

APLICACIÓN DE TÉCNICAS ESTADÍSTICAS DE ANÁLISIS MULTIVARIANTE AL ESTUDIO DE VARIEDADES DE CASTAÑO

P. Álvarez Álvarez ¹, M. Barrio Anta ² y R. Cao Abad ³

¹ Departamento de Producción Vegetal. Escuela Politécnica Superior. Universidad de Santiago de Compostela. Campus Universitario s/n. 27002-LUGO (España). Correo electrónico: palvarez@lugo.usc.es

² Departamento de Ingeniería Agroforestal. Escuela Politécnica Superior. Universidad de Santiago de Compostela. Campus Universitario s/n. 27002-LUGO (España). Correo electrónico: barrio@lugo.usc.es

³ Departamento de Matemáticas. Universidad de A Coruña. Facultad de Informática Campus de Elviña s/n. 15071-A CORUÑA (España). Correo electrónico: rcao@udc.es

Resumen

El objetivo de este trabajo ha sido mostrar, mediante un ejemplo práctico, la aplicación de tres técnicas de análisis multivariante al estudio de tres variedades de castaño (*Castanea sativa*) valoradas por su calidad de fruto. Dichas técnicas buscan reducir el número de variables originales y realizar clasificaciones a partir de grupos homogéneos conocidos. Se ha demostrado la existencia de diferencias estadísticamente significativas entre los valores de distintas variables (de hojas, frutos y erizos) medidas en las tres variedades y se ha podido establecer una función discriminante que permite asignar una futura observación a alguna de las tres variedades con sólo medir las variables originales.

Palabras clave: *Castanea sativa*, Variedades, Análisis multivariante

INTRODUCCIÓN

El castaño aparece en el medio natural de dos formas: silvestre y domesticada. En Galicia la forma silvestre se presenta como monte alto, en mezcla con carballos (*Quercus robur*) y abedules (*Betula alba*), siendo raramente el castaño la especie dominante en la masa. Este tipo de bosque mixto se denomina “fraga” y puede incluir además a otras especies como *Prunus avium*, *Acer pseudoplatanus*, etc. Actualmente es mucho más abundante el castañar domesticado (souto), aunque en las últimas décadas los soutos se han ido abandonando en gran parte del territorio. Esta especie ha sido domesticada desde antiguo en Galicia mediante la selección y propagación de variedades con castañas de buena calidad, o que presenta-

ban buenas características combinadas para la obtención de madera y fruto (ÁLVAREZ et al., 2000). Actualmente están catalogadas multitud de variedades de castaño, siendo su conocimiento una herramienta importante para poder estudiar sus posibilidades productivas y de mejora.

MATERIAL Y METODOS

Datos empleados

Los datos empleados proceden de dos parcelas experimentales instaladas en masas de castaño del país (*Castanea sativa*) en el municipio de A Fonsagrada (Lugo). Ambas parcelas están separadas entre sí 10 kilómetros y en ellas existen únicamente árboles de tres variedades de castaño:

Pared (variedad 1), *Lemos* (variedad 2) y *Portuguesa* (variedad 3). En cada una de las parcelas se han medido las variables de 9 árboles (3 de cada una de las variedades). Las variables medidas estaban relacionadas con las dimensiones de las hojas, erizos y frutos. Se anotaron 40 mediciones por variable, lo que para el conjunto de la muestra ha supuesto un total de 5.575 mediciones. Las variables medidas relativas a las hojas han sido: su longitud (LH), su ancho (AH) y la longitud del pecíolo (PH). Las relativas al fruto: el ancho (AF), alto (HF) y longitud (LF). Las relativas al erizo: ancho (AE) y alto (HE). En la tabla 1 se muestran los estadísticos descriptivos de la muestra de datos empleados en el estudio.

Métodos estadísticos empleados

Al no haberse diseñado ningún dispositivo experimental para la toma de datos, cabe la posibilidad de aplicar distintos análisis entre los que se han elegido tres técnicas multivariantes: análisis de componentes principales (PCA), análisis factorial discriminante (FDA) y análisis discriminante (DA). También se ha realizado un estudio de correlación entre las 8 variables estudiadas para examinar el grado de dependencia entre ellas. La finalidad de estos análisis es doble: por un lado discriminar entre distintos grupos definidos para la variable dependiente y por otro utilizar los valores de las variables independientes para adscribir cada nuevo individuo a alguno de los grupos o categorías definidos. Los análisis se han realizado empleando el programa estadístico SPSS 11.5.1 para Windows.

