno de los grupos se toman como elementos suplementarios en el análisis global. De hecho, los elementos no son suplementarios dado que contribuyen a la construcción de los ejes.

En la representación superpuesta (Abdessemed & Escofier 1992), las coordenadas factoriales para los individuos  $i^1$  (grupo de frecuencias) y  $i^2$  (grupo de variables continuas) sobre el eje  $s$  obtenido en el ACP global son:

$$F_s(i^1) = \frac{1}{\mu_1 \sqrt{\gamma_s}} \sum_{j=1}^J \frac{f_{.j}}{f_{.i}} G_s(j); \quad F_s(i^2) = \frac{1}{\nu_1 \sqrt{\gamma_s}} \sum_{k=1}^K \left( \frac{z_{ik} - m_k}{s_k} \right) H_s(k) \quad (3.5)$$

Donde:

- $\mu_1$ , es el primer valor propio asociado al análisis de correspondencias simples de la tabla de frecuencias,
- $\nu_1$ , es el primer valor propio asociado al análisis en componentes principales ponderado de la tabla de variables continuas,
- $\gamma_s$ , representa los valores propios asociados al AFM( $\mathbf{T}, \mathbf{Z}$ ) en el eje  $s$ ,
- $G_s(j)$  y  $H_s(k)$  son los factores de orden  $s$  para la frecuencia  $j$  y la variable continua  $k$  en el AFM( $\mathbf{T}, \mathbf{Z}$ ), respectivamente.

## 3.4. Análisis del ejemplo *Gorgona* con AFM( $\mathbf{T}, \mathbf{Z}$ )

### 3.4.1. Análisis separados

En el ejemplo *Gorgona* (Urbina & Londoño 2003), la inercia y el primer valor propio del ACS de especies son menores que los del ACP ponderado de variables de clima y habitat. El AFM( $\mathbf{T}, \mathbf{Z}$ ) equilibra la contribución de los grupos para evitar el dominio de las variables continuas en la construcción del primer eje (figura 3.1).

La gráfica de valores propios para los grupos, en los análisis separados, muestra que los dos grupos de variables tienen una primera dirección de inercia dominante, el primer plano factorial en cada uno de ellos explica alrededor del 50 % de variabilidad. La tabla de correlación entre los factores de los ACP separados muestra una correlación media (0.56) entre los primeros factores (figura 3.1).

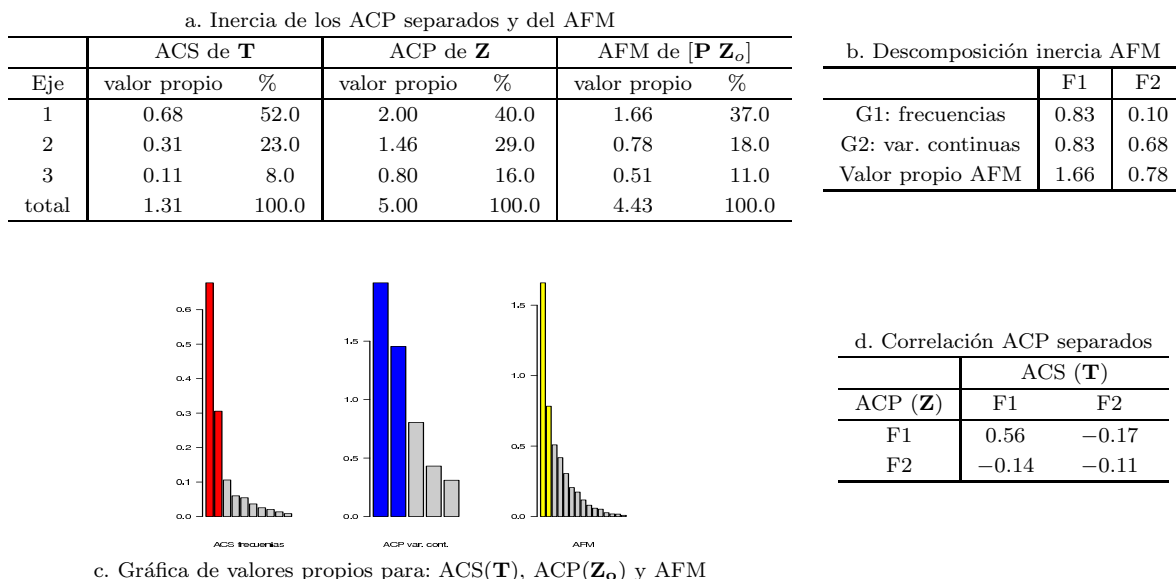


Figura 3.1: Resultados para los análisis separados y global del análisis factorial múltiple (AFM)

### 3.4.2. Resultados preliminares para determinar estructura común

Los indicadores iniciales del  $AFM(\mathbf{T}, \mathbf{Z})$  ponen de manifiesto la estructura común o semejanza global que tienen las dos tablas en el primer eje y estructura específica para el segundo eje. Este resultado puede observarse en las ayudas siguientes:

**Correlaciones entre los factores parciales de cada grupo y el factor global del AFM:** muestra un factor común a los dos grupos en el primer eje, las correlaciones son cercanas a 1 (0.90 para frecuencias y 0.86 para variables continuas), mientras que el segundo factor del AFM está más relacionado con las variables de clima y habitat en forma inversa (-0.89). Así, el primer plano proporcionado por el AFM es similar al de cada análisis separado, invirtiendo el segundo eje del ACP ponderado de las variables de clima-habitat.

**Contribución de los grupos a la formación de los ejes:** los dos grupos activos contribuyen de forma similar a la formación del primer eje (la inercia de cada grupo es de 0.83), mientras que al segundo eje contribuye más el grupo de variables de clima y habitat.

**Coefficientes RV y Lg:** el coeficiente RV de relación entre grupos es de 0.33, manifiesta una baja similitud entre los dos grupos en términos generales, mientras que el coeficiente Lg muestra igual dimensionalidad para las variables de clima-habitat (2.32) que para las especies de anfibios y reptiles (2.30), estos resultados coinciden con la dimensionalidad del AFM (2.28).

**Inercia del primer factor del AFM:** la inercia del primer factor ( $\gamma_1 = 1.66$ ) del AFM, indica la existencia de una estructura común.

### 3.4.3. Análisis global

**Inercia y valores propios:** la inercia total de la nube de secciones y de variables en el  $AFM(T, Z)$  es 4.43, la inercia del ACS de especies pasa de 1.31 a 1.93; mientras que la inercia del ACP de variables de clima y habitat cae de 5 a 2.5; lo que hace el  $AFM(T, Z)$  es equilibrar la contribución de los dos grupos a la formación del primer eje. Teniendo en cuenta los objetivos del estudio *Gorgona* y el histograma de valores propios (figura 3.1 c.), se interpretan los dos primeros ejes (65 % de la inercia total).

**Ejes factoriales.** De la figura 3.2 se destacan los resultados siguientes:

*Individuos.* El primer factor opone las secciones del área boscosa de las secciones *Ps1* y *Ps2*.

*Columnas.* El primer eje, se interpreta como la contraposición de las especies *R.brincona*, *R.arlequin*, y las variables de clima-habitat cobertura arbustiva y cobertura de dosel con respecto a la especie *R.venenosa* y la temperatura. Para el segundo factor, las variables que contribuyen pertenecen al grupo de variables continuas (cobertura arbustiva, cobertura herbácea y temperatura en el lado negativo).

*Individuos-Columnas.* El primer factor está relacionado a ubicación geográfica, altamente correlacionado con variables que pertenecen a los dos grupos. El segundo factor está más ligado a las variables continuas, está muy poco relacionado a la repartición de especies.

**Planos factoriales:** las relaciones entre especies y variables de clima-habitat sobre las secciones se puede leer en la figura 3.2:

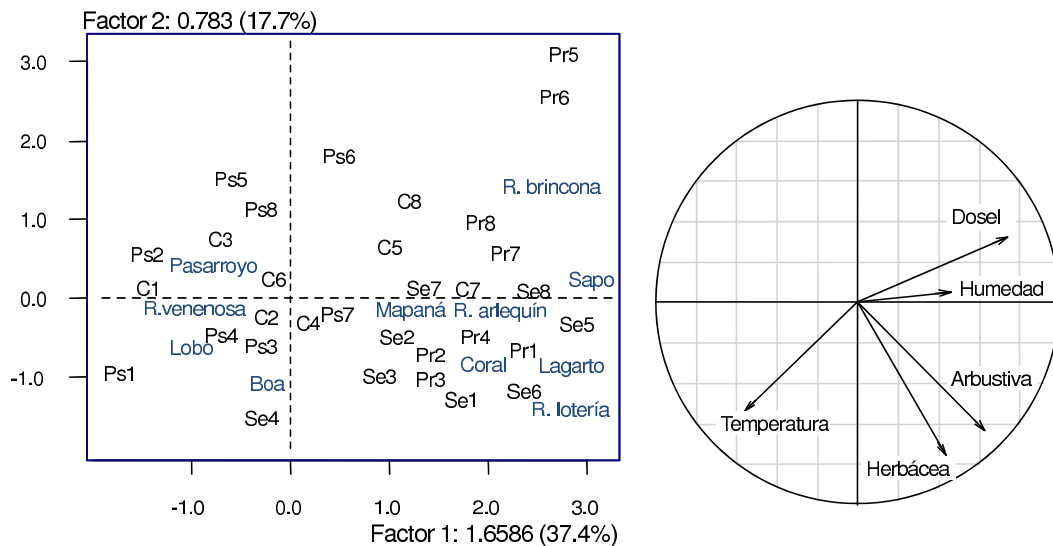


Figura 3.2: Plano Factorial 1-2 del AFM. Secciones, especies y círculo de correlaciones

- El área boscosa (bosque primario y secundario) presenta altos porcentajes de cobertura herbácea y arbustiva, como también bajas temperaturas; se encuentra la mayor riqueza de especies, habitan anfibios como: Lagartos, Corales y Mapanás, y reptiles como R.loteria y R.arlequin.
- Los sectores  $Ps1$  y  $Ps2$  presentan altas temperaturas y baja cobertura arbustiva y dosel, se presenta la especie R.venenosa.

**Representación superpuesta de los individuos descritos por cada grupo de variables por separado:**  $C1, 2, 3, 6$  opone las secciones de bosques primarios ( $Pr$ ), cualesquiera sea el conjunto de variables considerado (figura 3.3). Es otra manera, de poner de relieve un factor común entre los grupos.

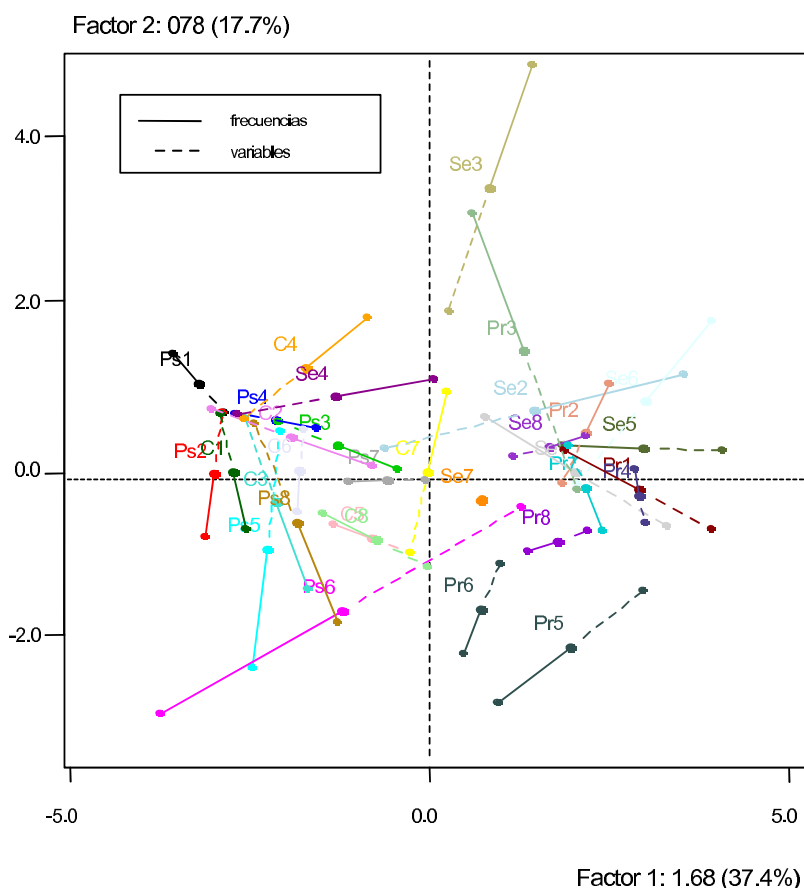


Figura 3.3: Plano Factorial 1-2 del AFM. Individuos: Puntos medios y Puntos parciales

## Capítulo 4

# Comparación entre los métodos: ACC y AFM, aplicados a tablas frecuencias-variables continuas

Los métodos factoriales: análisis canónico de correspondencias (capítulo 2) y análisis factorial múltiple (capítulo 3), permiten estudiar las relaciones que existen entre un grupo de frecuencias y un grupo de variables continuas descritos sobre un mismo conjunto de individuos (figura 1.1, pág. 4, sección 1.1).

