Revista FABICIB • año 2015 • volumen 19 • PÁGS. 113 a 120

Divulgación

Análisis factorial múltiple para la caracterización de variedades de trigo pan en diferentes ambientes

RECIBIDO: 28/05/2015
REVISION: 27/07/2015
ACEPTADO: 29/09/2015

Vitelleschi, M.S. • Chavasa, V.

Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Facultad de Ciencias Económicas y Estadística. Universidad Nacional de Rosario. Bvrd. Oroño 1261. (2000) Rosario, Santa Fe, Argentina. Teléfono (0341)-4802794 interno 151.
E-mail: mvitelle@fcecon.unr.edu.ar

RESUMEN: En muchas situaciones experimentales las observaciones de varias variables sobre un conjunto de individuos u objetos se realizan en distintas condiciones experimentales, temporales o ambientales, dando origen a datos de tres modos o vías: individuos, variables y condiciones. Los métodos multivariados que permiten analizar tablas de tres o más modos recogen la verdadera estructura presente en los datos y así, generan conclusiones más completas que las obtenidas al realizar, en forma aislada, los análisis multivariados tradicionales a tablas de dos modos (individuos y variables). El Análisis Factorial Múltiple (AFM) permite abordar esta problemática. En este trabajo se aplica dicha técnica a un conjunto de datos proporcionados por la Estación Experimental Agropecuaria del INTA de Marcos Juárez; que provienen de ensayos comparativos de variedades de trigo pan de ciclo largo, realizados en Corral de Bustos y Cavanagh, campaña 2011/2012. Se consideraron 21 variedades de trigo pan y se evaluaron 8 variables cuantitativas referidas a la calidad y al rendimiento. Lo que constituyó una tabla múltiple de tres modos: individuos, variables y ambientes. El AFM permitió observar qué variedades estaban más afectadas por el ambiente y posibilitó estudiar qué variables resultaron más sensibles a los cambios ambientales.

PALABRAS CLAVES: Datos de tres modos, Análisis Factorial Múltiple, Caracterización de trigo pan.

SUMMARY: *Multiple factor analysis for the characterization of varieties of bread wheat in different environments.*
This paper aims at characterizing 21 varieties of bread wheat conserved at the

germplasm bank of the INTA Marcos Juárez Experimental Station. To this purpose, we analyze together 8 quantitative variables in two different environmental situations (Corral de Bustos and Cavanagh). The experimental design generated three-way or three-mode data, repeated observations of a set of attributes for a set of individuals in different conditions. The information was displayed in a three-dimensional array. The structure of the data was explored using Multiple Factorial Analysis. In conclusion, this method provided useful analytic and graphic tools to study and characterize varieties of bread wheat, specially when the characterization was based on the study of agronomic variables that were affected by environmental conditions.

KEYWORDS: Three-way data; Multiple factorial analysis; Characterization of bread wheat.

Introducción

Las técnicas estadísticas multivariadas posibilitan el estudio simultáneo de un grupo de variables intercorrelacionadas medidas sobre un conjunto de individuos u objetos, permitiendo obtener representaciones simplificadas de bases de datos voluminosas. Dichas técnicas son utilizadas como herramientas para sintetizar la información (1).

Los datos multivariados son arreglados en una tabla o matriz en la que cada fila corresponde a una unidad de observación y cada columna a una variable en estudio; es decir son "datos de dos modos o vías". Denominándose "modo o vía" al conjunto de índice de la tabla; siendo un modo el conjunto de variables y otro el de las observaciones.

En muchas investigaciones las observaciones de un conjunto de variables sobre un grupo de individuos u objetos pueden presentar diferentes estructuras de comportamiento, asociadas principalmente a variables de caracterización como distintas condiciones experimentales, momentos en el tiempo o puntos geográficos, entre otras. Estas diferentes estructuras pueden quedar ocultas en los análisis de la información en su conjunto, si son analizadas como datos de dos modos. Por tal motivo, esta información puede ser estudiada desde la óptica de tablas múltiples; es decir, teniendo en cuenta la existencia de diversos grupos, lo que requiere realizar, por un lado, análisis parciales de cada uno de ellos y, por otro, un análisis global en el que la influencia individual de cada uno de los grupos esté equilibrada (2). El AFM es uno de los métodos utilizados para analizar tablas múltiples (individuos, variables y condiciones).

