Análisis exploratorio de base de datos de AACREA aplicando técnica estadística multivariada Clúster, utilizando programa R Studio

Jesica Ciminari
Cátedra de Estadística, Facultad de Ciencias Agrarias, Universidad Nacional de Rosario
Zavalla, Argentina
jesicaciminari@hotmail.com

Abstract

In Argentina there are important dairy basins, for this reason it is important to carry out studies in order to analyze the economic and productive growth of the country's drums. However, except for the national agricultural census carried out by the INDEC, and studies carried out by the INTA, it is difficult to have information and databases obtained properly from the drums and over the years. At a private level, one of the databases that exists in the country is that of the Argentine Association of Regional Agricultural Experimentation Consortiums (AACREA), supported by the advisors and members of AACREA. The main objective of this association is to ensure the proper functioning of the groups, so that they are economically profitable and sustainable over time. This work aims to perform an exploratory analysis of the AACREA database by applying the Cluster multivariate technique through the R Studio statistical program, to group and, thus, reduce the number of variables. In addition, we want to compare the results obtained from applying the Cluster technique with the results obtained by applying, years ago, the Main Components technique in the same database.

Keywords: Exploratory Analysis, Cluster Analysis, Main Components, Drums.

Resumen

En Argentina se encuentran importantes cuencas lecheras, por esta razón es importante llevar adelante estudios a fin de poder analizar el crecimiento económico y productivo que presentan los tambos del país. Sin embargo, salvo por el censo nacional agropecuario que realiza el INDEC, y estudios realizados por el INTA, es difícil contar con información y bases de datos obtenidas propiamente de los tambos y a través de los años. A nivel privado, una de las bases de datos que existe en el país es la de la Asociación Argentina de Consorcios Regionales de Experimentación Agrícola (AACREA), sostenida por los asesores e integrantes de AACREA. El objetivo principal de esta asociación es asegurar el buen funcionamiento de los grupos, para que éstos sean económicamente rentables y sustentables en el tiempo. Este trabajo apunta a realizar un análisis exploratorio de la base de datos de AACREA aplicando la técnica multivariada Clúster a través del programa estadístico R Studio, para agrupar y, de esta manera, lograr reducir el número de variables. Además se quiere comparar los resultados obtenidos de aplicar la técnica Clúster con los resultados obtenidos al haber aplicado, años atrás, la técnica de Componentes Principales en la misma base de datos.

Palabras claves: Análisis Exploratorio, Análisis de Clúster, Componentes Principales, Tambos.

1. INTRODUCCION

En Argentina se encuentran importantes cuencas lecheras. La historia de la producción láctea argentina ha mostrado un crecimiento permanente, más allá de los pronunciados vaivenes derivados de las condiciones económicas internas que hicieron acelerar y/o retrasar, por momentos, este avance.

La producción de leche en la República Argentina participa de dos condiciones difíciles de encontrar en otras regiones del mundo, especialmente las dos juntas. La posibilidad, salvo circunstancias climáticas extraordinarias, de realizar pastoreo directo durante todo el año y la abundancia de concentrados energéticos proteicos a valores por debajo de los precios internacionales.

En forma similar a otras lecherías del mundo la evolución de la producción se llevó a cabo junto con, y a pesar de, una reducción muy importante de la cantidad de explotaciones lecheras. Se estima que en las últimas décadas el total de tambos se redujo en un cincuenta por ciento sin afectar la existencia de ganado y aumentando a su vez la producción de leche.

Paralelamente fue necesario aumentar la eficiencia de producción. En este sentido, el registro y análisis de datos en los tambos se ha convertido en una herramienta indispensable para mejorar la gestión y contribuir al principal objetivo de los tambos, el cual apunta al crecimiento económico y a mejorar la producción de leche año a año.

Sin embargo, salvo por el relevamiento estadístico que realiza el INDEC mediante el censo nacional agropecuario, y estudios realizados por el INTA, es difícil contar con información y bases de datos obtenidas propiamente de los tambos y a través de los años, a fin de poder analizar profundamente su crecimiento económico y productivo.

A nivel privado, una de las bases de datos que existe en el país es la de la Asociación Argentina de Consorcios Regionales de Experimentación Agrícola (AACREA), sostenida por los asesores e integrantes de AACREA.

AACREA es una organización civil sin fines de lucro que nuclea a empresas agropecuarias con el propósito de mejorar los resultados de sus organizaciones a través del intercambio de ideas y experiencias. Los miembros trabajan en conjunto para mejorar el proceso de trabajo de la empresa y responden a las necesidades técnicas, económicas y humanas. AACREA está compuesta por alrededor de 40 grupos CREA (Consorcios Regionales de Experimentación Agrícola) pertenecientes a las distintas regiones del país. (www.crea.org.ar. 2019)

El objetivo principal de la Asociación es asegurar el buen funcionamiento de los grupos, para que éstos sean económicamente rentables y sustentables en el tiempo. Promueve la prueba y la adopción de tecnología para luego transferirla al medio, contribuyendo de esta manera con el sector y el país.