En esta sección, a través de una comparación metodológica se ponen en paralelo algunas características técnicas de estos dos métodos (ver tabla 4.1), similar al artículo de Pagès (1996).

En ambos métodos, la tabla de datos se nota  $[\mathbf{T} \mathbf{Z}]$ ;  $\mathbf{T}$  es una tabla de frecuencias de dimensión  $I \times J$  y de término general  $t_{ij}$ , la tabla de frecuencias relativas asociada a la tabla  $\mathbf{T}$  se nota  $\mathbf{F}$  y su término general es  $f_{ij}$ . Las marginales fila y columna de la tabla  $\mathbf{F}$  se notan  $f_{i.}$  y  $f_{.j}$ .  $\mathbf{Z}$  es la tabla de variables continuas, de dimensión  $I \times K$  y de término general  $z_{ik}$ . La tabla de frecuencias estandarizadas  $\mathbf{P}$  tiene término general  $p_{ij}$ ; la tabla de variables continuas estandarizadas  $\mathbf{Z}_o$  tiene término general  $z_{o_{ik}}$ .

### 4.1. Elementos comunes

#### 4.1.1. Teoría ACP ponderado

El marco teórico general que permite definir métodos factoriales particulares es el análisis en componentes principales ponderado  $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$  (sección 1.3).

El ACP ponderado se denota  $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$  donde:  $\mathbf{X}$  es la matriz a analizar (matriz de datos transformada según el método específico),  $\mathbf{M}$  la matriz diagonal de pesos de las columnas, y  $\mathbf{D}$  la matriz diagonal de pesos de las filas.

*Ejemplos*

- análisis de correspondencias simples (ACS) de la tabla de frecuencias (sección 1.4, pág.8):  
 $\mathbf{X} = \mathbf{P}, \mathbf{M} = \mathbf{D}_J, \mathbf{D} = \mathbf{D}_I$
- análisis en componentes principales (ACP) de la tabla de variables continuas normado ponderado por las marginales fila de  $\mathbf{F}$  (sección 1.5, pág.9):  
 $\mathbf{X} = \mathbf{Z}_o, \mathbf{M} = \mathbf{D}_J, \mathbf{D} = \mathbf{D}_I$
- análisis canónico de correspondencias (ACC) de la tabla  $[\mathbf{T} \ \mathbf{Z}]$  (sección 2, pág.10):  
 $\mathbf{X} = \widehat{\mathbf{Y}}, \mathbf{M} = \mathbf{D}_J, \mathbf{D} = \mathbf{D}_I$
- análisis factorial múltiple (AFM) de la tabla  $[\mathbf{T} \ \mathbf{Z}]$  (sección 3, pág.16):  
 $\mathbf{X} = [\mathbf{P} \ \mathbf{Z}_o], \mathbf{M} = \text{diag} \left( \frac{1}{\mu_1} \mathbf{D}_J, \frac{1}{\nu_1} \mathbf{I}_K \right), \mathbf{D} = \mathbf{D}_I$

**4.1.2. Peso de los individuos**

Para este estudio, el  $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$  de los métodos particulares es:

- $ACS(\mathbf{T})$ :  $ACP(\mathbf{P}, \mathbf{D}_J, \mathbf{D}_I)$
- $ACP(\mathbf{Z})$ :  $ACP(\mathbf{Z}_o, \mathbf{D}_J, \mathbf{D}_I)$
- $ACC(\mathbf{T}, \mathbf{Z})$ :  $ACP(\widehat{\mathbf{Y}}, \mathbf{D}_J, \mathbf{D}_I)$
- $AFM(\mathbf{T}, \mathbf{Z})$ :  $ACP([\mathbf{P} \ \mathbf{Z}_o], \text{diag} \left( \frac{1}{\mu_1} \mathbf{D}_J, \frac{1}{\nu_1} \mathbf{I}_K \right), \mathbf{D}_I)$

Los ACP ponderados individuales de los grupos de variables (frecuencias, variables continuas) y de los métodos factoriales a comparar ( $ACC(\mathbf{T}, \mathbf{Z})$  y  $AFM(\mathbf{T}, \mathbf{Z})$ ) tienen en común la matriz diagonal de pesos de las filas (individuos) y matriz de métrica de las columnas:  $\mathbf{D} = \mathbf{D}_I = \text{diag}(f_i)$ .

**4.1.3. Primera etapa común: análisis separados**

En los métodos  $ACC(\mathbf{T}, \mathbf{Z})$  y  $AFM(\mathbf{T}, \mathbf{Z})$ , se realiza primero un análisis de correspondencias simples para la tabla de frecuencias ( $\mathbf{T}$ ) y un ACP normado ponderado para la tabla de variables continuas ( $\mathbf{Z}$ ).

**Comparación:** en el  $AFM(\mathbf{T}, \mathbf{Z})$  en la primera etapa, se observa la gráfica de valores propios de cada grupo por separado, esencialmente para evaluar el número de dimensiones que intervendrán de manera significativa en el análisis de la tabla global  $[\mathbf{T} \ \mathbf{Z}]$ , un grupo de mayor dimensionalidad tendrá una mayor influencia global en el sentido que contribuirá a un mayor número de ejes. Para el  $ACC(\mathbf{T}, \mathbf{Z})$ , se mira la inercia y valores propios solamente del grupo de frecuencias, si la proporción de inercia ( $\lambda_s/\mu_s$ ) en cada eje asociada al  $ACC(\mathbf{T}, \mathbf{Z})$  con respecto a la inercia del ACS asociada al mismo eje es  $\geq 40\%$  (Ter-Braak 1986, Chessel et al. 1987, Lebreton et al. 1988), se puede considerar que ninguna variable continua ha sido pasada por alto, y no resultan fundamentalmente diferentes la distribución de frecuencias en el  $ACS(\mathbf{T})$  y en el  $ACC(\mathbf{T}, \mathbf{Z})$ .

## 4.2. Elementos diferentes

### 4.2.1. Objetivos de los métodos

Los objetivos en el  $AFM(\mathbf{T}, \mathbf{Z})$  no se limitan a la obtención de una tipología de los individuos definida a través del conjunto de variables, sino que busca posibles relaciones entre las estructuras obtenidas en cada uno de los dos grupos. En el  $ACC(\mathbf{T}, \mathbf{Z})$  el objetivo no es sólo estudiar las asociaciones entre *individuos* y *frecuencias* al igual que el análisis de correspondencias simples sino también estudiar las relaciones de dependencia que tengan estas frecuencias con el grupo externo de variables continuas, es decir, la obtención de una tipología de individuos definida en una parte restringida del espacio de las frecuencias, que es la parte explicada por la relación con las variables continuas.

### 4.2.2. Ponderación de variables

En el  $AFM(\mathbf{T}, \mathbf{Z})$ , las variables están representadas por frecuencias y variables continuas, en el  $ACC(\mathbf{T}, \mathbf{Z})$  sólo por frecuencias (estimadas). En el  $ACC(\mathbf{T}, \mathbf{Z})$ , las frecuencias (estimadas) al igual que las frecuencias en el ACS de la tabla  $\mathbf{T}$  no se ponderan, el peso de las columnas y matriz de métrica en el espacio de las filas es  $\mathbf{M} = \mathbf{D}_J = \text{diag}(f_{.j})$ . En el  $AFM(\mathbf{T}, \mathbf{Z})$  se equilibra la influencia de cada grupo de variables en el análisis global ponderando por el inverso del primer valor propio obtenido en el análisis separado de cada grupo, por lo tanto, el peso de las columnas y matriz de métrica de las filas,  $\mathbf{M} = \text{diag}(\frac{1}{\mu_1} \mathbf{D}_J, \frac{1}{\nu_1} \mathbf{I}_K)$ . Está ponderación contrae la nube de las variables sin alterar la estructura interna del grupo. Iguala a 1 la inercia del primer eje de cada tabla impidiendo que el grupo de variables continuas pueda determinar por sí sólo el primer eje del análisis global.

### 4.2.3. $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$ de los métodos

- $ACC(\mathbf{T}, \mathbf{Z})$ :  $ACP(\hat{\mathbf{Y}}, \mathbf{D}_J, \mathbf{D}_I)$

sólo las frecuencias son los elementos activos, las variables continuas son proyectadas como variables suplementarias o ilustrativas, las cuales ayudan a la construcción de las coordenadas factoriales de individuos.

- $AFM(\mathbf{T}, \mathbf{Z})$ :  $ACP([\mathbf{P} \ \mathbf{Z}_o], \text{diag}\left(\frac{1}{\mu_1} \mathbf{D}_J, \frac{1}{\nu_1} \mathbf{I}_K\right), \mathbf{D}_I)$

las frecuencias y las variables continuas son elementos activos, y ambos grupos contribuyen a la formación de los ejes.

En los dos métodos se obtienen para cada eje factorial: las coordenadas, las contribuciones y los cosenos cuadrados para *individuos* y *columnas-frecuencias*, y los coeficientes de correlación entre las *columnas – variables continuas* y los factores. La diferencia entre ellos, es que las *columnas – variables continuas* en el  $ACC(\mathbf{T}, \mathbf{Z})$  no contribuyen a la formación de los ejes directamente pero sí a través de la proyección de las frecuencias sobre el subespacio generado por las variables continuas.

Tabla 4.1: Comparación teórica entre los métodos ACC y AFM

Método	Análisis canónico de correspondencias (ACC)	Análisis factorial múltiple (AFM)
Nube de Individuos	$N_{I1}$	$N_I$
Espacio de Individuos	$\mathbb{R}^{J^*}$	$\mathbb{R}^{J \oplus K}$
Nube de Variables	$N_J$	$N_{J \cup K}$
Espacio de Variables	$\mathbb{R}^{I1}$	$\mathbb{R}^I$
Peso de los Individuos matriz $\mathbf{X}$	$\mathbf{D}_I = \text{diag}(f_{i.})$	$\mathbf{D}_I = \text{diag}(f_{i.})$
Ponderación de las Variables	$\hat{\mathbf{Y}} = \mathbf{P}_{\mathbf{z}_o} \mathbf{D}_I^{-1} \mathbf{F} \mathbf{D}_J^{-1}$	$[\mathbf{P} \ \mathbf{Z}_o]$ $\text{diag}(\frac{1}{\mu_1} \mathbf{D}_J, \frac{1}{\nu_1} \mathbf{I}_K)$
$ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$	$ACP(\hat{\mathbf{Y}}, \mathbf{D}_J, \mathbf{D}_I)$	$ACP([\mathbf{P} \ \mathbf{Z}_o], \text{diag}(\frac{1}{\mu_1} \mathbf{D}_J, \frac{1}{\nu_1} \mathbf{I}_K), \mathbf{D}_I)$
Inercia	$\sum_{i=1}^I \sum_{j=1}^J f_{i.} f_{.j} (\hat{y}_{ij})^2$	$\frac{\phi^2}{\mu_1} + \frac{K}{\nu_1}$
Valor propio	$\lambda_s, 0 \leq \lambda_s \leq 1$	$\gamma_s, 1 \leq \gamma_1 \leq 2$
Fórmula de transición para filas	$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \left[ \sum_{j=1}^J f_{.j} \hat{y}_{ij} G_s(j) \right]$	$F_s(i) = \frac{1}{\sqrt{\gamma_s}} \left[ \sum_{j=1}^J \frac{f_{.j}}{\mu_1} \left( \frac{f_{ij} - f_{i.} f_{.j}}{f_{i.} f_{.j}} \right) G_s(j) \right]$ $+ \frac{1}{\sqrt{\gamma_s}} \left[ \sum_{k=1}^K \frac{1}{\nu_1} \left( \frac{z_{ik} - m_k}{s_k} \right) H_s(k) \right]$
Fórmula de transición para columnas	$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \left[ \sum_{i=1}^I f_{i.} \hat{y}_{ij} F_s(i) \right]$	$G_s(j) = \frac{1}{\sqrt{\gamma_s}} \sum_{i=1}^I \frac{f_{ij}}{f_{i.} f_{.j}} F_s(i)$ $H_s(k) = \frac{1}{\sqrt{\gamma_s}} \sum_{i=1}^I f_{i.} \left( \frac{z_{ik}}{s_k} \right) F_s(i)$
Representación de los grupos de variables	No aplica	Los grupos son representados en $\mathbb{R}^{I^2}$
Representación superpuesta	No aplica	La distribución de los individuos para cada grupo se toma como elementos suplementarios en el análisis global

#### 4.2.4. Inercia y valores propios

Las inercias de los métodos  $ACC(\mathbf{T}, \mathbf{Z})$  y  $AFM(\mathbf{T}, \mathbf{Z})$  no son comparables (ver tabla 4.1, pág.26). La inercia del  $ACC(\mathbf{T}, \mathbf{Z})$  siempre va tomar un valor menor o igual a la inercia del ACS de la tabla de frecuencias. Por lo tanto, la inercia del  $AFM(\mathbf{T}, \mathbf{Z})$  siempre será mayor a la del  $ACC(\mathbf{T}, \mathbf{Z})$ .