En este trabajo los datos utilizados fueron proporcionados por la Estación Experimental Agropecuaria del INTA de Marcos Juárez, sobre diferentes variedades de trigo pan de ciclo largo.

Métodos

• *Análisis Factorial Múltiple*

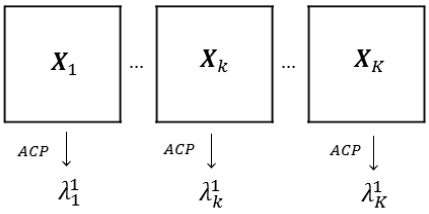
El AFM (3, 4 y 5), desarrollado por Escofier y Pagès, en el seno de la Escuela Francesa de Análisis de Datos, es un método factorial adaptado al tratamiento de tablas de datos (6) en las que un mismo conjunto

de individuos se describe a través de varios grupos de variables. Los mismos pueden ser diferentes, tanto en el número de variables que los componen como en la naturaleza (cuantitativa o cualitativa) o un mismo conjunto de variables medidas en distintos periodos de tiempo o ambientes. Para la aplicación del AFM se requiere que las variables que integran un grupo (o tabla) sean de la misma naturaleza.

La metodología del AFM, cuando todas las variables analizadas son cuantitativas, como en este trabajo, se basa en el Análisis de Componentes Principales (ACP) y se compone de dos etapas:

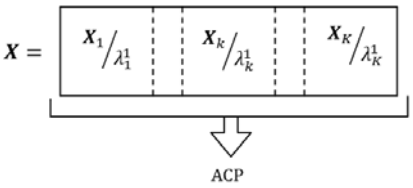
Etapa 1. Análisis parcial. Efectúa un ACP normado de cada tabla de datos (k=1, …, K) y retiene el primer valor propio de cada una de ellas (Figura 1).

Figura 1. Primera etapa del AFM



Etapa 2. Análisis global. Realiza un ACP de la tabla global que resulta de yuxtaponer todas las tablas, a las que previamente a cada una se las ponderó por el inverso del primer valor propio obtenido en la primera etapa (Figura 2). Esta ponderación permite mantener la estructura de cada tabla, ya que todas las variables han recibido la misma ponderación, pero consigue equilibrar la influencia de los grupos, ya que la inercia máxima de cada una de las nubes de individuos definida por los distintos grupos, vale 1 en cualquier dirección.

Figura 2. Segunda etapa del AFM



El objetivo principal de este método es analizar la estructura común de las distintas tablas de datos, poniendo de manifiesto cuáles son los elementos heterogéneos, es decir diferentes al resto.

El AFM proporciona, además de los resultados clásicos del ACP, medidas globales de relación entre los grupos, basados en los coeficientes RV y Lg, que permiten cuantificar la semejanza global existente entre grupos de indicadores parciales (7).

El coeficiente RV (7) puede ser utilizado como medida de similitud entre dos configuraciones; se define como el producto escalar entre pares de matrices (el producto de Hilbert-Schmidt); este producto escalar induce una norma y, por lo tanto, una distancia. Si la correlación vectorial entre dos matrices es igual a la unidad, eso significa que ambas matrices son equivalentes en el sentido de que ambas estructuras son congruentes; cuanto más próximo a uno, más similares las estructuras. Mientras que, si es igual a cero significa que no existe relación entre las variables de los dos grupos considerados. Esta medida es completada con los coeficientes Lg (3) que pueden ordenarse en una matriz de orden KxK y miden la dimensionalidad (número de factores de inercia considerable) de cada grupo. Estos coeficientes toman el valor cero cuando no existe relación entre los grupos y no tienen cota superior; es decir, son más grandes cuanto más multidimensionales sean las

tablas analizadas y presenten mayor cantidad de dimensiones comunes y próximas a las direcciones de inercia más importante de cada tabla.