Asimismo, se encarga de atender las demandas de los grupos y ayuda a trabajar eficazmente; desarrolla y lleva adelante proyectos de capacitación, experimentación y transferencia buscando anticiparse a las necesidades futuras. De esta misma manera, propicia el desarrollo comunitario.

Además, promueve el intercambio y el trabajo en conjunto con expertos y organismos de investigación nacionales y extranjeros.

2. OBJETIVOS

2.1 Objetivo General

Analizar de manera exploratoria la base de datos de AACREA aplicando la técnica multivariada Clúster a través del programa estadístico R Studio, para agrupar y, de esta manera, lograr reducir el número de variables.

2.2 Objetivos Específicos

Comparar resultados al aplicar la técnica de Clúster con resultados obtenidos al haber aplicado, años atrás, la técnica de Componentes Principales en la misma base de datos.

3. MATERIALES

Se cuenta con una base de datos que incluye datos correspondientes a relevamientos realizados en tres años consecutivos: 2006-2007, 2007-2008 y 2008-2009. Cada año abarca el periodo comprendido entre el primero de julio de un año y el treinta de junio del año siguiente. Cada observación corresponde a los datos de una empresa perteneciente a los 40 grupos CREA. La base de datos cuenta con un total de 446 observaciones. Las variables incluidas dentro de la base de datos son 17. Dichas variables se presentan en la Tabla 1, observándose el nombre de la variable, su abreviatura, clasificación de la variable y el área de interés al cual pertenece.

Tabla 1: Variables incluidas en la base de datos.

VARIABLE	**ABREVIATURA**	**CLASIFICACION DE LA VARIABLE**	**AREA DE INTERES**
Año	Año	Cualitativa	Identificación de las empresas tamberas
Región	Región	Cualitativa	Identificación de las empresas tamberas
Número de tambos	Nro tambos	Cuantitativa Discreta	Identificación de las empresas tamberas
Litros de ejercicio	Lejer	Cuantitativa Continua	Producción y calidad de la leche
Litros por vacas en ordeñe por día	L/VO/día	Cuantitativa Continua	Producción y calidad de la leche
Kilogramos de grasa butirosa	kgGB	Cuantitativa Continua	Producción y calidad de la leche
Porcentaje de grasa butirosa	%GB	Cuantitativa Continua	Producción y calidad de la leche

Porcentaje de proteínas	%P	Cuantitativa Continua	Producción y calidad de la leche
Vacas en ordeñe	VO	Cuantitativa Discreta	Rodeo
Vacas secas	VS	Cuantitativa Discreta	Rodeo
Vacas totales por hectárea	VT/haVO+VS	Cuantitativa Continua	Rodeo
Porcentaje de vacas de rechazo	%VR	Cuantitativa Continua	Rodeo
Porcentaje de vacas muertas	%VM	Cuantitativa Continua	Rodeo
Kilogramos peso vivo equivalente vaca adulta	kgPVEVA	Cuantitativa Continua	Rodeo
Hectáreas de vacas en ordeñe más vacas secas	haVO+VS	Cuantitativa Continua	Superficie
Equivalente grano forrajes conservados	EGFC	Cuantitativa Continua	Suplementación
Equivalente grano concentrados	EGC	Cuantitativa Continua	Suplementación

4. METODOLOGIA

Para poder responder al objetivo general se utilizó la técnica estadística multivariada Clúster. El análisis Clúster tiene como objetivo agrupar las variables disponibles en la base de datos en clústeres, de manera que dentro de cada uno queden las variables más similares, logrando de esta manera reducir el número de variables a trabajar. Para responder al objetivo específico se utilizó la técnica estadística de Componentes Principales. Dentro del objetivo específico se busca comparar los resultados obtenidos al aplicar la técnica de Clúster, con resultados obtenidos años atrás, al haber aplicado la técnica multivariada de Componentes Principales en la misma base de datos. A continuación se desarrolla cada una de estas técnicas.

4.1 Análisis Clúster

El análisis de clúster es una técnica cuya idea básica es agrupar un conjunto de observaciones (o variables) en un número dado de clústeres o grupos no conocidos de antemano pero sugeridos por la propia esencia de los datos, de manera que observaciones (o variables) que puedan ser considerados similares sean asignados a un mismo clúster, mientras que las diferentes (disimilares) se localicen en clústeres distintos. De esta manera, se logra la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos. Este agrupamiento se basa en la idea de distancia o similitud entre las observaciones. La obtención de dichos clústeres depende del criterio o distancia considerados. La variedad de formas de medir diferencias o distancias entre casos proporciona diversas posibilidades de análisis. El empleo de ellas, y el de las que continuamente siguen apareciendo, así como de los algoritmos de clasificación, o diferentes reglas matemáticas para asignar los individuos a distintos

grupos, depende del fenómeno estudiado y del conocimiento previo de posible agrupamiento que de él se tenga.