#### 4.2.5. Distancias

*Distancia entre individuos:*

- En el  $ACC(\mathbf{T}, \mathbf{Z})$ :  $d^2(i, l) = \sum_{j=1}^J f_{.j} (f_{i.} \hat{y}_{ij} - f_{l.} \hat{y}_{lj})^2$
- En el  $AFM(\mathbf{T}, \mathbf{Z})$ :  $d^2(i, l) = \sum_{j=1}^J \frac{f_{.j}}{\mu_1} \left( \frac{f_{ij}}{f_{i.} f_{.j}} - \frac{f_{lj}}{f_{l.} f_{.j}} \right)^2 + \sum_{k=1}^K \frac{1}{\nu_1} \left( \frac{z_{ik} - z_{lk}}{s_k} \right)^2$

*Distancia entre columnas*

- En el  $ACC(\mathbf{T}, \mathbf{Z})$ :  $d^2(j, q) = \sum_{i=1}^I f_{i.} (f_{.j} \hat{y}_{ij} - f_{.q} \hat{y}_{iq})^2$
- En el  $AFM(\mathbf{T}, \mathbf{Z})$ :  $d^2(j, q) = \sum_{i=1}^I \frac{f_{i.}}{\mu_1} \left( \frac{f_{ij}}{f_{i.} f_{.j}} - \frac{f_{iq}}{f_{i.} f_{.q}} \right)^2$ , y,  $d^2(k, r) = \sum_{i=1}^I \frac{f_{i.}}{\nu_1} \left( \frac{z_{ik} - m_k}{s_k} - \frac{z_{ir} - m_r}{s_r} \right)^2$

En ambos métodos, la distancia entre individuos o entre frecuencias se traduce en términos de la distancia  $\chi^2$ , y se interpretan en términos de proximidad o relación. La relación entre dos variables continuas o entre una frecuencia y una variables continua se expresa en términos de relación como en un ACP clásico.

#### 4.2.6. Relaciones de transición

En este sentido, hay que señalar que las relaciones de transición para un individuo  $i$  y para una frecuencia  $j$  permiten su interpretación análoga a un análisis de correspondencias simples. Las *filas – individuos* o *columnas – frecuencias* correspondientes a categorías de menor frecuencia son las más alejadas del origen de la representación.

#### 4.2.7. Gráficas y ayudas a la interpretación

##### Mapas factoriales

Los *Individuos* y las *columnas-frecuencias* son representados simultáneamente en planos factoriales igual que en el análisis de correspondencias simples.

La diferencia radica en que:

- En el  $ACC(\mathbf{T}, \mathbf{Z})$ , el mapa factorial se hace con *coordenadas factoriales estandarizadas* de individuos y las *coordenadas factoriales* de frecuencias (biplot con escalamiento tipo 2).
- En el  $AFM(\mathbf{T}, \mathbf{Z})$ , el mapa factorial se hace con *coordenadas factoriales* para individuos y frecuencias.

Las *columnas-variables continuas* caracterizadas por los individuos se representan en un círculo de correlaciones y se interpretan igual que en un ACP.

La diferencia radica en que:

- En el  $ACC(\mathbf{T}, \mathbf{Z})$ : las variables continuas son tomados como elementos suplementarios.
- En el  $AFM(\mathbf{T}, \mathbf{Z})$ : las variables continuas son tomadas como elementos activos.

##### Ayudas a la interpretación

El método  $AFM(\mathbf{T}, \mathbf{Z})$  es más exhaustivo para la detección de estructura comunes o específicas, cuenta con representaciones gráficas e indicadores que ayudan a esto: resultados del análisis separado de cada grupo (inercia, valores propios, correlación entre los factores de los grupos individuales, mapas factoriales para los grupos separados (nube parcial)), la descomposición de la inercia en cada eje del AFM (coordenadas de grupos), correlación entre factores parciales y factores globales; y, medidas de asociación (coeficientes  $Lg$  y  $RV$ ) que permiten cuantificar la semejanza global existente. También, sobre la representación global presenta trayectorias parciales de los individuos

vistos a través de los grupos por separado (representación superpuesta), de bastante interés si el objetivo del estudio es este.

El  $ACC(\mathbf{T}, \mathbf{Z})$  cuenta con la prueba de permutación Montecarlo (Greenacre 2007) para determinar la relación existente entre las frecuencias y las variables continuas, lo cuál complementa el análisis.

### 4.3. Criterios para analizar la tabla $[\mathbf{T} \ \mathbf{Z}]$

En ambos métodos, al analizar una tabla  $[\mathbf{T} \ \mathbf{Z}]$  la distribución de los individuos en cada eje pueden ser similares cuando los grupos de variables están relacionados o tienen estructuras comunes.

En primera instancia determinar *estructuras comunes* significa realizar un  $AFM(\mathbf{T}, \mathbf{Z})$  que cumpla las siguientes condiciones:

- Inercia en el primer eje superior a 1.4.
- Correlación entre factores parciales de cada grupo y factores globales del  $AFM(\mathbf{T}, \mathbf{Z})$  cercanos a  $\pm 1$ .
- Un coeficiente  $RV$ , que se interpreta como un coeficiente de correlación entre las tablas  $\mathbf{T}$  y  $\mathbf{Z}$ , tenga un valor superior a 0.5.
- Las coordenadas de los dos grupos de variables en cada eje factorial del  $AFM(\mathbf{T}, \mathbf{Z})$  cercanas a uno (sección 3.2).

Estos valores salen de bases teóricas (Escofier & Pagès 1984, Escofier & Pagès 1992, Abdessemed & Escofier 1992), resultados de las aplicaciones realizadas en este trabajo y aplicaciones referenciadas (Chessel et al. 1987, Lebreton et al. 1988, Lebreton et al. 1991, Doledec & Chessel 1991, Abdessemed & Escofier 1992, Birks & Austin 1994, Villalobos et al. 2000, Pavoine et al. 2003, Urbina & Londoño 2003, Sánchez-González & López-Mata 2003).

Después de encontrar *estructuras comunes* con el  $AFM(\mathbf{T}, \mathbf{Z})$  se debe realizar un análisis más fino con el método factorial análisis canónico de correspondencias,  $ACC(\mathbf{T}, \mathbf{Z})$ , para determinar las posibles relaciones entre las frecuencias y las variables continuas, si se tiene conocimiento que el grupo de frecuencias es explicado por el grupo de variables continuas. En caso contrario, que no se tenga conocimiento de dependencia entre los grupos se sigue con el análisis global que ofrece el análisis factorial múltiple,  $AFM(\mathbf{T}, \mathbf{Z})$ .

### 4.4. Comparación entre el $ACC(\mathbf{T}, \mathbf{Z})$ y el $AFM(\mathbf{T}, \mathbf{Z})$ para el ejemplo *Gorgona*

En esta sección se va a realizar un análisis comparativo de los resultados proporcionados por las dos métodos considerados en el estudio Urbina & Londoño (2003). Aunque, dada la naturaleza de los datos de esta aplicación, éstos son susceptibles de ser analizados mediante los dos procedimientos descritos.

**Inercia y valores propios:** las inercias de los métodos no son comparables. La inercia del  $ACC(\mathbf{T}, \mathbf{Z})$  es 0.398 mientras que la inercia global del  $AFM(\mathbf{T}, \mathbf{Z})$  es 4.43. En el primer eje, el  $ACC(\mathbf{T}, \mathbf{Z})$  explica un 81.1 % ( $\lambda_1 = 0.320$ ) de la variabilidad total, y el  $AFM(\mathbf{T}, \mathbf{Z})$  sólo explica el 37 % ( $\gamma_1 = 1.66$ ).

La ponderación aumenta sistemáticamente en el  $AFM(\mathbf{T}, \mathbf{Z})$  la importancia de la tabla de frecuencias, la inercia global cambia de 6.31 a 4.43; la ponderación en el  $AFM(\mathbf{T}, \mathbf{Z})$  lo que hace es equilibrar la influencia de los dos grupos de variables, para que el grupo de variables continuas no domine la construcción del primer eje del  $AFM(\mathbf{T}, \mathbf{Z})$  global (Escofier & Pagès 1992)

**Detección de estructuras comunes**<sup>1</sup>: al realizar el  $AFM(\mathbf{T}, \mathbf{Z})$  (sección 3) del ejemplo Urbina & Londoño (2003), muestra la *estructura común* o la semejanza global que tienen los dos grupos en el primer eje: primer valor propio igual a 1.66, las correlaciones entre los factores parciales de cada grupo y el factor global del  $AFM(\mathbf{T}, \mathbf{Z})$  son cercanas a 1 (0.90 para frecuencias y 0.86 para variables continuas). Los dos grupos activos contribuyen de forma similar (0.83 es la inercia de cada grupo) a la formación del primer eje del  $AFM(\mathbf{T}, \mathbf{Z})$ . El coeficiente RV es de 0.33, lo cuál manifiesta una baja similitud entre los dos grupos en términos generales. Por eso, es que el primer factor en los dos métodos es el mismo y la inercia explicada en el primer eje del  $ACC(\mathbf{T}, \mathbf{Z})$  es alta (81.1%), la relación entre las frecuencias y las variables continuas explicadas en el espacio restringido es satisfactorio.

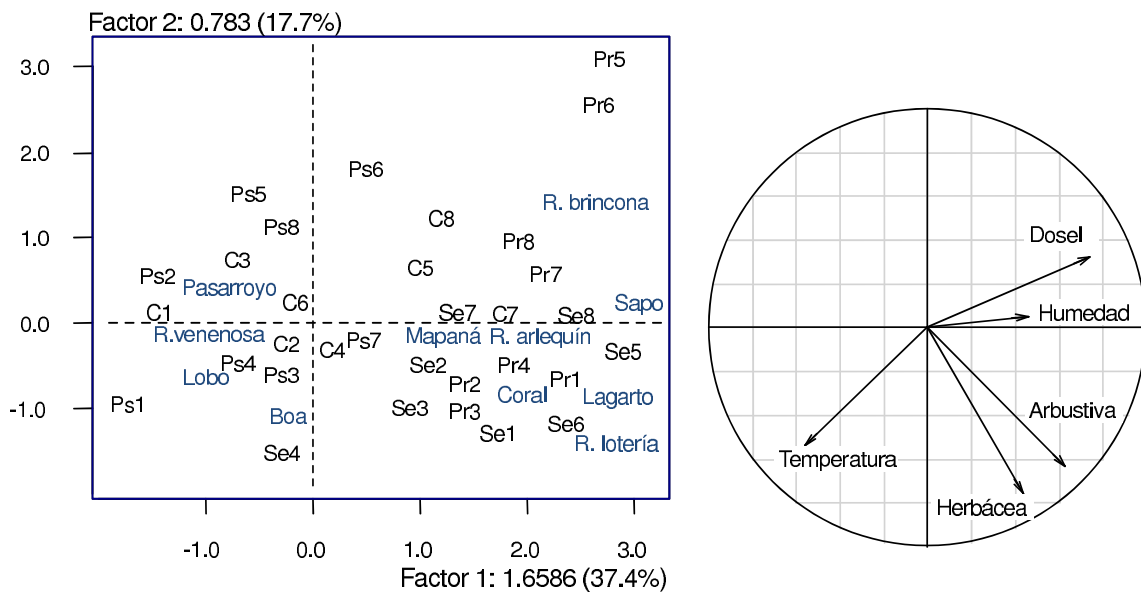


Figura 4.1: Plano Factorial 1-2 del AFM. Columnas e individuos

<sup>1</sup> Criterios para analizar la tabla  $[\mathbf{T}, \mathbf{Z}]$ , pág. 42

Para el segundo eje es diferente, los indicadores presentan *estructura específica* en el  $AFM(\mathbf{T}, \mathbf{Z})$  para el grupo de variables continuas.

**Ejes factoriales:** al observar el plano factorial para los dos métodos (figura 4.1 y 4.2), se observa que el primer factor es el mismo para los dos análisis. Esto se explica puesto que el primer factor del  $AFM(\mathbf{T}, \mathbf{Z})$  es un factor absolutamente común a los dos grupos. Al contrario, de los segundos factores de los dos métodos; el segundo factor del  $AFM(\mathbf{T}, \mathbf{Z})$  es un factor relacionado a las variables de *clima – habitat* que está muy poco relacionado con la distribución de las especies, y que no puede aparecer en el  $ACC(\mathbf{T}, \mathbf{Z})$ .

**Planos factoriales:** en el análisis de este conjunto de datos los planos factoriales de los dos métodos (figuras 4.1 y 4.2) son muy similares y permiten mas o menos las mismas conclusiones.