Materiales

El datos analizados en este trabajo corresponde a un conjunto de 21 variedades de trigo pan de ciclo largo (8 y 9) proporcionados por la Estación Experimental Agropecuaria del INTA de Marcos Juárez. Los ensayos fueron realizados en campo de productores de las localidades de Corral de Bustos y Cavanagh, durante el ciclo agrícola 2011/2012. Se evaluaron las variables: Rendimiento (REND, Kg/ha), Peso hectolítrico (PESOh, Kg/hl), Proteína grano (PROTg, %), Rendimiento harina (RENDh, %), Gluten húmedo (GLUTh, %), Alveograma W (W, 10^{-4} Julios), Alveograma P/L (PL, mm. de agua) y Volumen de Panificación (VOL, cm^3). Cada localidad representa un *ambiente*, por lo tanto los modos de la matriz de datos resultante son: variedades, características y ambientes.

Resultados

Todos los resultados se obtuvieron a través del software R (versión 2.12.0). Se denotó con A1 al conjunto de datos que pertenecen a Corral de Bustos y con A2 a Cavanagh. Análogamente, a cada etiqueta de las variables se le agregó el número 1 o 2, para referenciar al ambiente.

En la primera etapa se realizó un ACP normado sobre cada tabla de datos obteniendo el primer autovalor de cada una de ellas, $\lambda_1^{(1)} = 3{,}8$ y $\lambda_1^{(2)} = 3{,}7$, respectivamente. El primer eje parcial proyecta un porcentaje de inercia del 48% para A1 y 46% para A2.

En la segunda etapa, el primer autovalor ($\lambda_1 = 1{,}8$) recoge el 42% de la inercia total, mientras que el segundo ($\lambda_2 = 0{,}9$) un 20%. El valor del primer autovalor está próximo a la cantidad de condiciones analizadas, lo que indica que el primer eje principal es la dirección global de mayor inercia común a las dos tablas.

La correlación entre los factores parciales de cada una de las tablas y los correspondientes a la tabla global (Tabla 1) muestran que ninguno de los grupos tiene mayor protagonismo en el análisis global.

Tabla 1. Coeficientes de correlación entre los factores parciales y los factores del análisis global.

	Correlación	
Ambientes	**Eje 1**	**Eje 2**
A1	0,942	0,946
A2	0,943	0,947

En este trabajo se considera para el análisis, el plano factorial generado por las dos primeras componentes dado que en el mismo se pueden observar las principales características, semejanzas y diferencias de las variedades de trigo pan.

El coeficiente RV resultó ser igual a 0,6, lo cual sugiere que los dos ambientes presentan una estructura con más similitudes que diferencias.

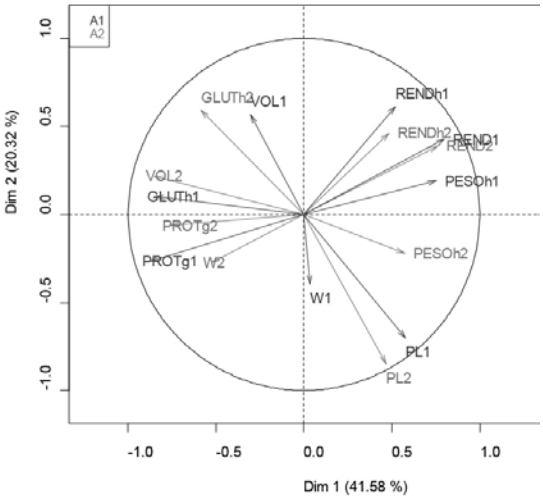
En la matriz de coeficientes Lg (Tabla 2) se puede observar que el coeficiente $Lg(A_1,A_2)$ es igual a 0,843, lo que indica que las variables del ambiente 1 están relacionadas con las variables del ambiente 2.

Tabla 2. Matriz de coeficientes Lg

	A1	A2
A1	1,374	0,843
A2	0,843	1,365

En el análisis de los vectores–variables (Figura 3) se puede apreciar que el primer eje global, marca diferencias entre las variables rendimiento de ambos ambientes y peso hectolítrico del ambiente 1 con volumen de panificación del ambiente 2, gluten húmedo del ambiente 1 y proteína en grano de ambos ambientes. En lo que se refiere al segundo eje global, las variables que más contribuyen son alveograma P/L de ambos ambientes, rendimiento de harina del grupo 1, gluten húmedo del ambiente 2 y volumen de panificación del ambiente 1. Marcando diferencias entre ellas, las dos primeras contribuyen en forma negativa y las tres restantes en forma positiva.