Existen dos grandes tipos de análisis de clúster: Jerárquicos y No Jerárquicos.

4.1.1 *Método Análisis Clúster Jerárquico*

En la práctica, no se pueden examinar todas las posibilidades de agrupar los elementos, incluso con los ordenadores más rápidos. Una solución se encuentra en los llamados métodos jerárquicos. Se tienen dos posibles formas de actuar:

- ✓ *Métodos jerárquicos aglomerativos*: se comienza con los objetos de modo individual; de este modo, se tienen tantos clústeres iniciales como objetos (o variables). Luego se van agrupando de manera que los primeros en hacerlo son los más similares y al final, todos los subgrupos se unen en un único clúster. Las medidas de similitud entre observaciones son calculadas por distancias entre ellas. Las medidas más utilizadas son:

  - Euclídea:

$$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

  - Manhattan:

$$d_{man}(x,y) = \sum_{i=1}^{n}\left|(x_i - y_i)\right|$$

  - Correlación de Pearson:

$$d_{cor}(x,y) = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  - Correlación de Spearman:

$$d_{spear}(x,y) = 1 - \frac{\sum_{i=1}^{n}(x'_i - \bar{x'})(y'_i - \bar{y'})}{\sqrt{\sum_{i=1}^{n}(x'_i - \bar{x'})^2 \sum_{i=1}^{n}(y'_i - \bar{y'})^2}}$$

donde:

$$x'_i = rank(x_i) \text{ and } y'_i = rank(y).$$

  - Correlación de Kendall:

$$d_{kend}(x,y) = 1 - \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

donde:

n_c: total number of concordant pairs
n_d: total number of discordant pairs
n: size of x and y

- ✓ *Métodos jerárquicos divididos:* se actúa de manera contraria. Se parte de un grupo único con todas las observaciones (o variables) y se van dividiendo según lo lejanos que estén.

En cualquier caso, de ambos métodos se deriva un dendograma, que es un gráfico que ilustra cómo se van haciendo las subdivisiones o los agrupamientos, etapa a etapa.

Consideremos aquí los métodos aglomerativos con diferentes métodos de unión (linkage methods). Los más importantes son:

- Encadenamiento Simple: Mínima disimilitud entre-clúster. Calcula todos los pares de disimilitudes entre las observaciones en el clúster A y las observaciones en el clúster B, y registra la menor de esas disimilitudes.
- Encadenamiento Completo: Máxima disimilitud entre-clúster. Calcula todas las disimilitudes entre pares de observaciones en un clúster A y las observaciones en el clúster B, y se queda con la máxima de esas disimilitudes.
- Promedio Simple: Disimilitud Media entre-clúster. Calcula la disimilitud entre pares de observaciones en el clúster A y las observaciones en el clúster B, y registra el promedio de esas disimilitudes.
- Distancia Centroide (o prototipo): Disimilitud entre el centroide del clúster A (el vector de medias de dimensión p) y el centroide del clúster B, y registra el promedio de esas disimilitudes.
- Método de Mínima Variancia de Ward: Minimiza la variancia total intra-clústeres. En cada paso el par de clústeres con mínima distancia entre clústeres se unen.

El valor de la medida de distancia está relacionado con la escala en la cual se midieron las variables. Por esto es recomendable estandarizar las variables antes de calcular las distancias, sobre todo si están medidas en escalas diferentes.

La estandarización permite que las variables sean comparables. Generalmente se estandariza de manera tal que tengan desviación 1 y media 0.

Definidas las distancias anteriores, se puede considerar el algoritmo básico, dados N objetos o individuos:

1. Empezar con N clústeres (el número inicial de elementos) y una matriz N × N simétrica de distancias o similitudes. $D = [d_{ik}]_{ik}$.

2. Dentro de la matriz de distancias, buscar aquella entre los clústeres U y V que sea la menor entre todas, d_{uv}.

3. Juntar los clústeres U y V en uno solo. Actualizar la matriz de distancias:

(i) Borrando las filas y columnas de los clústeres U y V.

(ii) Formando la fila y columna de las distancias del nuevo clúster (UV) al resto de clústeres.

4. Repetir los pasos (2) y (3) un total de (N − 1) veces.

Al final, todos los objetos están en un único clúster cuando termina el algoritmo. Además, se guarda la identificación de los clústeres que se van uniendo en cada etapa, así como las distancias a las que se unen. Finalmente se construye un dendograma.

Estos métodos se pueden usar para clasificar no sólo observaciones, sino también variables usando como medida de similitud algún coeficiente de correlación.