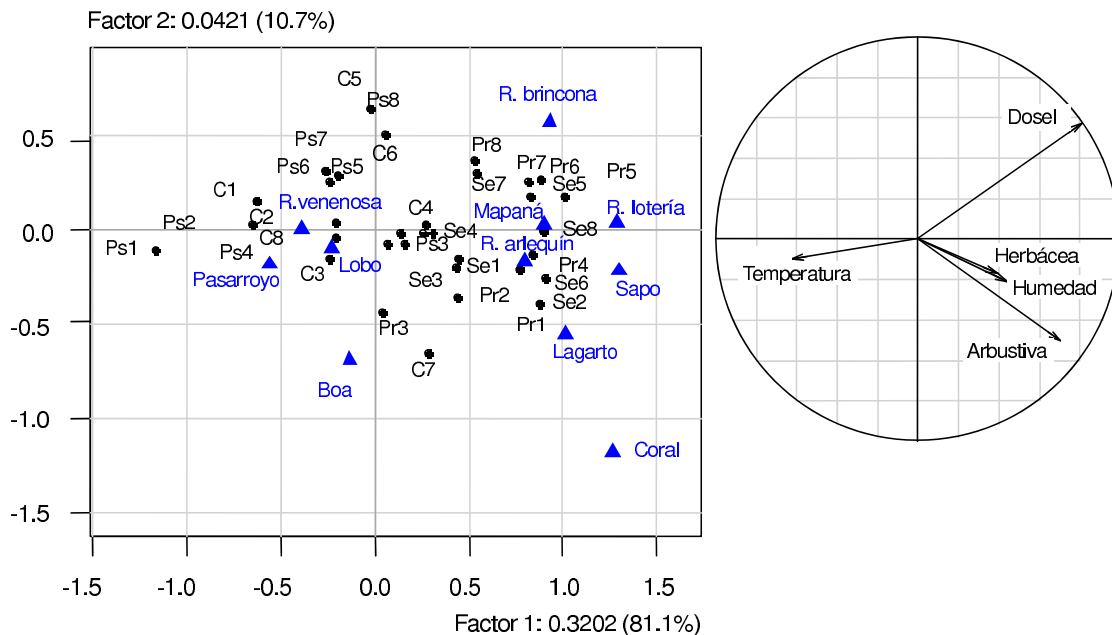


Figura 4.2: Plano Factorial 1-2 del ACC. Individuos, frecuencias y variables continuas

- Las secciones  $Ps1$  y  $Ps2$  son los que más contribuyen y mejor calidad presentan en el primer eje. Este primer eje en ambos métodos, enfrenta estos dos secciones ( $Ps1$  y  $Ps2$ ) con las secciones de la zona boscosa.
- El primer eje factorial esta altamente correlacionado con variables que pertenecen a los dos grupos, están  $R.brinconna$ ,  $R.venenosa$  y  $R.arlequin$  del grupo de especies; cobertura arbustiva y cobertura de dosel de las variables de *clima – habitat*. Las variables cobertura herbácea y cobertura arbustiva tienen correlación positiva en ambos métodos, similar a los resultados de la matriz de correlación, aunque se ve con mayor intensidad está correlación en el  $AFM(\mathbf{T}, \mathbf{Z})$ . El segundo eje tiene marcadas diferencias: las variables que más contribuyen son cobertura

arbustiva, cobertura herbácea y humedad en el  $AFM(\mathbf{T}, \mathbf{Z})$ ; y en el  $ACC(\mathbf{T}, \mathbf{Z})$  la frecuencia que más contribuye es R.brincona.

- La proyección conjunta de frecuencias e individuos permite observar aproximadamente dos centros de gravedad constituidos por las especies, en torno a los cuales se agrupan las secciones. En un lado está la especie R.brincona y R.arlequin que parecen ser los centros de gravedad de las secciones de bosques primarios y secundarios, en ambos métodos presentan alto porcentaje de cobertura de dosel y cobertura arbustiva y bajas temperaturas, difiere en la importancia que tiene la cobertura herbácea en el  $AFM(\mathbf{T}, \mathbf{Z})$ . En ese mismo gráfico se ha definido aproximadamente otro agrupamiento con la especie R.venenosa compuesto por algunas secciones de prisión y de cultivos.

# Capítulo 5

## Ejemplos de aplicación

Este capítulo tiene como objetivo presentar una guía metodológica para decidir cuándo aplicar  $AFM(\mathbf{T}, \mathbf{Z})$ ,  $ACC(\mathbf{T}, \mathbf{Z})$  o ambos a tablas de frecuencias-variables continuas descritas sobre el mismo conjunto de individuos, y realizar la ejecución práctica de los métodos en cada uno de ellos, utilizando para esto dos ejemplos de aplicación.

Los ejemplos de aplicación son en otras áreas diferentes a la ecología, ya que en investigación medio-ambiental se utiliza frecuentemente el análisis canónico de correspondencias (Chessel et al. 1987, Lebreton et al. 1988, Lebreton et al. 1991, Doledec & Chessel 1991, Birks & Austin 1994, Villalobos et al. 2000, Pavoine et al. 2003, Urbina & Londoño 2003, Sánchez-González & López-Mata 2003, Berti et al. 2004).

### 5.1. Primera aplicación: calidad de la educación media en Colombia en relación a indicadores socio-educativos

#### 5.1.1. Datos y objetivos del análisis

Los datos se muestran en la tabla 5.1: la tabla  $\mathbf{T}$  es la tabla de contingencia que clasifica los planteles educativos de Colombia de 23 departamentos (filas) y la calificación dada por el ICFES para cada plantel según los resultados de sus estudiantes en las pruebas de estado del 2007 (super: muy superior - superior, alta, media, baja, infer: inferior - muy inferior).

En la tabla  $\mathbf{Z}$  de variables continuas se tienen, para 23 departamentos colombianos, algunos indicadores socio-educativos (tasa de analfabetismo (%): analfab, Gasto promedio por alumno en el 2000 (transferencias \$): gasto.al00, Relación alumno-docente 2000: R.a.d02of, Coeficiente GINI 2004: GINI, Producto Interno Bruto percapita 2004: PIBperc, Necesidades Básicas Insatisfechas 2004 (%): NBI, y tasa de desempleo 2004 (%): desempleo).

En esta aplicación se pretende realizar un análisis descriptivo del comportamiento del sector educativo en Colombia, con énfasis en la educación media para el año 2007. El análisis está orientado por las siguientes preguntas:

Tabla 5.1: Datos de calidad e indicadores socio-educativos en los departamentos colombianos

	super	alta	media	baja	infer	analfab	gasto.al00	R.a.d02of	GINI	PIBperc	NBI	desempleo
Antioquia	118	127	298	456	108	6.4	570345	32.5	0.53	2.27	18.2	14.6
Atlántico	53	51	111	200	160	4.7	616850	27.9	0.49	1.67	17.5	14.1
Bolívar	34	26	59	184	135	9.6	481945	23.5	0.48	8.59	31.2	9.5
Bogotá	362	254	506	245	12	1.9	1259490	31.9	0.56	0.20	7.8	14.8
Boyacá	32	36	162	85	20	9.4	1188044	23.3	0.59	2.44	27.0	14.0
Caldas	24	21	87	103	17	7.2	948517	25.2	0.52	1.75	16.5	15.5
Caquetá	3	7	29	34	14	10.5	722412	24.1	0.52	4.14	26.6	10.7
Cauca	19	27	84	121	56	11.5	605402	23.8	0.53	0.36	28.2	8.9
Cesar	20	31	64	65	34	14.1	609747	23.1	0.46	1.36	35.2	7.4
Córdoba	14	15	65	128	59	17.1	506982	28.2	0.57	1.20	45.2	14.8
Cundinamarca	75	80	238	238	36	5.6	898589	25.4	0.51	0.82	20.5	13.8
Huila	20	21	91	88	17	6.4	730861	27.5	0.55	0.32	23.3	17.4
La Guajira	9	11	14	44	43	12.7	599260	28.3	0.41	3.02	32.1	8.0
Magdalena	9	11	28	128	108	11.1	534853	21.6	0.47	0.75	39.6	6.7
Meta	16	15	56	80	14	7.1	600392	28.4	0.50	1.78	22.7	10.2
Narino	30	55	123	92	35	8.6	747744	21.6	0.53	0.89	27.7	10.1
NSantander	23	22	82	113	41	10.0	782557	25.0	0.44	1.07	23.9	14.7
Quindio	13	18	41	45	8	5.7	927606	27.9	0.56	2.48	17.8	20.2
Risaralda	20	27	59	71	5	6.3	825089	26.1	0.49	0.70	16.8	15.7
Santander	71	73	148	143	22	7.7	873211	24.5	0.50	2.48	12.7	14.9
Sucre	7	13	41	87	45	15.5	524165	25.7	0.46	0.80	40.5	8.1
Tolima	22	30	118	149	39	10.4	859283	26.1	0.52	1.56	24.0	16.9
Valle	118	111	243	335	116	5.0	683732	28.4	0.51	2.11	13.0	15.1

Fuente: ICFES, DANE, e Iregui et al. (2006)

1. Cómo es la tipología de departamentos desde el punto de vista de *calidad de la educación* y desde el punto de vista de *indicadores socio-educativos*?
2. La distribución de *calidad educativa* en los diferentes *departamentos*, depende de los *indicadores socio-educativos*?

### 5.1.2. Análisis factorial múltiple (AFM)

Este conjunto de datos es interesante desde un punto de vista metodológico: la similitud entre los grupos de variables justifica el análisis simultáneo; las diferencias entre los grupos son suficientemente importantes para justificar la utilización de un método específico que ponga de relieve los rasgos comunes y los rasgos específicos.

#### Análisis separados

En este ejemplo, la inercia total asociada al ACS de la tabla de Calidad es 0.173, los dos primeros valores propios se destacan sobre los demás y retienen el 94.8% de está inercia (tabla de inercia de los ACP separados, figura 5.1). El primer eje ordena a los departamentos colombianos según el perfil de sus planteles educativos por la calificación del ICFES a partir de las pruebas de estado del 2007.

Al realizar el ACP ponderado de indicadores socio-educativos, los dos primeros ejes retienen una inercia del 63.9% (tabla de inercia de los ACP separados, figura 5.1), el primer eje separa departamentos (Bogotá, Quindio) que tienen bajos porcentajes de analfabetismo y necesidades básicas insatisfechas con departamentos (Magdalena, Sucre y Cesar) que tienen altos porcentajes de estos indicadores; el segundo eje separa el departamento de Bolivar que tiene PIB percapita alto de Boyacá y Córdoba que tienen bajo este indicador.

La inercia y el primer valor propio del análisis de correspondencias simples de la tabla de calidad educativa ( $Inercia_{ACS} = 0.173$ ,  $\mu_1 = 0.138$ ), son menores que los del ACP ponderado de la

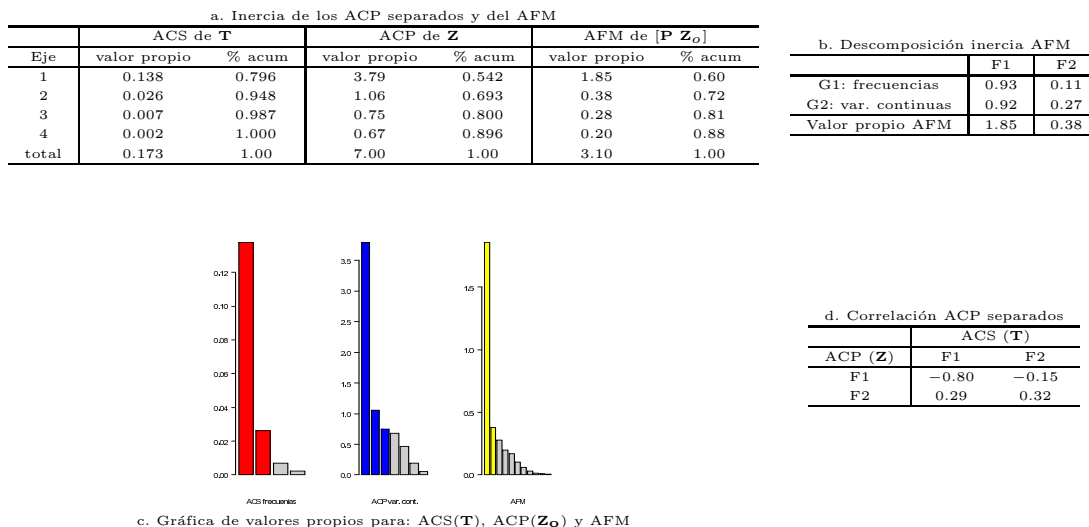


Figura 5.1: Resultados para el análisis parcial y global del AFM. Calidad Educativa e indicadores socio-educativos

tabla de indicadores socio-educativos ( $Inercia_{ACP} = 7$ ,  $\nu_1 = 3.79$ ) (ver figura 5.1). Equilibrar la contribución de los dos grupos de variables es útil para evitar la dominación de las variables continuas en la construcción del primer eje.

Los valores propios para los grupos separados (figura 5.1, parte a. y d.), muestran que los dos grupos tienen una primera dirección de inercia dominante, el primer plano factorial en cada uno de ellos explica más del 65% de variabilidad. Además, las correlaciones entre los factores de los ACP individuales muestra que los factores homólogos (iguales) están correlacionados unos con otros ( $F1_{(ACS-ACP)} = -0.80$  y  $F2_{(ACS-ACP)} = 0.32$ ).

### Detección de estructuras comunes

Algunos indicadores (ver criterios para analizar la tabla  $[\mathbf{T} \mathbf{Z}]$ , en la página 42) ponen de manifiesto la *estructura común* que tienen los grupos en el primer eje:

- *Correlaciones entre los factores parciales de cada grupo y el factor global del AFM (figura 5.2):* el primer factor del  $AFM(\mathbf{T}, \mathbf{Z})$ , compromiso entre los dos factores de rango 1 de los dos análisis separados está correlacionado con el primer factor de cada análisis separado ( $-0.95$  frecuencias,  $0.95$  variables continuas), la correlación con las *columnas-frecuencias* que más contribuyen al primer eje del ACS de Calidad educativa (Infer, Super) son las mismas en el AFM. Y, las variables relacionadas a indicadores socio-educativos que más contribuyen al primer eje del ACP ponderado son las que más contribuyen al primer eje del  $AFM(\mathbf{T}, \mathbf{Z})$ , pero intercambiadas en ese eje. El segundo factor del AFM está más relacionado con indicadores socio-educativos. Así, el primer plano factorial proporcionado por el  $AFM(\mathbf{T}, \mathbf{Z})$  es similar al de cada análisis separado, invirtiendo el primer eje del ACP ponderado de indicadores socio-educativos.