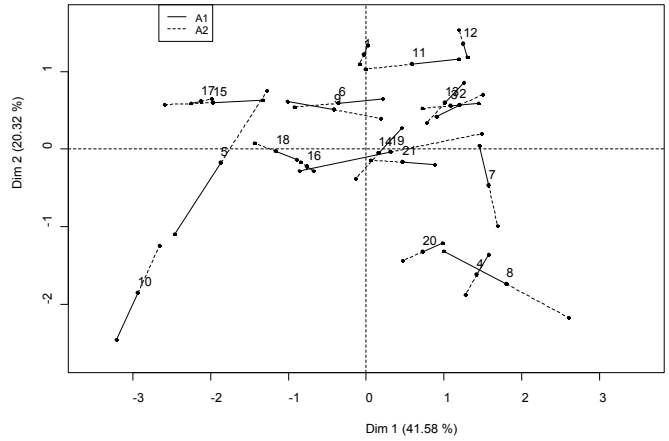
Figura 3. Proyección de las variables sobre los dos primeros ejes factoriales



Además, en la Figura 3 se puede observar que la mayoría de los vectores que representan a la misma variable en ambos ambientes presentan un ángulo pequeño, es decir tienen un comportamiento estable en los dos ambientes. Se producen algunas excepciones, los vectores que representan a las variables W, GLUTh y VOL exhiben un ángulo mayor entre los dos ambientes, sugiriendo que las mismas presentan un comportamiento menos estable. Pudiendo concluir que serían las variables más afectadas por el ambiente.

En relación al plano de los individuos, en la Figura 4 se muestra la trayectoria de cada variedad de trigo pan en los dos ambientes, proyectadas sobre los dos primeros ejes factoriales. Las trayectorias están representadas por tres puntos: los de los extremos corresponden a cada posición relativa que ocupa la variedad en cada uno de los dos ambientes (individuos parciales) y el punto medio es el centro de gravedad (individuos medios).

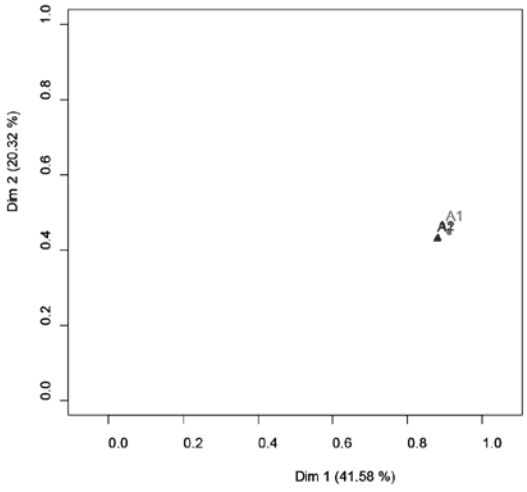
Figura 4. Proyección de los individuos medios y parciales sobre los dos primeros ejes factoriales.



Se puede observar que las variedades 17, 16, 1 y 12 no se ven afectadas por el ambiente. En cambio, las variedades 10, 5, 7, 15, 19 y 8, entre otras, son las que presentan mayor efecto ambiente. En 5, 10 y 7 las mayores diferencias se relacionan al rendimiento de la harina, volumen de panificación, alveograma P/L y gluten húmedo. Mientras que, en las variedades 15, 19 y 8 se diferencian de un ambiente a otro por tener distintos contenidos de proteína en grano, peso hectolítrico del grano y rendimiento.

En la representación de los dos ambientes sobre los dos primeros ejes globales (Figura 5) se pone de manifiesto que los ejes presentados recogen una realidad común a las tablas consideradas, ya que la contribución es la misma para todas y cada una de ellas. Los dos factores globales extraídos para el análisis están igual-

Figura 5. Representación de los ambientes sobre los dos primeros ejes globales



mente afectados por los dos ambientes. Se puede apreciar que en el primer eje global, los valores de las coordenadas para los dos ambientes son próximos a 1, constituyéndose en una dirección de inercia importante para cada ambiente. El segundo eje global está relacionado moderadamente con los ambientes.

