TRABAJO PRÁCTICO
Parte III

Ejercicio 1: Para obtener una única caracterización agronómica incluyendo toda la información brindada en los archivos CUALITATIVAS y CUANTITATIVAS vistos en el TP Parte I, una alternativa válida es trabajar con el Coeficiente General de Similaridad de Gower que permite considerar variables cuantitativas y cualitativas en forma simultánea.

a) Calcule a mano este coeficiente para medir el grado de asociación entre las variedades A_1 y A_2, y F_4 y F_5 (filas 1, 2, 12 y 13 respectivamente).

b) Recurra al software para encontrar la matriz de similaridad de Gower entre todos los individuos.

c) Aplique Análisis de Coordenadas Principales para obtener la configuración conjunta cuali-cuantitativa de las variedades de pepino

d) Halle un cluster con encadenamiento UPGMA en función de la distancia de Gower

e) Compare resultados de (c) con los hallados en ejercicio 4 de Parte I. Cuantifique la relación con la configuración de consenso (bidimensional) encontrada con APG

Ejercicio 2: Con el objetivo de tipificar los establecimientos frutícolas del Alto Valle de Río Negro en función de factores socio-económicos, tecnológicos y productivos, se seleccionó una muestra aleatoria de 86 chacras de la región. Recolectándose información acerca a 40 puntos de una encuesta frutícola que corresponden tanto a variables dicotómicas, cualitativas como cuantitativas. Además, se presenta el inconveniente de contar con información faltante, hecho que complica la utilización de la mayoría de las estrategias clásicas de caracterización. Por ello se decidió aplicar el coeficiente de similaridad de Gower que por otro lado permite asignar diferente importancia a los caracteres a través de ponderaciones, alternativa válida para eliminar pesos espurios en encuestas donde sobre una misma temática se emplea un mayor número de preguntas respecto a otras.

A continuación, se listan las variables relevadas clasificándolas según la característica o ítem que se deseaba estudiar:

1. VARIABLES RELACIONADAS A LA ESTUCTURA DE CHACRA (E)
   1.1 SUPERFICIE
   1.2 REGIMEN DE TENENCIA (P: propietario, D: arrendatario, A: aparcería)
   1.3 % DE LA SUP CON CULTIVO DE MANZANA
   1.4 % DE LA SUP CON CULTIVO DE PERA
   1.5 % DE LA SUP CON CULTIVO DE CAROZO
   1.6 % DE LA SUP CON OTROS CULTIVOS
   1.7 % DE LA SUP SIN CULTIVO
   1.8 EDAD MEDIA DE LOS MONTES DE PERA Y MANZANA
   1.9 CANTIDAD TOTAL DE MANO DE OBRA PERMANENTE Y TEMPORARIA
2. VARIABLES RELACIONADAS A LA COMERCIALIZACION (C)
   2.1 EMPAQUE PROPIO (No, Si)
   2.2 COMERCIALIZACION PROPIA (No, Nacional, Internacional, Ambas) *
   2.3 INDUSTRIALIZACIÓN PROPIA: (No, Si)
3. VARIABLES RELACIONADAS A SITUACIÓN LABORAL (L)
   3.1 MANO DE OBRA FAMILIAR (0: nadie, 1: productor, 2: productor y familia, familia) *
   3.2 MESES HOMBRE DEL PRODUCTOR EN EL ESTABLECIMIENTO

3.3 MESES HOMBRE DEL PRODUCTOR FUERA DEL ESTABLECIMIENTO
3.4 ACTIVIDAD DEL PRODUCTOR FUERA DEL PREDIO (Frutícola: C. Comerciante, T. Contratista, X. Otros. No Frutícola: M. Comerciante, P. Profesional, E. Empleado, O. Otros. N: ninguna)
3.5 MANO DE OBRA NO FAMILIAR PERMANENTE POR HECTÁREA
3.6 MANO DE OBRA NO FAMILIAR TEMPORARIA POR HECTÁREA
3.7 % MANO OBRA PERMANENTE RESPECTO MANO DE OBRA TOTAL (NO FAMILIAR)

4. VARIABLES RELACIONADAS CON EL PARQUE DE MAQUINARIA (M)
4.1 NUMERO DE TRACTORES
4.2 EDAD MEDIA DE TRACTORES (modelo)
4.3 POTENCIA MEDIA DE TRACTORES
4.4 NUMERO DE PULVERIZADORAS
4.5 EDAD MEDIA DE PULVERIZADORAS
4.6 CAPACIDAD MEDIA DE PULVERIZADORAS
4.7 NUMERO DE TRACTOELEVADORES
4.8 EDAD MEDIA DE LOS TRACTOELEVADORES