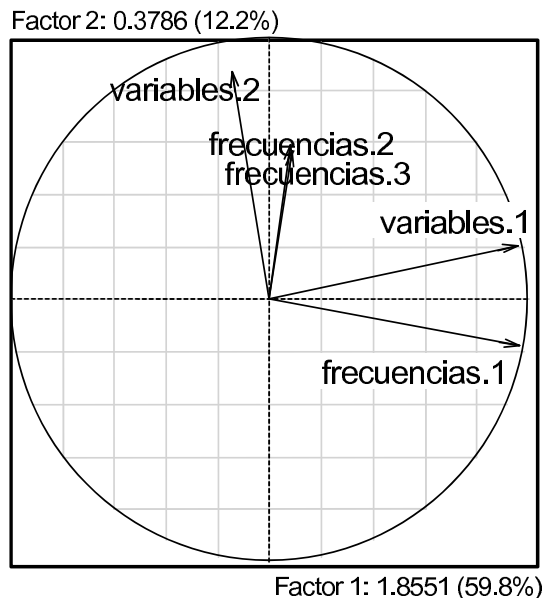


Figura 5.2: Plano factorial 1-2 en el AFM. Factores parciales

- *Medidas de asociación entre las tablas  $\mathbf{T}$  y  $\mathbf{Z}$* : el coeficiente  $RV$  de relación entre grupos es 0.70, lo que pone de manifiesto la notable similitud que las dos tablas analizadas mantienen en términos generales. La matriz  $Lg$ , de relación entre grupos, muestra igual dimensionalidad para los dos grupos (calidad educativa,  $Lg = 1.5$ ; indicadores socio-educativos,  $Lg = 1.1$ ).
- *Inercia del AFM( $\mathbf{T}, \mathbf{Z}$ )*: la inercia total de la nube de departamentos y de variables en el  $AFM(\mathbf{T}, \mathbf{Z})$  es 3.10, la inercia del grupo de Calidad educativa pasa de 0.173 a 1.26, mientras que la inercia del grupo de Indicadores de educación cae de 7 a 1.85, se equilibra la contribución de los dos grupos.
- *Contribución de los grupos a la formación de los ejes*: la inercia del primer factor ( $\gamma_1 = 1.85$ ) es cercano al máximo ( $2 =$  número de grupos en este caso), la descomposición de la inercia del primer factor para cada uno de los grupos son cercanas al valor máximo ( $= 1$ ), lo que indica la existencia de una estructura común a los grupos de variables, por lo tanto, los dos grupos activos contribuyen de forma similar a la formación del primer eje. La situación es diferente para el segundo factor ( $\gamma_2 = 0.378$ ), ambos grupos tienen una baja contribución (ver tabla 5.2).

Tabla 5.2: Coordenadas y ayudas a la interpretación de los grupos activos

Grupos			coordenadas		contribuciones		Cos <sup>2</sup>	
	p.rel.	Disto	F1	F2	F1	F2	F1	F2
g1: frecuencias	0.5	0.02	0.93	0.11	50.3	28.9	43.24	0.61
g2: var. continuas	0.5	16.78	0.92	0.27	49.7	71.1	0.05	0.004

En conclusión, los dos grupos de variables tienen *estructura común* en el primer eje y *estructura específica* para el segundo eje, por lo tanto, para determinar la relación existente entre ellas se

completa el análisis descriptivo realizando el análisis canónico de correspondencias para la tabla  $[\mathbf{T} \ \mathbf{Z}]$ , por creer que la calidad educativa de los departamentos depende de indicadores socio-educativos.

### 5.1.3. Análisis canónico de correspondencias (ACC)

La inercia del  $ACC(\mathbf{T}, \mathbf{Z})$  es 0.138 (tabla 5.3). Esto significa, que el 79.8 % (0.138/0.173) de la inercia total del ACS de calidad educativa es explicada por indicadores socio-educativos.

La decisión para saber cuántos ejes es conveniente analizar en el  $ACC(\mathbf{T}, \mathbf{Z})$  está soportada en los valores propios (Tabla 5.3), en este caso se decide utilizar los dos primeros ejes para la tipología de los departamentos con respecto a la calidad educativa (87.1 % ( $\lambda_1 = 0.120$ ) para el eje 1; 9.6 % ( $\lambda_2 = 0.013$ ) para el eje 2) por que acumulan el 96.7 % de la inercia total. El primer eje del  $ACC(\mathbf{T}, \mathbf{Z})$  explica el 87.0 % (0.12/0.138) de la inercia proyectada por el mismo eje en el análisis de correspondencias simples de calidad educativa, indicando que las variables continuas relacionadas con este factor explican las frecuencias satisfactoriamente.

Tabla 5.3: Resultados del  $ACS(\mathbf{T})$  y del  $ACC(\mathbf{T}, \mathbf{Z})$

Método	ACS de $\mathbf{T}$		ACC de $(\mathbf{T}, \mathbf{Z})$		
Ejes	Inercia ( $\mu_s$ )	Acum.	Inercia ( $\lambda_s$ )	% Acum.	$\lambda_s/(\mu_s)$
1	0.138	0.138	0.1200	87.1	87.0
2	0.026	0.164	0.0130	96.7	50.0
3	0.007	0.171	0.0040	99.6	53.6
4	0.002	0.173	0.0006	100.0	26.2
Total	0.173		0.138		

El resultado de la estadística *Pseudo - F* es de 3.96 para el primer eje canónico del ACC, con un  $P_{value} = 0.005$ , significando que la tabla de calidad educativa y la tabla de indicadores socio-educativos tienen relación lineal al nivel del 0.5 %. Para la aplicación, la distribución de calidad educativa en el  $ACS(\mathbf{T})$  y en el  $ACC(\mathbf{T}, \mathbf{Z})$  son muy similares.

Para observar las relaciones entre calidad educativa e indicadores socio-educativos en departamentos colombianos se utiliza el biplot y el círculo de correlaciones (figura 5.3).

**Variables socioeconómicas:** las variables que presentan mayor asociación con respecto a la distribución de calidad educativa en los departamentos colombianos para el eje 1 del  $ACC(\mathbf{T}, \mathbf{Z})$  son transferencias por alumno, necesidades básicas insatisfechas, analfabetismo y coeficiente GINI, en el segundo eje son desempleo y analfabetismo. Las variables necesidades básicas insatisfechas y analfabetismo presentan correlación positiva entre ellas.

**Departamentos:** el primer eje, se interpreta como la contraposición de departamentos de la costa caribe (Magdalena, Sucre, Bolívar, Atlántico) asociados a tasas de analfabetismo y necesidades básicas insatisfechas altas como también bajas transferencias, con Bogotá que presenta transferencias altas. En el segundo factor se presentan departamentos como Tolima, Cauca, Huila y Risaralda asociados a altas tasas de analfabetismo y necesidades básicas insatisfechas.

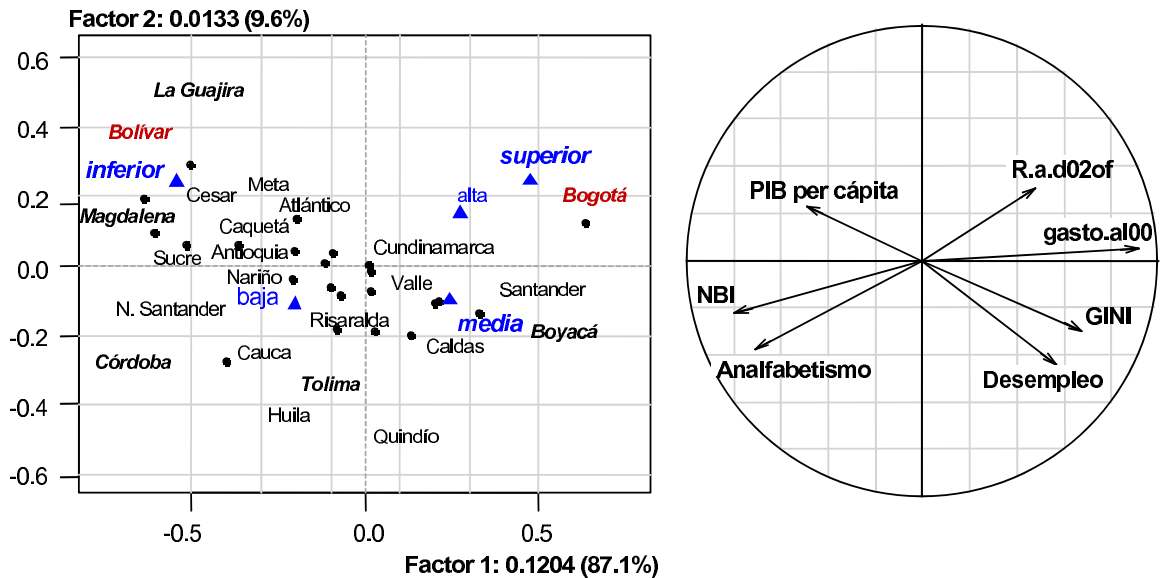


Figura 5.3: Plano Factorial 1-2 del ACC. Individuos, frecuencias y variables continuas

**Lectura simultánea:** en los departamentos que presentan tasas de analfabetismo y necesidades básicas insatisfechas altas como también bajas transferencias, su calidad educativa no es favorable (baja e infer), y Bogotá presenta tasas de analfabetismo y necesidades básicas insatisfechas bajas como también transferencias altas tiene calidad educativa favorable (media, alto y super). En la lectura del primer eje, hay un ordenamiento de los departamentos por calificación del ICFES (superior, alta, media, baja, inferior)

## 5.2. Segunda aplicación: estudio de Mortalidad en edades prematuras en comunidades autónomas de España

### 5.2.1. Datos y objetivo del análisis

Se utilizan para la tabla de frecuencias **T**, los datos de mortalidad del año 2005 suministrada por la Eurostat, correspondiente a los adultos con muertes prematuras (entre 35 y 64 años). Estos datos son calculados a partir de la información sobre las tasas de mortalidad estandarizada para cada una de las comunidades autónomas de España. No se tienen en cuenta las regiones de Ceuta y Melilla. Las causas de mortalidad se encuentran codificadas en la tabla 5.4

La tabla **Z** de variables continuas cruza las mismas comunidades autónomas (filas) y variables relacionadas a aspectos socioeconómicos (columnas) analizadas como variables suplementarias en un estudio de mortalidad por Bécue et al. (2003): Producto Interno Bruto “PIB” (millones de \$), Tasa de desempleo “Desempleo” (%), Titulados (%), Analfabetismo (%), Hacinamiento (%). La información recolectada de la tabla de variables continuas es del año 2004.

Tabla 5.4: nombre y codificación de causas de mortalidad

Nº	Descripción	Código masculino	Código femenino
1	Neoplasmas		
	Cáncer de estómago	m1CEsto	f1CEsto
	Cáncer de colon	m1CCol	f1CCol
	Cáncer de pulmón	m1CPulm	f1CPulm
	Cáncer de pancreas	m1CPanc	f1CPanc
	Cáncer de esófago	m1CEsof	
	Cáncer de hígado	m1CHig	f1CPulm
	Cáncer de boca	m1CBoca	
	Cáncer de pecho		f1CSeno
	Cáncer de utero		f1CUteroO
	Cáncer de Ovario		f1COvario
2	Enfermedades inmunológicas		
	SIDA	m1Sida	
3	Enfermedades de la sangre y órganos de formación de la sangre	m1CLinfH	f1CLinfH
4	Enfermedades del sistema circulatorio		
	Enfermedad isquémica del corazón	m1IsqC	f1IsqC
	Otras enfermedades del corazón	m1OtrC	f1OtrC
	Enfermedades cerebro vasculares	m1CerVasc	f1CerVasc
5	Enfermedades del Sistema Respiratorio		
	Enfermedad pulmonar obstructiva crónica	m1EPulmO	
6	Enfermedades del sistema Digestivo		
	Enfermedad crónica del hígado	m1ECHig	f1ECHig
7	Heridas y envenenamientos		
	Lesiones por accidentes de trafico	m1AcTra	f1AcTra
	Suicidio	m1Suic	f1Suic

La diversidad de las regiones de España (industrialización, nivel económico, condiciones climáticas, costumbres dietéticas, etc.) entraña una desigualdad frente a la mortalidad que interesa conocer tanto para establecer la planificación de las compañías de seguro como para mejorar la política sanitaria (Bécue et al. 2003).

El análisis de este ejemplo está orientado por las siguientes preguntas:

1. ¿Cuales son las comunidades autónomas que globalmente, es decir desde el punto de vista de las causas de mortalidad y de las variables socioeconómicas, se parecen, si intervienen igualmente las variables de ambos grupos?
2. ¿Qué comunidades autónomas en particular cuyo perfil de causas de mortalidad no corresponden al de las variables socioeconómicas?
3. La distribución de causas de mortalidad en las diferentes comunidades autónomas de España, está relacionado a indicadores socioeconómicos?