Comentarios finales

Las investigaciones en las que se dispone de información de tres vías tienen objetivos más ambiciosos, ya que no se limitan a la búsqueda de relaciones entre variables y tipologías de los individuos, sino que se amplían al análisis comparativo de las realidades presentes en el seno de cada una de las tablas (6).

El tratamiento de tablas múltiples supone un enfoque mucho más completo que el de tablas a doble entrada. En el que cada una tiene identidad propia, esto es, tiene un papel activo en los resultados globales; proporcionando, además, indicadores apropiados para medir las semejanzas y las diferencias entre las estructuras internas de cada uno de los grupos considerados (5).

El AFM se ha convertido en una metodología con una gran versatilidad para el tratamiento de información de tres dimensiones.

Los resultados obtenidos a través del AFM, poseen información mucho más rica en relación a la interpretación del efecto ambiente y sus interacciones, que las que se hubieran obtenido al analizar las tablas de datos a dos modos con las técnicas multivariadas tradicionales. Se logró identificar a las variables que resultaron más sensibles a los cambios ambientales, siendo las mismas: volumen de panificación, gluten húmedo y alveograma W. Además, se consiguió identificar a las variedades que fueron menos afectadas por el ambiente, siendo las mismas 12, 1, 17 y 16 entre otras; como así también a las variedades 5, 10, 8,7 15 y 19, entre otras, que resultaron ser las más afectadas por el ambiente. Las mayores diferencias de las variedades 7, 5 y 10 se relacionan al rendimiento de la harina y volumen de panificación; mientras que en las 15, 8 y 19 se deben al peso hectolítrico del grano y gluten húmedo.

En síntesis, el AFM permitió caracterizar a las variedades de trigo pan, sin que haya un grupo de variables más preponderante que otro. Se logró obtener una representación superpuesta de las variedades de trigo pan vistas a través de cada ambiente, permitiendo observar qué variedades estaban más afectada por el mismo. Además, posibilitó estudiar qué variables resultaron más sensibles a los cambios ambientales; en este caso particular resultaron ser las variables alveograma W, gluten húmedo y volumen de panificación.

Agradecimientos

Las autoras agradecen a la Estadística Beatriz Masiero y a la Ingeniera Leticia Mir por brindar la base de datos utilizada en este trabajo.

Parte de este trabajo fue presentado a la XIX Reunión Científica del Grupo Argentino de Biometría (2014).

Referencias bibliográficas

1. Cuadras, C., 2012. "Nuevos Métodos de Análisis Multivariente". CMC Editions, Barcelona.

2. Kroonenberg, P. M., 2008. "Applied Multiway Data Analysis". John Wiley & Sons, Inc. Hoboken. New Jersey.

3. Escofier, B.; Pagès, J., 1992. "Análisis Factoriales Simples y Múltiples". Ed. Universidad del País Vasco.

4. Escofier, B.; Pagès, J., 1994. Multiple Factor Analysis (AFMULT package). Computational Statistics and Data Analysis **18:** 121-140.

5. Pagès, J., 2004. Multiple Factor Analysis: Main Features and Application to Sensory Data. Rev. Colombiana de Estadística **27:** 1-26.

6. Fernández Aguirre, K; Landaluce Calvo, M.; Modroño Herrán, J. 2013. Nuevo procedimiento metodológico para el análisis exploratorio de una tabla estructurada en diversos conjuntos de individuos. Estadística Española, 55, **182:** 305-322.

7. Abdi, H., 2007. RV Coefficient and Congruence Coefficient. Encycl. of Measuremente and Statistics. Thousand Oaks (CA): Sage. 849-853.

8. Abbate, P.; Gutheim, O.; Milisich, H.; Cuniberti, M., 2010. "Fundamentos para la clasificación del trigo argentino por calidad: efectos del cultivar, la localidad, el año y sus interacciones". Agriscientia **17:** 1-9.

9. Cuniberti, M.; Mir, L.; Masiero, B.; Fraschina, J., 2012. Influencia varietal en parámetros de calidad y rendimiento en trigo. Interacción GxA. Informe de Actualización Técnica **23:** 43-54.