Los ejemplos de este tipo de análisis son muy numerosos en la literatura forestal (e.g. HOUSTON *et al.*, 2002; ROMANYÁ & VALLEJO, 2004; SÁNCHEZ RODRÍGUEZ *et al.*, 2002).

Análisis de componentes principales

El objetivo de esta técnica consiste en encontrar las sucesivas combinaciones lineales (incorre-

ladas dos a dos y de módulo 1) de las variables de partida, de modo que expliquen la mayor variabilidad posible, al objeto de que con sólo unas pocas de ellas pueda explicarse casi la misma variabilidad (o, al menos, un porcentaje muy relevante) de la varianza de las variables originales. El procedimiento de análisis consiste en calcular los autovectores y autovalores de la matriz de varianza-covarianza. Los autovectores se eligen de forma que su módulo sea 1 (es decir de forma que la suma del cuadrado de sus componentes sea 1). La primera componente principal es la combinación lineal correspondiente al mayor autovalor y su varianza es precisamente dicho autovalor. La segunda componente principal es la asociada al segundo mayor autovalor y así sucesivamente (PEÑA, 2002). En este trabajo se ha realizado este análisis sobre las variables estandarizadas para dar a todas las variables la misma importancia y que los resultados no dependan de las unidades elegidas para medir las variables.

Análisis factorial discriminante

Esta técnica trata de encontrar las sucesivas combinaciones lineales de las variables de partida que mejor permitan distinguir o discriminar entre distintos grupos presentes (y conocidos) en una población. Más concretamente, se trata de maximizar el poder discriminante de las combinaciones lineales, entendido dicho poder discriminante como el cociente entre la varianzas entre grupos y la varianza dentro de los grupos, para la combinación lineal en cuestión. El procedimiento de análisis se basa en calcular los autovectores y autovalores de la matriz $W^{-1}B$, donde B es la matriz de varianzas covarianzas entre grupos y W es la matriz de varianzas covarianzas dentro de los grupos (PEÑA, 2002). El primer factor discriminante viene dado por la combinación lineal correspondiente al autovector asociado al mayor autovalor. Dicho autovalor representa precisamente el poder discriminante de dicho factor. El segundo factor

Estadísticos	LH	AH	PH	AF	HF	LF	AE	HE
Media	16,91	6,92	1,45	2,99	1,70	2,91	5,60	4,28
Máximo	32,0	10,5	3,0	4	3,1	4	8,0	6
Mínimo	6,6	2,0	0,6	2,0	1,0	2	3,5	3
Desviación estándar	3,49	1,21	0,34	0,33	0,26	0,22	0,84	0,50

Tabla 1. Estadísticos descriptivos de la muestra de datos empleados

discriminante es el que se calcula a partir del auto-vector asociado al segundo autovalor (en orden de magnitud). Se procede de igual forma para los sucesivos factores discriminantes hasta tantos como variables originales existan en el problema pero sin superar nunca el número de grupo menos 1. Al igual que en el PCA se ha realizado el análisis sobre las variables estandarizadas por las mismas razones que se esgrimieron antes.