### 5.2.2. Análisis factorial múltiple (AFM)

#### Análisis separados

El siguiente análisis se hace a partir de la tabla 5.5 parte (a): La inercia y el primer valor propio del análisis de correspondencias simples de causas de mortalidad ( $Inercia_{ACS} = 0.035$ ,  $\mu_1 = 0.0084$ ),

son mucho más pequeños que los del ACP normado ponderado del grupo de variables continuas ( $Inercia_{ACP} = 5, \nu_1 = 3.5$ ). La tabla de valores propios para los grupos por individual, muestra que los dos grupos tienen una primera dirección de inercia dominante. En el primer eje, hay un mayor porcentaje de inercia explicado por el grupo de variables socioeconómicas (69.3 %) que por el grupo de mortalidad por causas (23.9 %); el primer plano factorial en cada uno de ellos explica más del 40 %.

Tabla 5.5: Resultados del análisis parcial y global en el AFM de la segunda aplicación

a. Inercia de los ACP separados y del AFM							b. Descomposición inercia AFM		
Eje	ACS de $\mathbf{T}$		ACP de $\mathbf{Z}$		AFM de $[\mathbf{P} \mathbf{Z}_o]$			F1	F2
	valor propio*1000	% acum	valor propio	% acum	valor propio	% acum			
1	8.4	23.9	3.5	69.3	1.70	30.3	G1: frecuencias	0.79	0.99
2	7.8	46.2	0.9	88.2	1.04	48.8	G2: var. continuas	0.91	0.05
3	4.4	58.7	0.4	95.3	0.60	59.4	Valor propio AFM	1.70	1.04
4	3.6	68.8	0.2	98.4	0.47	67.8			
total	35.3	100.0	5.0	100.0	5.62	100.0			

### Detección de estructuras comunes o específicas

Algunos indicadores iniciales ponen de manifiesto la *estructura común* para los el primer eje y específica del segundo eje:

- *Correlaciones entre los factores parciales de cada grupo y el factor global del AFM*: el primer factor del  $AFM(\mathbf{T}, \mathbf{Z})$  se confunde con el primer factor del análisis individual del grupo de variables socioeconómicas y el segundo factor del análisis individual del grupo de causas de mortalidad, las correlaciones entre los primeros factores parciales de cada grupo y el primer factor global fue  $-0.07$  para el grupo de frecuencias y  $0.95$  para el grupo de variables socioeconómicas. El segundo factor global está relacionado con el primer factor del análisis de correspondencias simples del grupo de causas de mortalidad ( $0.995$ ).
- *Contribución de los grupos a la formación de los ejes*: los dos grupos contribuyen de forma similar a la formación del primer eje (figura 5.5 parte b.), indica que el primer factor corresponde a una dirección de inercia común a los dos grupos. Las coordenadas a lo largo del segundo eje muestran que el segundo factor se debe principalmente al grupo de mortalidad por causas (frecuencias:  $0.99$  y no debido a las variables continuas:  $0.05$ ).
- *Medidas de asociación entre las tablas  $\mathbf{T}$  y  $\mathbf{Z}$* : el coeficiente  $RV$  de relación entre grupos es  $0.42$ , lo que manifiesta una similitud media entre los dos grupos en términos generales. La matriz  $Lg$ , de relación entre grupos, en el sentido del  $AFM(\mathbf{T}, \mathbf{Z})$  muestra menor dimensionalidad para el grupo de variables socioeconómicas ( $Lg = 1.2$ ) que para el grupo de mortalidad por causas ( $Lg = 3.4$ ).
- *Inercia del  $AFM(\mathbf{T}, \mathbf{Z})$* : La inercia del primer factor ( $\gamma_1 = 1.70$ ) concluye que este factor es la dirección principal de dispersión de las dos nubes (que por tanto se confunden), y representa una dirección de dispersión común para ambos grupos.

### Representación superpuesta de los individuos descritos por cada grupo por separado

Cualesquiera sea el conjunto de variables considerado, el primer factor del AFM opone las comunidades autónomas de Madrid, Navarra y País Vasco con Andalucía, C. Mancha y Extremadura (ver figura 5.4).

La representación superpuesta permite una comparación precisa de los dos grupos, el primer factor del AFM está correlacionado con los primeros factores de cada uno de los análisis separados, la representación superpuesta da una buena idea de la representación obtenida en los análisis separados (Pagès 2004).

En conclusión, los dos grupos de variables tienen estructura común en el primer eje y estructura específica para el segundo eje, por lo tanto, para determinar la relación existente entre ellas se completa el análisis descriptivo realizando el análisis canónico de correspondencias para la tabla  $[T \ Z]$ , por creer que la mortalidad por causas de las comunidades autónomas depende de indicadores socioeconómicos.

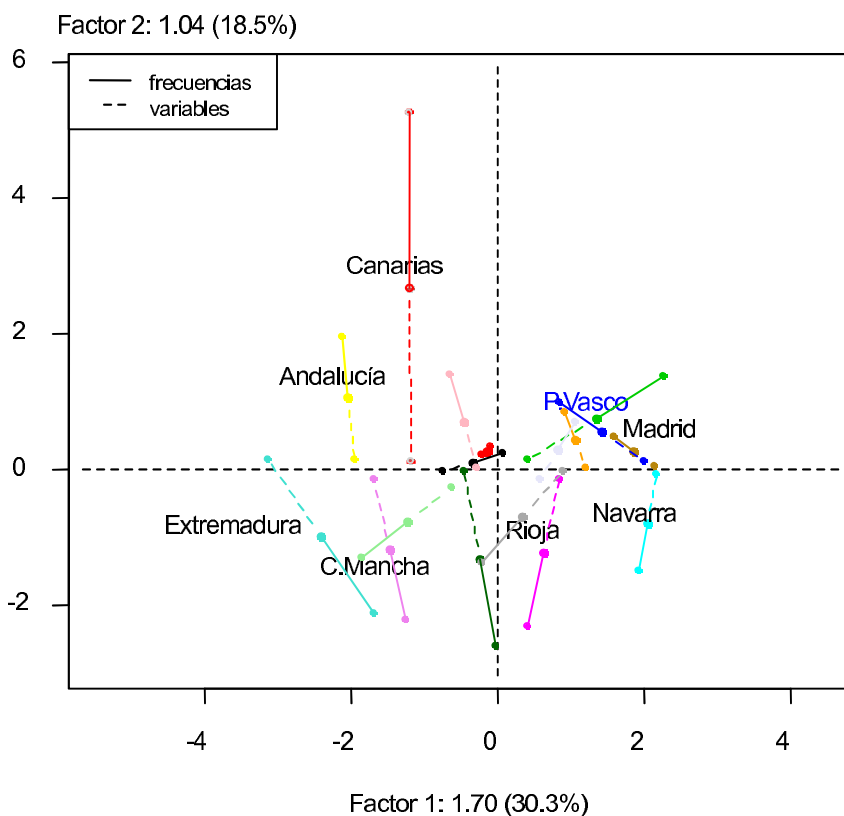


Figura 5.4: Plano factorial 1-2 en el AFM. Algunas comunidades autónomas. Puntos medios y puntos parciales.

### 5.2.3. Análisis canónico de correspondencias (ACC)

La inercia total asociada al  $ACS(\mathbf{T})$  de mortalidad prematura por causas es  $0.0353$  (figura 5.5).

Las causas de mortalidad prematura y las variables socioeconómicas tienen una relación lineal significativa (estadística  $Pseudo-F = 1.66$ ,  $p-value = 0.017$ ). Esto significa, que la mortalidad por causas es explicada por indicadores socioeconómicos.

El 43% ( $0.0152/0.0353$ ) de la inercia total del  $ACS(\mathbf{T})$  es explicada por los indicadores socioeconómicos tomados en esta aplicación. El primer eje del  $ACC(\mathbf{T}, \mathbf{Z})$  representa el 71.2% ( $0.00601/0.00844$ ) de la inercia proyectada por el mismo eje del análisis de correspondencias simples de la mortalidad prematura, indicando que las variables socioeconómicas relacionadas con este factor explican las causas por mortalidad prematura satisfactoriamente.

El primer valor propio del  $ACC(\mathbf{T}, \mathbf{Z})$  vale  $\lambda_1 = 0.00601$  (39.5% de la inercia total). La secuencia de los valores propios sugiere la presencia de tres ejes interpretables. Estos tres ejes acumulan el 80.5% de la inercia, suficiente para resumir la información de la relación entre las causas de mortalidad prematura y las variables socioeconómicas.

Para observar las relaciones entre mortalidad por causas e indicadores socioeconómicos en comunidades autónomas de España se utiliza el biplot y el círculo de correlaciones (figura 5.5).

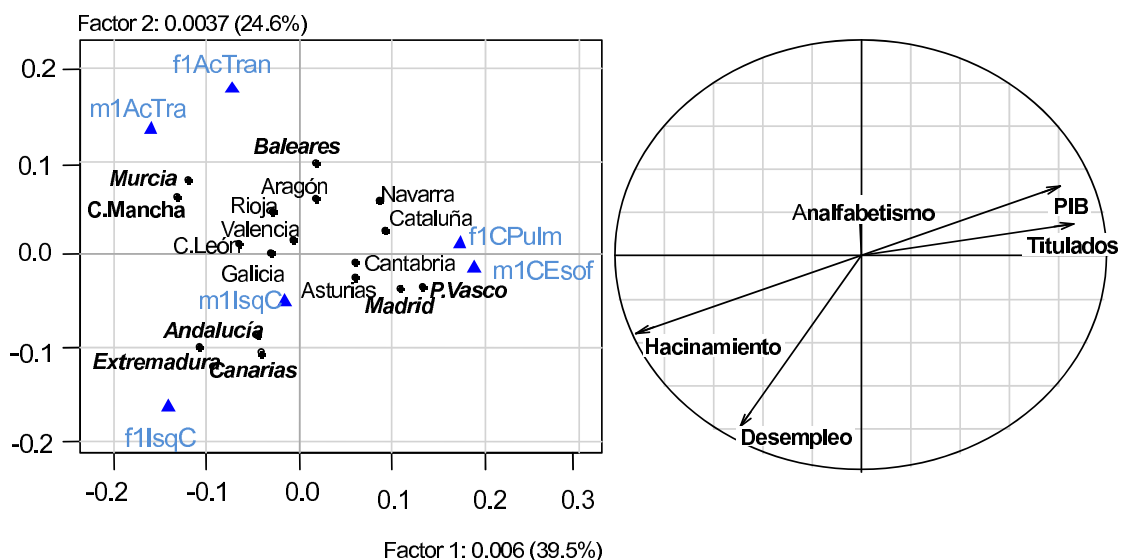


Figura 5.5: Plano Factorial 1-2 del ACC. Individuos, frecuencias y variables continuas

Se observa que el primer eje retiene una primera dirección de inercia importante, retiene el 39.5% de la inercia total, se interpreta como la contraposición de las comunidades de Murcia, C. Mancha y Extremadura con las comunidades de Madrid y País Vasco.

El primer eje, tiene una correlación positiva con el índice de hacinamiento (0.92) y menos fuerte con el índice de desempleo (0.49) y analfabetismo (0.85); presenta correlación negativa con el porcentaje de diplomados superiores sobre la población de egresados del sistema escolar en los últimos 10 años (-0.86), y con el Producto Interno Bruto percapita (-0.81).

Las comunidades de C. Mancha, Extremadura y Murcia tienen condiciones de hacinamiento altos, porcentaje de diplomados superiores sobre la población de egresados del sistema escolar en los últimos 10 años y Producto Interno Bruto percapita bajos; presenta mortalidad prematura en hombres por accidentes de tránsito, mientras que, Madrid y País Vasco presenta condiciones socioeconómicas con mucha más frecuencia mortalidad en hombres por cancer de esófago y en mujeres con cáncer de pulmón.

El segundo eje, explica el 24.6 % de la inercia total del  $ACC(\mathbf{T}, \mathbf{Z})$ . Este eje, esta relacionado con comunidades como Extremadura, Andalucía y Canarias tienen condiciones de desempleo altos, presentan mortalidad prematura tanto en hombres como mujeres por isquemias cardíacas, y enfermedad crónica del hígado en hombres. En la comunidad de Baleares tiene condiciones de desempleo bajo, presenta mortalidad prematura por accidentes de tránsito tanto en hombres como mujeres.

El tercer eje ( $\lambda_3 = 0.00249$ , 16.4 % de la inercia total), pone en evidencia rasgos específicos de algunas regiones, como por ejemplo, la alta incidencia de cáncer de hígado en hombres en las comunidades de Cataluña y Baleares; y la alta incidencia sólo en hombres por muertes de cáncer de esófago o accidentes de tránsito en Asturias. Este eje presenta correlación negativa no muy fuerte con el PIB percapita (-0.29) y con el porcentaje de diplomados superiores sobre la población de egresados del sistema escolar en los últimos 10 años (0.21).

### 5.3. Guía de análisis

Para el análisis descriptivo multivariado de una tabla  $[\mathbf{T} \ \mathbf{Z}]$  por los métodos factoriales estudiados, se presenta a continuación una guía de como hacerlo.