5. VARIABLES RELACIONADAS CON LA TECNOLOGÍA PRODUCTIVA (T)
5.1 DENSIDAD MEDIA DE LOS CUADROS DE PERAS Y MANZANAS
5.2 % DE LA SUP CON CONDUCCIÓN LIBRE
5.3 % DE LA SUP CON CONDUCCIÓN EN ESPALDERA
5.4 % DE LA SUP SIN DEFENSA CONTRA HELADAS
5.5 % DE LA SUP CON DEFENSA CONTRA HELADAS POR ASPERSIÓN
5.6 % DE LA SUP CON DEFENSA CONTRA HELADAS POR CALEFACCIÓN
5.7 % DE LA SUP CON OTROS SISTEMAS DE LUCHA CONTRA HELADAS
5.8 % DE LA SUP CON PLANTACIONES JÓVENES
5.9 % DE LA SUP CON VARIEDADES 'NUEVAS'

6. VARIABLES NETAMENTE PRODUCTIVAS (P)
6.1 RENDIMIENTO MEDIO DE LOS CUADROS DE RED DELICIOUS
6.2 % DE LA SUP EN ESTADO BUENO
6.3 % DE LA SUP EN ESTADO REGULAR
6.4 % DE LA SUP EN ESTADO MALO

a) Aplique el coeficiente de similaridad de Gower para cuantificar la semejanza entre unidades productivas considerando que se debe asignar igual importancia (1/6) a cada grupo de variables. Tenga en cuenta además que las variables 5.2 y 5.3 miden un único concepto que es el tipo de conducción, las variables 5.4, 5.5, 5.6 y 5.7 también cuantifican una sola característica que es el tipo de defensa contra heladas, y por último las variables 6.2, 6.3 y 6.4 miden el estado de la plantación (se podría haber puesto una multinomial con niveles Bueno, Regular y Malo, pero se perdería información)
b) Sobre la matriz de similaridad hallada aplique Análisis de Coordenadas Principales y conforme grupos de explotaciones agrícolas según la primer coordenada y subgrupos en función de la segunda coordenada.
c) Caracterice grupos y subgrupos en forma descriptiva y somera en función de las variables originales
d) Verifique si encuentra esos grupos en un cluster UPGMA realizado a partir de la matriz de similaridad de Gower

REEB, P. D.; BRAMARDI, S. J.; ALVAREZ, O. "Tipificacion de unidades productivas del Alto Valle de Río Negro basada en el coeficiente de similaridad de Gower". V Congreso Latinoamericano de Sociedades de Estadística, Buenos Aires, 28-30 de Octubre 2002. Resúmenes pág. 112. Trabajo completo en Actas en CD.

Ejercicio 3: El archivo PANOJA contiene datos provenientes de la evaluación morfológica-fisiológica de planta y panoja de 125 poblaciones nativas de maíz del Banco de Germoplasma de la EEA INTA Pergamino. Los caracteres evaluados corresponden a 9 variables cuantitativas y 2 cualitativas que se describen a continuación:

DFLORMASC: Días a la floración masculina
DFLORFEM: Días a la floración femenina
ALTPLA: Altura de la planta (cm)

ALTMAZ: Altura de la mazorca (cm)
NROHJAS: Número total de hojas
NRHOJARR: Número de hojas arriba de la mazorca
ANCHHOJMAZ: Ancho de la hoja de la mazorca (cm)
LARGOHOJMAZ: Largo de la hoja de la mazorca (cm)
DIAMTALLO: Diámetro del tallo (mm)
COLTALL: Color del tallo (1 Verde, 2 Púrpura diluido, 3 Púrpura, 4 Rojo sol diluido, 7 Rojo)
POSHOJ: Posición de las hojas (1 Semivolcada, 2 Normal, 3 Semierguida)

El objetivo del trabajo más que caracterizar y clasificar a los individuos fue describir la asociación entre caracteres cuantitativos y cualitativos.

a) Aplicando una medida de distancia y/o similaridad adecuada a ambos conjuntos de variables cuantificar el grado de asociación entre los caracteres cualitativos y cuantitativos.

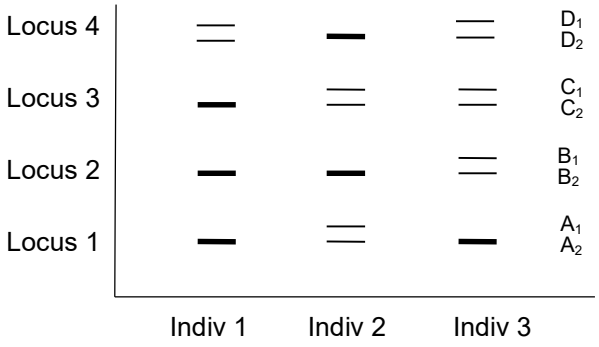
b) Recurra a la discretización de Escofier sobre las variables cuantitativas para poder aplicar Análisis de Correspondencias Múltiples y en función de los resultados obtenidos concluya sobre la relación entre los distintos niveles de las variables cualitativas y las variables cuantitativas.