Análisis discriminante

La técnica del análisis discriminante pretende encontrar una regla de clasificación que permita asignar (lo más fiablemente posible) una futura observación a uno de los grupos preestablecidos en una población, utilizando únicamente la información suministrada por un conjunto de variables auxiliares. El procedimiento consiste en calcular los factores discriminantes y encontrar los centroides de cada grupo, que no son más que los vectores formados por las medias de cada uno de los factores discriminantes para el grupo en cuestión (es decir, el punto medio del grupo, en lo tocante a los factores discriminantes). Hay varias formas de asignar una nueva observación a un grupo. Así, la regla lineal de Fisher consiste en asignar la nueva observación al grupo cuyo centroide es el más cercano a dicha observación (considera la matriz de varianzas covarianza igual en todos los grupos); esta regla, cuando se supone que la distribución del vector de variables originales, condicionada a cada uno de los grupos es una normal multivariante con vector de medias distinto para cada grupo pero idéntica matriz de varianzas covarianza, la regla discriminante de Fisher coincide con la regla de máxima verosimilitud, que es aquella que asigna la nueva observación al grupo que presenta una mayor densidad de probabilidad en los valores observados de las variables para esa nueva observación. Si se permite que las matrices de varianzas covarianzas sean distintas para cada grupo, la regla de máxima verosimilitud desemboca en una regla de tipo cuadrático que, por lo tanto, no coincide con la de Fisher. Otra regla de clasificación alternativa es la de Bayes (o de máxima probabilidad a posteriori), que consiste en ponderar las verosimilitudes mediante las estimaciones de las probabilidades de pertenencia a priori a cada grupo (PEÑA, 2002).

RESULTADOS Y DISCUSIÓN

Del estudio de correlaciones lineales, se observa que hay muchas significativamente distintas de cero. Así, dentro de las variables que miden distintas distancias sobre el mismo elemento (hoja, fruto y erizo) se observa correlación positiva entre ellas. Por ejemplo, cualesquiera dos de las variables relativas a la hoja (AH, LH y PH) tienen correlación significativa. Algunas variables apenas tienen correlación más que con las que miden otra dimensión de ese mismo elemento (como la altura del erizo, HE). De hecho esta variable (HE) sólo presenta correlación estadísticamente significativa con la anchura del erizo (AE) y con la anchura de la hoja (AH). Del estudio de las gráficas de dispersión (para cada par de variables), que permiten inspeccionar visualmente los resultados de las correlaciones, se observa una relación lineal clara (creciente) entre anchura de fruto (AF) y longitud de fruto (LF), mientras que AE y HF no muestran una gráfica de dispersión con aparente dependencia. Como ya se comentó, el PCA se ha realizado sobre la matriz de correlaciones (es decir las variables se han tipificado). En el caso de que las variables originales fueran incorreladas (sin dependencia) el PCA sería superfluo porque las variables originales ya son las componentes principales (esfericidad de la población). El contraste de esfericidad de Bartlett nos informa sobre la esfericidad de las variables, es decir sobre la relación entre las variables analizadas. En este caso se ha obtenido un p-valor del contraste de Bartlett de 0,000, lo que indica que las variables no son incorreladas y, que por tanto las propias variables no pueden ser ellas mismas componentes principales. La tabla 2 muestra todos los autovalores, el porcentaje de varianzas que explica cada componente y sus valores acumulados. Las tres últimas columnas nos dan la misma información para las componentes seleccionadas en el modelo: en este caso tres pues esos son los autovalores mayores que 1.

Se observa, como con tan sólo tres componentes (combinaciones lineales de las 8 variables de partida) se puede explicar el 68 % de la variabilidad de los datos.

En la matriz de componentes de la tabla 3 se puede identificar lo que significa cada compo-

Componente	Autovalores iniciales			Suma de las saturaciones al cuadrado		
	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	2,180	27,248	27,248	2,180	27,248	27,248
2	1,664	28,805	48,054	1,664	20,805	48,054
3	1,604	20,056	68,109	1,604	20,056	68,109
4	0,773	9,659	77,768			
5	0,672	8,398	86,166			
6	0,436	5,451	91,617			
7	0,338	4,225	95,842			
8	0,333	4,158	100,000			

Tabla 2. Varianza total explicada por cada una de las componentes

nente principal, ya que se recogen los coeficientes de correlación entre cada variable original y las tres componentes principales. Se puede observar como la tercera componente está mayormente correlada (positivamente, además) con las dos variables del erizo y mucho menos con las demás, podría interpretarse como una medida “del tamaño del erizo”. Las otras dos componentes principales son de más difícil interpretación. Mientras la segunda crece al aumentar el tamaño de la hoja y del fruto y decrece al disminuir estos (podría medir simplemente “tamaño de todo conjuntamente”), la primera crece al crecer las dimensiones del fruto, pero decrece al aumentar las de la hoja. Por lo anterior, la primera componente principal podría interpretarse como “una medida de la relación entre tamaño de fruto y el tamaño de la hoja”. En la matriz de coeficientes de la tabla anterior se muestran los coeficientes por los que hay que multiplicar cada una de las variables originales para obtener la combinación lineal que represen-

ta cada componente principal. Como en este caso los cálculos se han realizado sobre la matriz de correlaciones (variables tipificadas) los cuadrados de los coeficientes de cada columna no suman 1.