4.1.2. *Método Análisis Clúster No Jerárquico*

Se usan para agrupar objetos, pero no variables, en un conjunto de k clústeres ya predeterminado. No se tiene que especificar una matriz de distancias ni se tienen que almacenar las iteraciones. Todo esto permite trabajar con un número de datos mayor que en el caso de los métodos jerárquicos. Se parte de un conjunto inicial de clústeres elegidos al azar, que son los representantes de todos ellos; luego se van cambiando de modo iterativo. Se usa habitualmente el método de las k-medias.

Método de las k-medias: Es un método que permite asignar a cada observación el clúster que se encuentra más próximo en términos del centroide (media). En general, la distancia empleada es la euclídea.

Pasos:

1. Se toman al azar k clústeres iniciales.

2. Para el conjunto de observaciones, se vuelve a calcular las distancias a los centroides de los clústeres y se reasignan a los que estén más próximos. Se vuelven a recalcular los centroides de los k clústeres después de las reasignaciones de los elementos.

3. Se repiten los dos pasos anteriores hasta que no se produzca ninguna reasignación, es decir, hasta que los elementos se estabilicen en algún grupo. Usualmente, se especifican k centroides iniciales y se procede al paso (2) y, en la práctica, se observan la mayor parte de reasignaciones en las primeras iteraciones.

4.2 ANALISIS DE COMPONENTES PRINCIPALES

El Análisis de Componentes Principales (ACP) es una técnica estadística de síntesis de la información, o reducción de la dimensión (número de variables). Es decir, ante una base de datos con muchas variables, el objetivo es reducirlas a un menor número perdiendo la menor cantidad de información posible. Las nuevas componentes principales o factores serán una combinación lineal de las variables originales, y además serán independientes entre sí.

Un aspecto clave en ACP es la interpretación de los factores, ya que ésta no viene dada *a priori*, sino que será deducida tras observar la relación de los factores con las variables iniciales (habrá que

estudiar tanto el signo como la magnitud de las correlaciones). Esto no siempre es fácil, y será de vital importancia el conocimiento que el experto tenga sobre la materia de investigación.

En el ACP existe la opción de usar la matriz de correlaciones o bien, la matriz de covarianzas. La primera opción se puede utilizar cuando las variables están expresadas en distintas escalas de medidas. Contrariamente, la segunda opción se aplica cuando todas las variables se encuentran expresadas en la misma unidad de medida.

Sea **X** un vector de *p* variables, tal que E(**X**)=0 y ∑=E(**XX**'). Se define al componente lineal o eje de proyección como: **Z** = **A**'**X**.

Los elementos del vector **A**' = ($a_{11}, a_{12}, ..., a_{np}$) son los cosenos directores. Utilizando multiplicadores de Lagrange para maximizar la variabilidad de **Z**, sujeta a la condición que $\|\mathbf{A}\|$=1, se obtiene:

Var (Z_i) = λ = **A**'∑**A**

La componente principal i-ésima se define como: $Z_i = a_{i1}x_1 + a_{i2}x_2 + ... + a_{ip}x_p$ con variancia λ_i. Donde, $a_{i1}, a_{i2}, ..., a_{ip}$ son dados por el i-ésimo autovector y λ_i por el autovalor de la matriz de covariancias, de modo que $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$.

Las componentes principales se ordenan en función del porcentaje de la variancia explicada, la cual estará indicada por los autovalores $\lambda_1, \lambda_2, \ldots, \lambda_p$. La elección del número de componentes principales se realiza de forma que la primera componente explique la máxima proporción posible de la variabilidad total (indicada por el autovalor λ_1), la segunda componente explique la máxima variabilidad posible (indicada por el autovalor λ_2) no reflejada en la primera y así sucesivamente. En este sentido, la primera componente será la más importante por ser la que explica el mayor porcentaje de la variabilidad de los datos. Queda a criterio del investigador decidir cuántas componentes se elegirán en el estudio. Generalmente se busca que las componentes principales expliquen al menos el 75%-80% de la variabilidad original y que los autovalores $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ sean mayores a 1.

5. RESULTADOS

5.1 Análisis Descriptivo De Las Variables

A continuación se presenta un análisis descriptivo de las variables mostradas en la Tabla 1.

5.1.1 Variables Cuantitativas

En la Tabla 2 se encuentran las medidas descriptivas correspondientes a las variables cuantitativas. Se presenta la media, mediana, desvío estándar, coeficiente de variación y los cuantiles. En el Apéndice 6.1 se presentan los gráficos con la distribución de cada variable, agrupados por área de interés.

Tabla 2: Medidas descriptivas de las variables cuantitativas continuas.