BRAMARDI, S.J.; REEB, P.; DE BERNARDIN, F.; TASILLE, V.; FERRER, M. (2006) Codificación de Escofier: una 'discretización' sin pérdida de información. VII Congreso Latinoamericano de Sociedades de Estadística. Rosario, Resúmenes pág. 40. Actas en CD, págs. 12. ISBN 650-673-564-6

Ejercicio 4: Supongamos el siguiente ejemplo donde tenemos tres individuos diploides bialélicos donde el genotipo correspondiente a cuatro loci es el siguiente:

Individuo	Locus1	Locus2	Locus3	Locus4
1	A_2A_2	B_2B_2	C_2C_2	D_1D_2
2	A_1A_2	B_2B_2	C_1C_2	D_2D_2
3	A_2A_2	B_1B_2	C_1C_2	D_1D_2

Si a estos individuos le aplicamos un marcador molecular codominante que pudiera describirnos la información contenida en cada locus quedaría el siguiente patrón de bandas:



a) Podría en función de esta información determinar un gradiente de similaridad entre los tres individuos?

b) Con el fin de utilizar un coeficiente de similaridad para cuantificar la semejanza entre individuos como codificaría el resultado obtenido del patrón de bandas ?

c) Utilizando de los coeficientes de similaridad SM, Jaccard, Dice y Russel-Rao cuantifique las similaridades entre los individuos 1-2 y 2-3.

d) Concuerda lo hallado en el inciso (c) respecto a lo esperado en (a) ? Comente al respecto.

e) Como sería el patrón de bandas si hubiese utilizado un marcador molecular dominante ? (tomemos a la alternativa alélica 1 como dominante sobre la 2)

f) Como codificaría esta información para poder utilizar como medida de semejanza un coeficiente de similaridad ?
g) Verificar resultados con R

Ejercicio 5: En el siguiente ejemplo más sencillo repetir un análisis como en el ejercicio anterior, pero comparando las distancias entre los individuos 1-2 y 1-3 utilizando distancias genéticas (Nei, Cuerda, Rogers y Prevosti).

Genotipo:

Individuo	Locus1	Locus2
1	A_1A_1	A_1A_1
2	A_1A_2	A_1A_2
3	A_2A_2	A_1A_1

Frecuencias relativas asociadas de alelos A_1 y A_2 por locus:

Individuo	Locus1		Locus2	
1	1	0	1	0
2	0.5	0.5	0.5	0.5
3	0	1	1	0

Ejercicio 6: A continuación, se presentan los resultados de la simulación del genotipo de individuos diploides. Se supuso que la población era apareada al azar con respecto a cada locus bialélico y lo suficientemente grande como para ignorar consanguinidad. Bajo estos supuestos los genotipos para cada locus seguirán el equilibrio de Hardy-Weinberg.

INDIV	LOC1	LOC2	LOC3	LOC4	LOC5	LOC6	LOC7	LOC8	LOC9	LOC10
1	Aa	Aa	aa	aa	aa	AA	aa	aa	Aa	AA
2	AA	Aa	aa	aa	Aa	AA	Aa	aa	AA	aa
3	Aa	aa	aa	Aa	aa	aa	AA	Aa	aa	Aa
4	Aa	AA	AA	aa	Aa	aa	AA	aa	aa	Aa
5	AA	aa	aa	aa	Aa	Aa	aa	aa	Aa	aa
6	Aa	aa	aa	aa	Aa	Aa	Aa	Aa	AA	Aa
7	AA	Aa	aa	aa	Aa	AA	Aa	Aa	Aa	AA
8	Aa	Aa	aa	aa	AA	Aa	Aa	Aa	Aa	aa
9	aa	aa	aa	aa	aa	Aa	Aa	Aa	Aa	Aa
10	aa	Aa	aa	aa	Aa	AA	AA	aa	Aa	AA
11	Aa	Aa	aa	Aa	Aa	Aa	AA	AA	Aa	aa
12	Aa	Aa	aa	aa	AA	Aa	aa	aa	aa	Aa
13	AA	aa	Aa	aa	AA	AA	aa	Aa	Aa	Aa
14	AA	AA	aa	Aa	AA	aa	AA	Aa	Aa	AA
15	Aa	Aa	aa	aa	AA	AA	Aa	AA	aa	Aa