El objetivo del análisis factorial discriminante ha sido investigar si las variedades de castaño pueden explicarse razonadamente bien teniendo presentes las 8 variables sobre la dimensión de la hoja, fruto y erizo. Se busca encontrar una función discriminante que nos permita aseverar con las 8 variables de entrada que tenemos, y con un nivel alto de probabilidad, ante que variedad estamos. El número de funciones discriminantes es el menor de los dos valores siguientes: número de grupos menos uno y el número de variables; en este caso el menor valor es el número de grupos menos uno (2 funciones discriminantes). En la tabla 4 se expresa el poder discriminante de cada función (autovalor) y el porcentaje de varianza (de discriminación que es capaz de explicar).

	Matriz de componentes			Matriz de coeficientes		
	Componente 1	Componente 2	Componente 3	Componente 1	Componente 2	Componente 3
LH	-0,691	0,554	0,0625	-0,317	0,333	0,039
AH	-0,636	0,365	-0,293	-0,292	0,219	-0,183
PH	-0,423	0,671	-0,039	-0,194	0,403	-0,025
AF	0,705	0,491	-0,246	0,323	0,295	-0,153
HF	0,507	0,328	-0,377	0,232	0,197	-0,235
LF	0,516	0,485	-0,276	0,237	0,292	-0,172
AE	0,236	0,353	0,778	0,108	0,212	0,485
HE	0,207	0,256	0,793	0,095	0,154	0,494

Tabla 3. Matriz de componentes y de coeficientes para el cálculo de coeficientes

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	0,841 ^a	60,7	60,7	0,676
2	0,545 ^a	39,3	100,0	0,594

Tabla 4. Poder discriminante de cada función y varianza que es capaz de explicar (^a se han empleado las dos primeras funciones discriminantes canónicas en el análisis)

El contraste de igualdad de grupos de la Lambda de Wilks ha mostrado como tanto los dos primeros factores discriminantes examinados conjuntamente, como el segundo de ellos por separado, presentan p-valores menores de una milésima. Esto implica que existen diferencias significativas entre las variedades, tanto en lo tocante a los dos factores discriminantes simultáneamente como respecto del segundo factor discriminante solamente. Es decir, ambos factores aportan discriminación estadísticamente significativa. La tabla 5 refleja los coeficientes de los dos primeros factores discriminantes (combinaciones lineales de las variables originales) que nos dan el mayor poder discriminante.

Los valores de la tabla 6, al igual que en el caso del PCA, ayudan a interpretar lo que significa la función discriminante. En esta tabla aparecen los coeficientes de correlación entre cada variable en cuestión y cada una de las funciones discriminantes, marcando con un asterisco aquella que es mayor de las dos (por filas) y ordenando las variables para que queden agrupadas con respecto a cuál de los factores parece asociarse. En este caso vemos como la anchura de fruto (AF) es la variable que más claramente se asocia al primer factor (además positivamente). La longitud y altura del fruto (LF y HF) están negativamente correladas con él (al igual que pasaba con la primera componente principal). Por su

parte, el segundo factor discriminante presenta una correlación muy moderada con AE (negativa) y con PH (positiva). La correlación de este segundo factor con el resto de las variables es aún menos importante. Por tanto el segundo factor es más difícil de identificar (AE, PH).

En la tabla 7 se expresa, para cada uno de los grupos, las coordenadas de los centroides, que son las medias de los dos factores discriminantes. Una representación gráfica de los mismos puede verse en el mapa territorial que el programa SPSS presenta también como salida. Se ha realizado una prueba de Box sobre la igualdad de las matrices de covarianzas de las funciones canónicas discriminantes, obteniéndose un p-valor de 0,000 lo cual indica que la estructura de covarianzas para los distintos grupos es diferente, es decir, no todos los grupos tienen la misma matriz de varianzas-covarianzas.