VARIABLES	MEDIA	MEDIANA	DESVIO ESTANDAR	COEF. VAR.	CUANTIL 25%	CUANTIL 75%
Lejer	3273758,92	2283075,00	3182174,02	0,97	1513259,89	3663117,47
L/VO/día	21,09	20,99	2,81	0,13	19,16	22,95
kgGB	120219,19	83240,39	116715,01	0,97	55156,65	135822,21
VO	418,81	294,90	391,81	0,94	197,73	463,85
VS	87,16	65	72,60	0,83	44,00	104,02
VT/haVO+VS	1,45	1,41	0,35	0,24	1,20	1,68
kgPVEVA	584,78	600,00	38,63	0,07	550,00	620,00
haVO+VS	353,59	253,00	328,43	0,11	180,25	380,50
EGFC	1108,47	671,17	1581,82	0,14	335,58	1207,44
EGC	3138,37	1174,43	5944,29	0,19	522,63	3586,94

Las variables Lejer, kgGB, VO, VS, EGFC y EGC presentan una distribución asimétrica hacia la derecha (ver Apéndice 1.1). Esto se puede corroborar observando que el valor que arroja la media es mayor que el valor de la mediana, en todos los casos. En cambio, la variable L/VO/día presenta una distribución simétrica, la media y la mediana arrojan valores similares, 21,09 y 20,99, respectivamente. La variable VT/haVO+VS también presenta una distribución simétrica. En cuanto a la variable haVO+VS se trata de una distribución moderadamente simétrica. Además se obtuvo un coeficiente de variación de 0,11, indicando mucha homogeneidad en los datos observados.

5.1.2 Variables Cualitativas

En la Tabla 3 se presentan las variables cualitativas estudiadas con sus respectivas frecuencias y porcentajes.

Tabla 3: Variables cualitativas estudiadas.

VARIABLE	CATEGORIAS	FRECUENCIA ABSOLUTA	PORCENTAJE (%)
Año	2006-2007	161	36,1
	2007-2008	138	30,9
	2008-2009	147	33,0
Región	1.Oeste	115	25,8
	2.Mar y Sierras	71	15,9
	3.Oeste Arenoso	21	4,7
	4.Este	39	8,7
	5.Santa Fe Centro	114	25,6

	6.Litoral Sur	10	2,2
	7. Norte de Bs. As.	0	0,0
	8.Centro	42	9,4
	9.Sur de Santa Fe	27	6,1
	10.NOA	7	1,6

A partir de la variable Región, se puede observar que la mayoría de los registros presentes en la base de datos final provienen de las regiones: Oeste, Santa Fe Centro, Mar y Sierras, Centro y Este acumulando el 85,4% de los datos. Vale aclarar que la Región Norte de Buenos Aires arroja una frecuencia cero debido a que no se observaron registros de esa región en la base de datos final, con lo cual se trabajará con un total de nueve regiones.

5.1.3 *Gráficos radiales por niveles de producción*.

En los Gráficos 1 y 2 se pueden observar los promedios estandarizados de cada una de las variables estudiadas por nivel de producción para la variable respuesta kilos de grasa butirosa por hectárea de vaca total (kgGB/haVO+VS) y para litros de ejercicios por hectárea de vaca total (Lejer/haVO+VS) respectivamente. El nivel de producción se determinó en Inferior, Medio o Superior de acuerdo a cada variable respuesta.

Grafico 1: Promedios estandarizados de las variables observadas, por nivel de producción, para kilos de grasa butirosa por hectárea de vaca total.