En las columnas encontramos los loci y en las filas los individuos.
a) Calcular las matrices de distancias y/o similaridades correspondientes a las medidas de asociación SM, Jaccard, Rao, Nei, Cuerda, Prevosti y Rogers, en caso de utilización de un marcador codominante (dos alelos alternativos, A_1 asociado con A y A_2 asociado con a)
b) Analizar la relación entre las distintas medidas de distancias/similaridad recurriendo a correlación entre matrices. Comentar al respecto y sacar conclusiones.
c) Repetir el ejercicio suponiendo la utilización de un marcador dominante.
d) Analizar para cada una de las distancias/similaridades estudiadas la correlación entre los resultados obtenidos para marcadores codominantes y dominantes.

Encontrará en el archivo COD2 y COD3 las codificaciones correspondientes a la aplicación de coeficientes de similaridad y distancias genéticas respectivamente para el caso de marcadores codominantes. Ídem con COD4 y COD5 para marcadores dominantes.

Ejercicio 7: Sobre 15 poblaciones locales de maíz (*Zea mays*) de las provincias de Buenos Aires y Santa Fe, más tres variedades mejoradas de polinización abierta (VM1, VM2 y VM3 las cuales se utilizaron como testigos) y una línea endocriada (LE), se realizó una caracterización molecular por la técnica de Microsatélites. Cada población estuvo representada por 25 individuos y se registró la frecuencia alélica absoluta para 6 microsatélites. En el archivo MICROSATELITES se encuentra la información donde en las filas la misma letra indica las alternativas alélicas para cada uno de los loci microsatélites. Este archivo se encuentra con el formato que fue entregado por el propietario de los datos.

a) Adecuar el archivo MICROSATELITES para poder trabajar sobre el con la sentencia dist.genpop

b) Hallar la distancia genética de Prevosti entre poblaciones con dist.genpoo.

c) Sobre la matriz de distancia hallada aplique Análisis de Coordenadas Principales para conformar grupos de poblaciones de Maíz según su semejanza (puede ayudarse haciendo un clúster UPGMA sobre las distancias euclídeas observadas en el plano principal).

d) Encontrar la distancia de Manhattan acotada y comparar con lo hallado en (b)

e) Podría aplicar el coeficiente de similaridad SM en este problema ? Porque ?

Ejercicio 8: El archivo LACAR contiene la caracterización de 29 sitios de la cuenca binacional Lacar - Hua Hum (San Martín de los Andes, Provincia de Neuquén) en función de la abundancia (en frecuencia) de 36 familias macroinvertebrados que sirven como bioindicadores de contaminación orgánica. Por otra parte, se registraron 9 variables fisicoquímicos in situ y de hábitat, para determinar la calidad ambiental del lugar. Estas variables fueron:

CE: conductividad eléctrica (µS/cm)	PH: nivel de acidez o alcalinidad
ALT: altitud (m.s.n.m.)	PRS: fósforo reactivo soluble (µg/l)
FE: hierro (mg/l)	SI: sílice (mg/l)
DU: dureza (mg/l)	ALC: alcalinidad (mg/l)

PEH: puntaje de evaluación del hábitat (a partir de las planillas de caracterización del hábitat de Barbour et al., 1999; UDGCA, 2012).

La segunda columna del archivo indica la subcuenca a la que pertenece cada sitio (L: Lacar , P: Pocahullo, TQ: Trabunco-Quitrahue).

a) Realizar un Análisis de Correspondencia para caracterizar los sitios en función de la abundancia de Lacar, macroinvertebrados y de Componentes Principales para ver la misma caracterización en función de variables ambientales. A través de la configuración de ambas caracterizaciones puede ver algún tipo de asociación entre las familias de macroinvertebrados y las variables ambientales ?

b) Cuantificar la relación entre los planos principales de las dos configuraciones halladas en (a)

c) Recurrir a AFM para caracterizar brevemente los sitios en función de la abundancia de macroinvertebrados y gradientes de las variables ambientales

d) Con Análisis de Redundancia profundizar sobre el estudio de asociación entre abundancia de familias de macroinvertebrados y variables ambientales. Relacionar (si es posible) con los grupos de sitios que venía estableciendo.

e) Y en cuanto a la subcuenca a la que pertenecen los sitios, encuentra algún patrón ?

f) Verificar, en base a la longitud del gradiente implícito en Y, si fue adecuado usar RDA

g) Repita lo realizado en (d) recurriendo a Análisis Canónico de Correspondencia y compare resultados