Para asignar una futura observación a uno de los dos grupos mediante la regla discriminante de Fisher, simplemente calcularíamos las tres combinaciones lineales y observaríamos cual es la de mayor valor. En la tabla 8 se presentan los coeficientes de las funciones discriminantes de Fisher que permiten clasificar un nuevo dato en una de las tres variedades. Sin embargo esta regla lineal no es la más adecuada por la existencia de diferencias entre las matrices de varianzas-covarianzas.

Función	LH	AH	PH	AF	HF	LF	AE	HE
1	-0,337	-0,210	0,122	0,819	0,056	-0,036	0,279	0,176
2	-1,153	0,571	0,966	-0,230	0,273	0,169	-0,355	0,080

Tabla 5. Coeficientes estandarizados de las funciones discriminantes canónicas

Función	LH	AH	PH	AF	HF	LF	AE	HE
1	0,812*	-0,389*	0,356*	0,350*	-0,350*	0,238*	0,336	-0,051
2	0,050	0,246	0,164	0,229	-0,296	-0,187	-0,372*	0,334*

Tabla 6. Matriz con la correlaciones intra-grupos combinadas entre las variables discriminantes y las funciones discriminantes canónicas tipificadas

Función	Variedad 1	Variedad 2	Variedad 2
1	-1,098	1,076	0,301
2	-0,312	-0,716	1,051

Tabla 7. Funciones en los centroides de grupo. Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos

También se puede obtener un gráfico territorial (Figura 1 izquierda), en este mapa se representan las tres regiones en las que se clasificará una futura observación a cada uno de los tres grupos (de las tres variedades en este caso). En este caso, los límites de las regiones no son líneas rectas ya que no es la región discriminante de Fisher (considera la misma matriz de varianzas covarianzas intragrupos para los tres grupos) la que se está usando, pues en el análisis (menú de entrada de datos en el SPSS) se ha permitido que utilice matrices de varianzas-covarianzas distintas, lo cual parece, en la práctica, más razonable. En este caso la regla de clasificación es de tipo cuadrático. En la figura 1 (derecha) se observa como la clasificación de las tres variedades es visualmente muy aceptable, pues los datos de las distintas variedades se entremezclan relativamente poco.

Los resultados de la clasificación (Tabla 9) muestra como la probabilidad de la clasificación incorrecta son del orden de un 24,5 %. El error mayor se comete al clasificar un 17,5 % de casos de variedad 2 como variedad 1 y justo otro 17,5 % de casos como variedades 3. Las variedades

3, y sobre todo la 1, se clasifican correctamente en una mayor proporción (un 77 % y un 82,5 % respectivamente).

CONCLUSIONES

A la vista de los resultados se observa que existen diferencias significativas entre las variables medidas en las tres variedades. Usando las 8 variables originales hemos llegado a resumirlas en 2 mediante un PCA (primera y segunda componente principal), explicando entre las dos un porcentaje razonable de varianza (un 68 %). Usando estas dos componentes se podrían hacer otros tipos de análisis como de regresión, teniendo la ventaja de que se ha reducido notablemente el número de variables. Mediante el FDA se ha visto que es posible discriminar de una manera satisfactoria las variedades en base a las dos funciones discriminantes canónicas. Es posible asignar una futura observación a uno de los grupos mediante la regla discriminante de Fisher, simplemente calculando las tres combinaciones lineales

Variedad	Constante	LH	AH	PH	AF	HF	LF	AE	HE
1	-179,256	0,578	4,694	1,305	12,755	20,358	53,167	4,936	15,864
2	-200,663	0,496	4,097	0,940	19,662	20,402	52,463	5,899	116,582
3	-195,256	0,0480	5,116	5,706	15,825	22,217	54,031	4,815	16,592

Tabla 8. Coeficientes de las funciones discriminantes lineales de Fisher

		VARIEDAD	Grupo de pertenencia pronosticado			Total
			1	2	3	
Original	Recuento	1	198	15	27	240
		2	33	123	33	189
		3	24	22	154	200
	%	1	82,5	6,3	11,3	100,0
		2	17,5	65,1	17,5	100,0
		3	12,0	11,0	77,0	100,0