- En primera instancia para determinar *estructuras comunes o específicas* se realiza el análisis parcial de cada grupo de variables y la detección de estructuras comunes ofrecidas por el método  $AFM(\mathbf{T}, \mathbf{Z})$ . La distribución de los individuos en cada eje pueden ser similares en los métodos factoriales estudiados cuando los grupos de variables están relacionados o tienen estructuras comunes.

*Análisis parcial de los grupos.* ¿Qué porcentaje de inercia es recogida por el primer (segundo) eje factorial y por el primer plano factorial para cada uno de los grupos?, ¿como es la correlación entre los factores parciales de cada grupo?

*Detección de estructuras comunes.*

*Correlaciones entre los factores parciales de cada grupo y los factores globales del  $AFM(\mathbf{T}, \mathbf{Z})$ :* ¿El factor global de orden  $s$  está próximo a cada uno de los factores de los grupos?, ¿Sobre qué factores globales están bien representados los primeros factores de los grupos?, ¿Cómo es la correlación entre los factores parciales de cada grupo y los factores globales del  $AFM(\mathbf{T}, \mathbf{Z})$ ?

*Coordenadas y ayudas a la interpretación de los grupos:* ¿Los dos grupos contribuyen de forma similar a la formación del primer (segundo) eje factorial del  $AFM(\mathbf{T}, \mathbf{Z})$ ?, ¿cuantas dimensiones intervienen de manera significativa en el análisis global del  $AFM(\mathbf{T}, \mathbf{Z})$ ?

*Medidas de asociación entre las tablas  $\mathbf{T}$  y  $\mathbf{Z}$ :* ¿Son similares los dos grupos a partir del coeficiente RV?, ¿Son similares los dos grupos a partir del coeficiente Lg?, ¿Los anteriores indicadores ponen de manifiesto una estructura interna similar entre los dos grupos?

*Inercia total y valores propios del análisis global:* ¿Qué porcentaje de la inercia total es recogida por el primer (segundo) eje factorial y por el primer plano factorial?, ¿el primer valor propio está cercano al número de grupos (2)?, ¿al hacer la descomposición de la inercia del primer eje para cada grupo, las inercias de las variables de cada uno de los grupos están próximas del valor máximo 1, o el primer factor está muy relacionada a uno de los grupos?

- Si se encuentran *estructuras comunes* con el  $AFM(\mathbf{T}, \mathbf{Z})$  y se tiene conocimiento que el grupo de frecuencias es explicado por el grupo de variables continuas (acercamiento no simétrico) se debe realizar un análisis más fino con el método factorial análisis canónico de correspondencias  $ACC(\mathbf{T}, \mathbf{Z})$  para determinar las posibles relaciones entre las frecuencias y las variables continuas. A continuación se presenta la guía a seguir con el método análisis canónico de correspondencias.

*Inercia total y valores propios.* ¿qué porcentaje de la inercia total es recogida por el primer (segundo) eje factorial y por el primer plano factorial?, ¿cuántos ejes factoriales considera razonable interpretar?, ¿Cuánto es el porcentaje de inercia explicado por las frecuencias en el ACC, por las frecuencias en el análisis de correspondencias simples, por la relación de las frecuencias y las variables continuas en el ACC?, ¿La prueba de permutación de Montecarlo permite corroborar la existencia de una relación significativa entre las frecuencias y las variables continuas para los individuos?

*Tipología de los individuos y principales factores de variabilidad.*

**Variables continuas:** ¿Puede decirse que es coherente la representación de las variables continuas en el círculo de correlación con la lectura de la matriz de correlación de las variables continuas?, ¿Qué variables continuas se puede decir que están más altamente correlacionadas con el primer factor?, ¿Puede identificar subconjuntos de variables altamente correlacionadas entre sí?

**Frecuencias:** ¿Cuáles son las categorías del grupo de frecuencias que más contribuyen y mejor calidad de representación al primer plano factorial?, ¿Cuáles son sus coordenadas y cuáles sus pesos relativos?, ¿cuál es la categoría que está más mal representada en el primer plano factorial y puede decirse que está muy mal representado? ¿A partir de estos resultados, puede darle algún significado a este primer factor?. Se puede hacer un análisis similar con el segundo factor.

**Individuos:** ¿Cuáles son los individuos que más contribuyen y mejor calidad de representación al primer eje (plano) factorial?, ¿Cuáles son sus coordenadas y sus pesos relativos?, ¿Cuáles son los individuos más (menos) distanciados entre sí?, ¿De qué manera estos resultados son útiles para ayudar a la caracterización del primer eje (plano) factorial, al primer factorial?

**Lectura simultánea:** Los análisis anteriores sugieren que pueden constituirse grupos de individuos caracterizados por las frecuencias en términos de las variables continuas?, ¿Podría sugerir grupos?

- Si se encuentran *estructuras comunes* con el  $AFM(\mathbf{T}, \mathbf{Z})$  y no se tiene conocimiento de dependencia, o, se encuentren *estructuras específicas* entre grupos se sigue con el análisis global que ofrece el análisis factorial múltiple,  $AFM(\mathbf{T}, \mathbf{Z})$ .

*Inercia total y valores propios del análisis global.* ¿Qué porcentaje de la inercia total es recogida por el primer (segundo) eje factorial y por el primer plano factorial?, ¿cuántos ejes factoriales considera razonable interpretar?.

*Tipología de los individuos y principales factores de variabilidad*

**Variables.** Para *frecuencias y variables continuas* ¿Cuáles son las categorías (variables) del grupo de frecuencias (variables continuas) que más contribuyen al primer plano factorial?, ¿Cuáles son sus coordenadas, pesos relativos y qué tan bien representadas están estas categorías (variables continuas) en el primer plano factorial?, ¿Puede identificar subconjuntos de variables altamente correlacionadas entre sí?, ¿A partir de los dos grupos (frecuencias, variables continuas), puede darle algún significado al primer (segundo) factor?.

**Individuos.** ¿Cuáles son los individuos que más contribuyen al primer eje factorial, al primer plano factorial?, ¿Cuáles son sus coordenadas y sus pesos relativos?, ¿Qué tan bien representados se encuentran estos individuos en el primer eje factorial, al primer plano factorial?, ¿Cuáles son los individuos más distanciados entre sí?, ¿Cuáles son los individuos menos distanciados entre sí?, ¿De qué manera estos resultados son útiles para ayudar a la caracterización del primer eje factorial, al primer plano factorial?.

**Lectura simultánea.** Los análisis anteriores sugieren que pueden constituirse grupos de individuos caracterizados por las variables?, ¿Podría sugerir algunos grupos?

*Representación superpuesta.* ¿Cuales son los individuos que globalmente, es decir desde el punto de vista del grupo de frecuencias y del grupo de variables continuas, se parecen, si intervienen igualmente las variables del grupo de frecuencias y las variables del grupo de variables continuas?, ¿Cuales individuos se asemejan por el grupo de frecuencias, cuáles se asemejan por el grupo de variables continuas?, ¿Qué individuos en particular cuyo perfil del grupo de frecuencias no corresponde al del grupo de variables continuas?.

## Software

Para encontrar los elementos del  $ACC(\mathbf{T}, \mathbf{Z})$  se utilizaron las siguientes funciones: *cca* del módulo *ade4* (Thioulouse et al. 1997), que hace el ACC como un ACPVI (Rao (1964) citado por Dray (2003)); *planfac* del paquete *FactoClass* (Pardo & DelCampo 2007) para realizar el *biplot* con escalamiento tipo 2, el cuál recibe un objeto *dudi* y produce un plano factorial similar a los del paquete *FactoMineR* (Husson et al. 2007) o a los de *ade4* (Chessel et al. 2005); *plot(cca)* para el gráfico triplot y *anova(cca(fre,var))* para la prueba de permutación Montecarlo, ambas del modulo *vegan* (Oksanen et al. 2007). En el análisis del  $AFM(\mathbf{T}, \mathbf{Z})$ , se utilizo la función *mfa* del módulo *ade4* (Chessel et al. 2005), teniendo en cuenta la ponderación de las filas de los ACP ponderados individuales en la opción *ktab.list.dudi* que es el objeto del *mfa* y también se programo el AFM de la tabla de frecuencias-variables continuas con la función *as.dudi*; para las ayudas a la interpretación en ambos métodos se utilizo la función *dudi.tex* del paquete *FactoClass* (Pardo & DelCampo 2007).

# Conclusiones

- Los métodos  $ACC(\mathbf{T}, \mathbf{Z})$  y  $AFM(\mathbf{T}, \mathbf{Z})$  aplicados a la tabla  $[\mathbf{T} \ \mathbf{Z}]$  son complementarios, sí el grupo de frecuencias depende del grupo de variables continuas.

**Primero:** realizar un  $AFM(\mathbf{T}, \mathbf{Z})$ , y aplicar los *criterios* para analizar la tabla  $[\mathbf{T} \ \mathbf{Z}]$  para determinar *estructuras comunes*, que se muestran en la página 42.

**Segundo:** si se encuentran *estructuras comunes* se debe realizar un  $ACC(\mathbf{T}, \mathbf{Z})$  para describir la dependencia entre las frecuencias y las variables continuas.

- Si la naturaleza de los datos permite elegir entre ambos métodos, el objetivo del estudio resulta ser el criterio más apropiado para la elección.

**Si la trayectoria de cada individuo es de gran interés:** esto es, si desea analizar el comportamiento de cada individuo tanto en el comportamiento medio, como el correspondiente a cada una de las situaciones consideradas (frecuencias, variables continuas), el  $AFM(\mathbf{T}, \mathbf{Z})$  es la técnica a utilizar.

**Si el interés se centra en la distribución de las frecuencias:** y su posible relación al grupo de variables continuas el  $ACC(\mathbf{T}, \mathbf{Z})$  es la técnica a utilizar.

- Aplicaciones en la que la información este estructurada en dos grupos de variables sobre un mismo conjunto de individuos, y las frecuencias no dependan de las variables continuas se realizará un  $AFM(\mathbf{T}, \mathbf{Z})$  para detectar *estructuras comunes* o *específicas*. El conocimiento de la dependencia se fundamenta en el contexto conceptual de la aplicación.

# Recomendaciones

- El  $ACC(\mathbf{T}, \mathbf{Z})$  se puede aplicar a otras disciplinas diferentes a la Ecología y afines.
- Para analizar tablas de frecuencias-variables continuas, y se tenga conocimiento de dependencia entre las frecuencias-variables continuas, y el grupo de variables continuas sea heterogéneo (diferentes temáticas) se recomienda:

Un análisis combinado entre el AFM y ACC; a las variables continuas agrupadas por temáticas hacerle en primera instancia un AFM, y después realizar el  $ACC(\mathbf{T}, \mathbf{Z})$  aplicado a tablas de frecuencias-variables continuas.

# Bibliografía

- Abdessemed, L. & Escofier, B. (1992), Generalisation de l'analyse factorielle multiple a l'etude des tableaux de frequence et comparaison avec l'analyse canonique des correspondances, Technical Report 688, INRIA.
- Bécue, M., Pagès, J., Álvarez, R. & Hernández, M. (2003), 'Análisis factorial múltiple para tablas de contingencia: Estudio de la mortalidad en las comunidades autónomas de España', *Congreso Nacional de Estadística e Investigación Operativa*.
- Berti, J., Gutierrez, A. & Zimmerman, R. (2004), 'Relaciones entre tipo de hábitat, algunas variables químicas y la presencia de larvas de *Anopheles aquasalis curry* y *Anopheles pseudopunctipennis theobald* en un área costera del Estado Sucre, Venezuela', *Entomotrópica* **2**(2), 14–30.
- Birks, H. & Austin, H. (1994), An annotated bibliography of canonical correspondence analysis and related constrained ordination methods (1986-1991), Technical report, Botanical Institute, Norway. All-Gaten 41, N-5007 Bergen, Bunch, K.J., Heneghan.
- Chessel, D., Dufour, A., Dray, S., Lobry, C., Ollier, S., Pavoine, S. & Thioulouse, J. (2005), *ade4: Analysis of Environmental Data Exploratory and Euclidean methods in Environmental sciences*. R package version 1.4-0.  
\*<http://pbil.univ-lyon1.fr/ADE-4>
- Chessel, D., Lebreton, J. & Yoccoz, N. (1987), 'Propriétés de l'analyse canonique des correspondances; une illustration en hydrobiologie', *Revue Statistique Appliquée* **35**(4), 55–72.
- Díaz, M. (2002), 'Preferencias alimentarias como alternativa al estudio de patrón dietético', *Rev. Esp. Nutr. Comunitaria* **8**(1-2), 29–34.
- Doledec, S. & Chessel, D. (1991), 'Recent developments in linear ordination methods for environmental sciences', *Advances in Ecology* **1**, 133–155. India.
- Dray, S. (2003), Eléments d'interface entre analyses multivariées, systèmes d'information géographique et observations écologiques, PhD thesis, Université Claude Bernard - Lyon 1.
- Escofier, B. & Pagès, J. (1984), L'analyse factorielle multiple: une méthode de comparaison de groupes de variables, in E. Diday, ed., 'Data Analysis and Informatics, III', Elsevier Science, Amsterdam, pp. 41–56. Proceedings of the Third International Symposium on Data Analysis and Informatics.