Tabla 9. Resultado de la clasificación del análisis discriminante. Clasificados correctamente el 75,5% de los casos agrupados originales

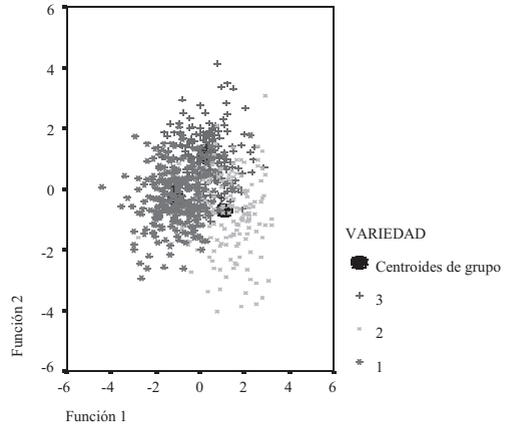
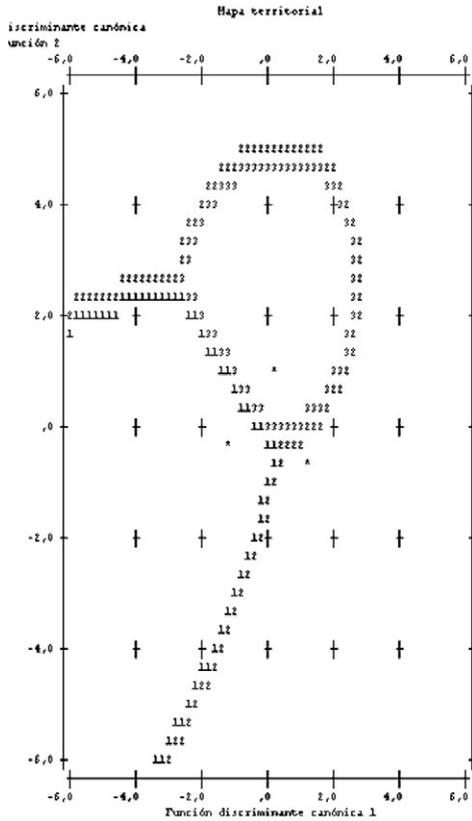


Figura 1. Funciones discriminantes canónicas para cada una de las tres variedades (derecha) y mapa territorial de las variedades (izquierda)

(dadas en la tabla de funciones discriminantes lineales de Fisher) y observando cual es la mayor.

BIBLIOGRAFÍA

ÁLVAREZ ÁLVAREZ, P.; BARRIO ANTA, M.; CASTEDO DORADO, F.; DÍAZ VARELA, R.A.; FERNÁNDEZ LORENZO, J.L.; MANSILLA VÁZQUEZ, P.; PÉREZ OTERO, R.; PINTOS VARELA, C.; RIESCO MUÑOZ, G.; RODRÍGUEZ SOALLEIRO, R.J. Y SALINERO CORRAL, M.C.; 2000. *Manual de selvicultura del castaño en Galicia*. Escuela Politécnica Superior de Lugo. Oviedo. Disponible: <http://www.agrobyte.com>.
 HOUSTON, T.J.; DURRANT, D.W.; BENHAM, S.E.; 2002. Sampling in a variable environmental:

selection of representative positions of throughfall collectors for volume and chemistry under three tree species in the U.K. *For. Ecol. Manage* 158: 1-8.
 PEÑA SÁNCHEZ DE RIVERA, D.; 2002. *Análisis de datos multivariantes*. McGraw-Hill. New York.
 ROMANYÀ, J. & VALLEJO, V.R.; 2004. Productivity of Pinus radiata plantations in Spain in response to climate and soil. *For. Ecol. Manage.* 195: 177-189.
 SÁNCHEZ RODRÍGUEZ, R.J.; RODRÍGUEZ SOALLEIRO, R.J.; ESPAÑOL, E.; LÓPEZ, C.A. & MERINO, A.; 2002. Influence of edaphic factors tree nutritive status on the productivity of Pinus radiata D. Don plantations in northwestern Spain. *For. Ecol. Manage.* 171: 181-189.