- Escofier, B. & Pagès, J. (1992), *Análisis factoriales simples y múltiples. Objetivos, métodos e interpretación*, Universidad del País Vasco, Bilbao.
- Grabiel, K. R. (1971), 'The biplot graphic display of matrices with application to principal component analysis', *Biometrika* **58**, 453–467.
- Greenacre, M. (2007), *Correspondence Analysis in Practice*, 2 edn, Chapman & Hall/CRC.
- Husson, F., Lê, S. & Mazet, J. (2007), *FactoMineR: Factor Analysis and Data Mining with R*. R package version 1.05.  
\*<http://factominer.free.fr>, Mailing list: [http://www.agrocampus-rennes.fr/math/Encoding latin1](http://www.agrocampus-rennes.fr/math/Encoding%20latin1)
- Iregui, A., Melo, L. & Ramos, J. (2006), *Evaluación y análisis de eficiencia de la educación en Colombia*, Banco de la República, Bogotá.
- Lebart, L., Morineau, A. & Piron, M. (1995), *Statistique exploratoire multidimensionnelle*, Dunod, Paris.
- Lebreton, J., Chessel, D., Prodon, R. & Yoccoz, N. (1988), 'L'analyse des relations espèces-milieu par l'analyse canonique des correspondances; i.- variables de milieu quantitatives', *Acta Ecológica* **9**(1), 53–67.
- Lebreton, J., Sabatier, R., Banco, G. & Bacou, A. (1991), 'Principal component and correspondence analyses with respect to instrumental variables. an overview of their role in studies of structure-activity and species-environment relationships', *Applied Multivariate Analysis in SAR and Enviromental studies* pp. 85–114.
- Oksanen, J., Kindt, R., Legendre, P., B., O. & Stevens, M. (2007), *VEGAN: Community Ecology Package*. vegan package version 1.8-8, suggests MASS, mgcv, lattice, cluster, scatterplot3d, rgl, ellipse.  
\*<http://cran.r-project.org/>, Mailing list: [//r-forge.r-project.org/projects/vegan/](mailto://r-forge.r-project.org/projects/vegan/)
- Pagès, J. (1996), 'Eléments de comparaison entre l'analyse factorielle multiple et la méthode STATIS', *Revue de Statistique Appliquée* **44**(4), 81–95.
- Pagès, J. (2004), 'Multiple factor analysis: Main features and application to sensory data', *Revista Colombiana de Estadística* **27**(4), 1–98.
- Pardo, C. & DelCampo, P. (2007), 'Combinacion de metodos factoriales y de analisis de conglomerados en r: el paquete factoclass', *Revista Colombiana de Estadística* **30**(2), 235–245.
- Pavoine, S., Dufour, A. & Chessel, D. (2003), Canonical correspondence analysis, a standard in ecology, in M.Greenacre & E. J. Blasius, eds, 'CARME 2003: International Conference on Correspondence Analysis and Related Methods', pp. 63–64.  
\*<http://pbil.univ-lyon1.fr/R/articles/arti112.pdf>
- R Development Core Team (2009), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.  
\*<http://www.R-project.org>

- Rao, C. (1964), 'The use and interpretation of principal component analysis in applied research', *Sankhya* **26**(1), 329–359.
- Sabatier, R., Lebreton, J. & Chessel, D. (1989), Principal Component Analysis with Instrumental Variables as a Tool for Modelling Composition Data, in R. Coppi & S. Bolasco, eds, 'Multiway Data Analysis', Elsevier, Amsterdam, pp. 341–350.
- Sánchez-González, A. & López-Mata, L. (2003), 'Clasificación y ordenación de la vegetación del norte de la Sierra Nevada, a lo largo de un gradiente altitudinal', *Anales del Instituto de Biología, Universidad Nacional Autónoma de México, serie Botánica* **74**(1), 47–71.
- Ter-Braak, C. (1986), 'Canonical correspondence analysis: A new technique for multivariate direct gradient analysis', *Ecology* **67**(5).
- Ter-Braak, C. & Smilauer, P. (2002), *CANOCO, Reference Manual and CANOCO-DRAW for Windows User's guide: Software for Canonical Community Ordination, versión 4.5*, Microcomputer Power, Ithaca, NY, USA.
- Thioulouse, J., Chessel, D., Dolédec, S. & Olivier, J. (1997), 'ADE-4: a multivariate analysis and graphical display software', *Stat. Comp.* **7**, 75–83.  
\*<http://pbil.univ-lyon1.fr/ADE-4/ADE-4F.html>
- Ulate-Montero, G. & Fernández-Ramírez, A. (2001), 'Relaciones del perfil lipídico con variables dietéticas, antropométricas, bioquímicas, y otros factores de riesgo cardiovascular en estudiantes universitarios', *Acta méd. costarric.* **43**(2), 70–76. ISSN 0001-6002.
- Urbina, J. & Londoño, M. (2003), 'Distribución de la comunidad de herpetofauna asociada a cuatro áreas con diferente grado de perturbación en la isla gorgona, pacífico colombiano', *Rev. Acad. Colom. Cienc.* **27**(102), 105–112.
- Villalobos, F., Ortíz-Pulido, R., Moreno, C., Pavón-Hernández, N., Hernández-Trejo, H., Bello, J. & Montiel, S. (2000), 'Patrones de la macrofauna edáfica en un cultivo de *Zea maiz* durante la fase postcosecha en "La Mancha", Veracruz, México', *Acta Zoo. Méx.* **80**, 167–183.

# Apéndice

## Apéndice A. Código en R para el AFM y el ACC

```
# Paquetes utilizados
library(ade4)
library(xtable)
library(FactoClass)
library(FactoMineR)

# Tabla de datos del grupo de frecuencias
frecuencias<-read.table("biolo.txt",header=TRUE); frecuencias

# ACS para el grupo de frecuencias por el paquete ade4
acs<-dudi.coa(frecuencias,scannf=F,nf=5); acs

# Plano factorial del ACS de frecuencias con el paquete FactoClass
planfac(acs, cex.row = 0.8, cex.col = 0.6)

# Ayudas a la interpretación
dudi.tex(acs,job="herpetofauna")

# Tabla de datos para el grupo de variables continuas
variables<-read.table("vargor.txt",header=TRUE); variables

# ACP para el grupo de variables continuas por el paquete ade4
acp<-dudi.pca(variables,acs$lw,scannf=F,nf=5); acp

# Gráficas del ACP del grupo de variables con el paquete ade4
s.corcircle(acp$co)
windows()
planfac(acp,Tcol=FALSE)

# Matriz de correlación de las variables continuas
cor(acp$tab)

# Ayudas a la interpretación
```

```

dudi.tex(acp,job="variables")

# Análisis factorial múltiple de la tabla frecuencias-variables continuas
ktabgor<-ktab.list.dudi(list(acs,acp),tabnames=c("frecuencias","variables"))
afm<-mfa(ktabgor,scannf=F); afm

# Correlaciones entre los factores parciales
cor(acs$li,acp$li)

# Correlación entre los factores globales del AFM y los factores parciales de
# los análisis individuales y gráfica
afm$T4comp; s.corcircle(afm$T4comp)

# Correlación entre los factores globales y parciales en el AFM
cor(afm$li,acp$li); cor(afm$li,acs$li)

# Análisis factorial múltiple por el FactoMineR
AFM.1<-MFA(cbind(acs$tab,acp$tab),group=c(11,5),name.group=c("frecuencias","variables"))

# Gráfica superpuesta de individuos
plot(AFM.1, choix = "ind", partial="all")

# Análisis factorial múltiple para la tabla frecuencias-v.continuas programada
# con la función as.dudi
afm.acp<-as.dudi(afm$tab,c(acs$cw/acs$eig[1],acp$cw/acp$eig[1]),acs$lw,
scannf=F,nf=5,c("afm"),c("coa")); afm.acp

# Plano factorial de individuos y columnas-frecuencias en el AFM
biplot(afm$li[,1:2],afm$co[1:11,1:2],cex=c(0.7,0.7),col=c("darkblue","black"),xlab="F1",
ylab="F2",main="AFM",las=1,abline(h = 0, v = 0, reg = NULL, lty=2.1,lwd=1.5))

# Gráficas del AFM para la tabla (T,Z) con planfac

# individuos, columnas-frecuencias
planfac(afm.acp)
planfac(afm.acp,Tcol=FALSE)
points(afm.acp$co[1:11,])
text(afm.acp$co[1:11,],rownames(afm.acp$co[1:11,]),1)
planfac(afm.acp)

# variables continuas
s.corcircle(afm.acp$co[12:16,])

# Ayudas a la interpretación en el AFM
dudi.tex(afm.acp,job="calidad.afm")

# Coeficientes Lg y RV
# Coeficiente Lg frecuencias

```

```

fre.lg<-as.matrix(acs$tab/((acs$eig[1])^2))%*%as.matrix(diag(acs$cw))
%*%t(as.matrix(acs$tab))%*%as.matrix(diag(acs$lw))%*%as.matrix(acs$tab)
%*%as.matrix(diag(acs$cw))%*%t(as.matrix(acs$tab))%*%as.matrix(diag(acs$lw))
Lg.fre<-sum(diag(fre.lg)); Lg.fre

# Coeficiente Lg variables continuas
var.lg<-as.matrix(acp$tab/((acp$eig[1])^2))%*%t(as.matrix(acp$tab))%*%as.matrix(diag(acs$lw))
%*%as.matrix(acp$tab)%*%t(as.matrix(acp$tab))%*%as.matrix(diag(acs$lw))
Lg.var<-sum(diag(var.lg)); Lg.var

# Coeficiente Lg frecuencias-variables continuas
frevar.Lg<-as.matrix(acs$tab/acs$eig[1])%*%as.matrix(diag(acs$cw))%*%t(as.matrix(acs$tab))
%*%as.matrix(diag(acs$lw))%*%as.matrix(acp$tab)%*%t(as.matrix(acp$tab/acp$eig[1]))
%*%as.matrix(diag(acs$lw))
Lg.frevar<-sum(diag(frevar.Lg)); Lg.frevar

# Coeficiente RV para frecuencias
fre.lg1<-as.matrix(acs$tab)%*%as.matrix(diag(acs$cw))%*%t(as.matrix(acs$tab))
%*%as.matrix(diag(acs$lw))%*%as.matrix(acs$tab)%*%as.matrix(diag(acs$cw))
%*%t(as.matrix(acs$tab))%*%as.matrix(diag(acs$lw))
Lg.fre1<-sum(diag(fre.lg1)); Lg.fre1
fre.RV<-Lg.fre1/(sum(acs$eig^2)); fre.RV

# Coeficiente RV para var.continuas
var.lg1<-as.matrix(acp$tab)%*%t(as.matrix(acp$tab))%*%as.matrix(diag(acs$lw))
%*%as.matrix(acp$tab)%*%t(as.matrix(acp$tab))%*%as.matrix(diag(acs$lw))
Lg.var1<-sum(diag(var.lg1)); Lg.var1
var.RV<-Lg.var1/(sum(acp$eig^2)); var.RV

# Coeficiente RV para frecuencias-var.continuas
frevar.Lg1<-as.matrix(acs$tab)%*%as.matrix(diag(acs$cw))%*%t(as.matrix(acs$tab))
%*%as.matrix(diag(acs$lw))%*%as.matrix(acp$tab)%*%t(as.matrix(acp$tab))%*%as.matrix(diag(acs$lw))
Lg.frevar1<-sum(diag(frevar.Lg1)); Lg.frevar1
frevar.RV<-Lg.frevar1/(sqrt(sum(acp$eig^2))*sqrt(sum(acs$eig^2))); frevar.RV

# Gráfica de valores propios para el análisis separado y valores propios del análisis global del AFM
par(mfrow=c(1,3))
barplot(acs$eig,col = c(rep("red", 2), rep(grey(0.8), 11)),las=1,pch=19,xlab="ACS frecuencias")
barplot(acp$eig,col = c(rep("blue", 2), rep(grey(0.8), 5)),las=1,pch=19,xlab="ACP var. cont.")
barplot(afm$eig,col = c(rep("yellow", 4), rep(grey(0.8), 16)),las=1,pch=19,xlab="AFM")

# Análisis canónico de correspondencias en el paquete ade4
acc<-cca(frecuencias, variablesc,scannf=F,nf=5); acc

# Análisis canónico de correspondencias programado con la función as.dudi
acc1<-as.dudi(acc$tab,acs$cw,acs$lw,scannf=FALSE,nf=2,c("acc"),c("coa")); acc1

# Plano factorial del ACC para frecuencias x individuos

```

```
planfac(acc1)
planfac(acc1,Trow=FALSE)
points(acc$l1)
text(acc$l1,rownames(acc$l1),2)

# Circulo de correlaciones del grupo de variables continuas en el ACC
s.corcircle(acc$cor)

# Coordenadas factoriales de las v. continuas en el ACC
acc$cor

# Ayudas a la interpretación
dudi.tex(acc,job="ACC")

# Librería Vegan
library(vegan)
# Prueba de permutación Montecarlo en ACC
anova(cca(frecuencias,variablesc))
# Gráfico triplot del ACC en el paquete vegan
plot(cca(frecuencias,variablesc))

# Presentación de los datos en el programa Latex
datos<-cbind(frecuencias, variablesc);datos

datosx<-xtable(datos); datosx
```