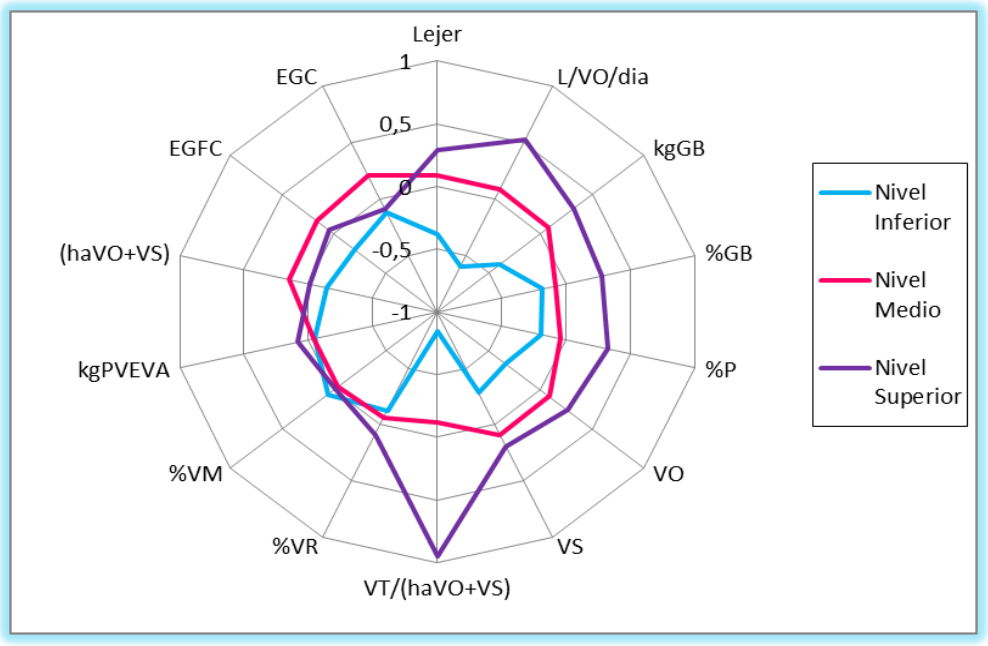
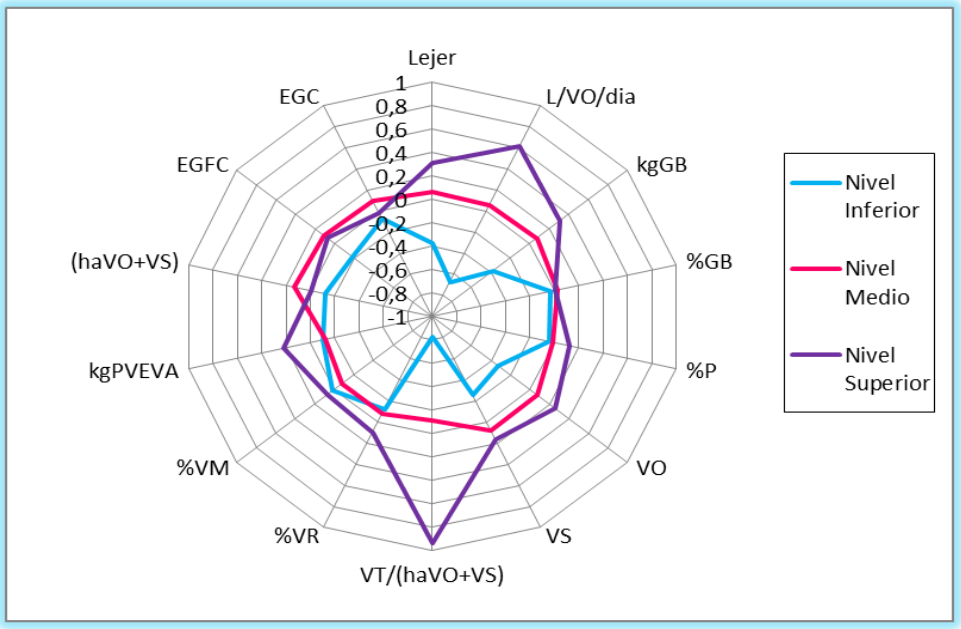


Grafico 2: Promedios estandarizados de las variables observadas, por nivel de producción, para litros de ejercicio por hectárea de vaca total.



A través de los gráficos radiales se puede visualizar que no existen diferencias entre las dos variables respuestas, dado que al observar ambos gráficos se puede ver, que es muy similar la forma que toma cada estrato (nivel inferior, medio y superior). En lo que sí se observan diferencias, es en el peso que toma cada variable en cada nivel de producción. Como es de esperar, los valores de las variables aumentan a medida que se pasa de nivel inferior, a nivel medio y luego a nivel superior, excepto en las variables %VM, haVO+VS, EGFC, EGC, para ambas variables respuestas y en

%GB y kgPVEVA en la variable respuesta Lejer/haVO+VS. Por lo tanto, a la conclusión que podemos arribar es que existen diferencias entre los estratos productivos dentro de cada variable respuesta, pero el comportamiento de cada nivel de producción es similar en cada gráfico. En el Apéndice 6.2 se pueden observar las tablas de los valores de las variables estandarizadas para ambas variables respuestas.

5.2 TECNICA ESTADISTICA MULTIVARIADA CLÚSTER

Para analizar la base de datos con la técnica estadística Clúster se utilizó el programa R Studio.
Se decidió utilizar el Método Jerárquico Aglomerativo. La medida de similitud utilizada fue la del coeficiente de correlación de Pearson y el Linkage utilizado fue el de Encadenamiento Completo.
Se estandarizaron las variables, debido a que estaban medidas en diferentes escalas de medidas.
A continuación se muestra los comandos utilizados en el programa R para obtener los resultados.

```
#Importar la base
library(readxl)
aacrea <- read_excel("C:/estadistica multivariada en r/tema 5/BASE.xls")
View(aacrea)
#Análisis de Clúster
pkgs <- c("factoextra",  "NbClust")
install.packages(pkgs)
library(factoextra)
library(NbClust)
library(Hmisc)
#Estandarizar los datos
aacreaest <- scale(aacrea)
View(aacreaest)
#Método Jerárquico
#Agrupar Variables
agr1 <- varclus(aacreaest, similarity="pearson",
              type="data.matrix",
              method="complete",
              data=NULL, subset=NULL, na.action=na.retain,
              trans="square")
agr1
#Gráfico Dendograma
plot(agr1)
```

La matriz de correlación obtenida a través del programa R fue:

Matriz de correlación de las variables

	Lejer	Lts/VO/D	KgGB	prGB	prP	VO	VS	VT/haVO+VS	prVR	prVM	KgPVEVA	haVO+VS	EGFC	EGC
Lejer	1.00	0.06	0.98	0.00	0.01	0.97	0.96	0.03	0.04	0.01	0.01	0.90	0.62	0.37
Lts/VO/D	0.06	1.00	0.04	0.07	0.02	0.02	0.01	0.01	0.02	0.02	0.09	0.01	0.02	0.01
KgGB	0.98	0.04	1.00	0.02	0.03	0.98	0.98	0.04	0.04	0.01	0.00	0.89	0.63	0.37
prGB	0.00	0.07	0.02	1.00	0.55	0.02	0.02	0.03	0.01	0.01	0.21	0.00	0.00	0.00
prP	0.01	0.02	0.03	0.55	1.00	0.03	0.03	0.02	0.01	0.00	0.12	0.01	0.01	0.00
VO	0.97	0.02	0.98	0.02	0.03	1.00	1.00	0.04	0.04	0.01	0.00	0.90	0.61	0.35
VS	0.96	0.01	0.98	0.02	0.03	1.00	1.00	0.04	0.04	0.01	0.00	0.90	0.60	0.35
VT/haVO+VS	0.03	0.01	0.04	0.03	0.02	0.04	0.04	1.00	0.00	0.01	0.00	0.00	0.00	0.00
prVR	0.04	0.02	0.04	0.01	0.01	0.04	0.04	0.00	1.00	0.02	0.00	0.04	0.04	0.02
prVM	0.01	0.02	0.01	0.01	0.00	0.01	0.01	0.01	0.02	1.00	0.02	0.01	0.01	0.01
KgPVEVA	0.01	0.09	0.00	0.21	0.12	0.00	0.00	0.00	0.00	0.02	1.00	0.00	0.00	0.00
haVO+VS	0.90	0.01	0.89	0.00	0.01	0.90	0.90	0.00	0.04	0.01	0.00	1.00	0.63	0.40
EGFC	0.62	0.02	0.63	0.00	0.01	0.61	0.60	0.00	0.04	0.01	0.00	0.63	1.00	0.79
EGC	0.37	0.01	0.37	0.00	0.00	0.35	0.35	0.00	0.02	0.01	0.00	0.40	0.79	1.00

En la Figura 1 se presenta el dendograma obtenido a partir del programa R Studio:

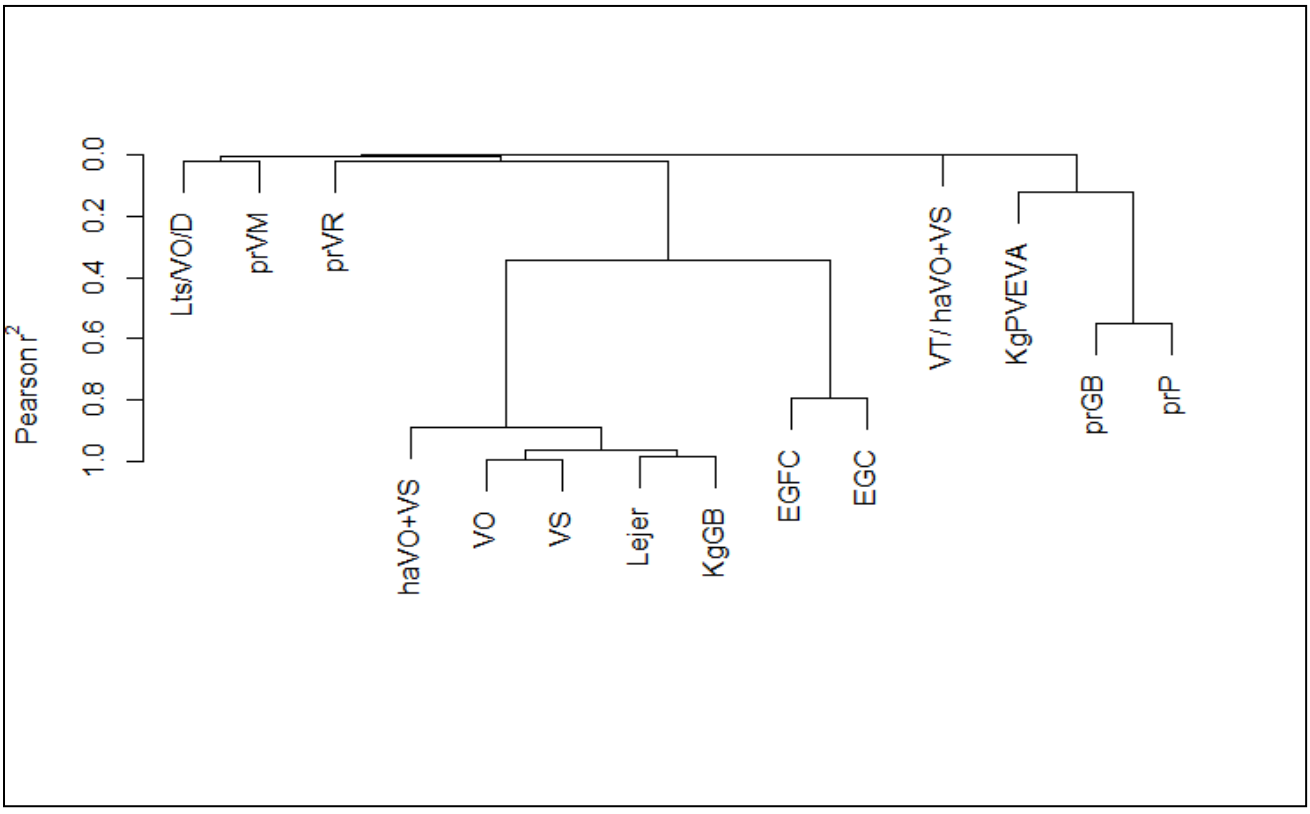


Figura 1: Dendograma obtenido a partir del programa R al aplicar Análisis Clúster

Como se puede observar tanto en la matriz de correlaciones como en el dendograma, las variables con mayor correlación son: Lejer, KgGB, haVO+VS, VO, VS.

Al analizar los resultados y apoyándonos en el dendograma se decide formar 4 clústeres; cada uno contiene las siguientes variables:

Clúster1: %VR, haVO+VS, VO, VS, Lejer, KgGB, EGFC y EGC

Clúster2: KgPVEVA, %GB y %P

Clúster3: Lts/VO/D y %VM

Clúster4: VT/haVO+VS

Se observa que en el primer clúster se agrupan aquellas variables que tienen que ver con la "Producción de leche"; el segundo clúster encierra las variables referidas a la "Calidad de la leche"; en el tercer clúster se observan variables que describen la "Producción en litros diarios" y el último clúster se relaciona con el "Ganado existente".

5.3 ESTIMACION DE LAS COMPONENTES PRINCIPALES

Se aplicó esta técnica multivariada ya que es de interés conocer si las variables medidas con respecto a la producción de leche en las empresas tamberas se encuentran agrupadas según algunas características de interés y de este modo, poder diferenciarlas en grupos bien definidos. A continuación se mostrarán las componentes principales obtenidas al analizar la base de datos a través del programa Statgraphics.

Debido a que las variables se encuentran medidas en distintas escalas y se sabe que las variancias de estas variables no resultan semejantes, es conveniente llevar a cabo dicho análisis con la matriz de correlaciones. En el Apéndice 6.3 se puede observar dicha matriz.

En la Tabla 4 se muestran los autovalores encontrados, el porcentaje de la variabilidad explicada y acumulada en cada autovalor.

Tabla 4: Autovalores, porcentaje de variabilidad explicada y acumulada.

Autovalor	**Valor del Autovalor**	**%Variabilidad Explicada**	**%Variabilidad Acumulada**
1	6,06	43,26	43,26
2	2,27	16,19	59,45
3	1,20	8,55	68,00
4	1,13	8,07	76,07
5	0,87	6,21	82,28
6	0,74	5,28	87,56

Dado que, los primeros cuatro autovalores explican el 76,07% de la variabilidad total de los datos y éstos son mayores que 1, se construyeron cuatro componentes principales. En cada componente, se resaltó en color rojo, aquellas variables que presentaron el mayor coeficiente estimado y a partir de las cuales se les asignó un nombre a cada componente principal.

- Componente Principal 1: **Producción de ejercicio**

$$CP1 = 0{,}3955\ Lejer + 0{,}07013\,(L/VO/día) + 0{,}39825\ kgGB + 0{,}05673\ \%GB + \\ + 0{,}0724\ \%P + 0{,}39609\ VO + 0{,}36604\ VT + 0{,}0594\ VT/(haVO + VS) + \\ + 0{,}10526\,\%\mathrm{VR} + 0{,}04801\,\%VM + 0{,}00017\ kgPVEVA + 0{,}3861\ (haVO + VS) + \\ + 0{,}35611\ EGFC + 0{,}29649\ EGC$$

- Componente Principal 2: **Calidad de la leche**

$$CP2 = -0{,}0599\ Lejer - 0{,}32348\,(L/VO/día) + 0{,}00288\ kgGB + 0{,}57878\ \%GB + \\ + 0{,}5219\ \%P + 0{,}01106\ VO + 0{,}08179\ VT + 0{,}17954\ VT/(haVO + VS) + \\ + 0{,}00587\,\%\mathrm{VR} - 0{,}16402\,\%VM - 0{,}45458\ kgPVEVA - 0{,}05432\ (haVO + VS) - \\ - 0{,}05432\ EGFC - 0{,}07658\ EGC$$

- Componente Principal 3: **Producción en litros**

$$CP3 = 0{,}09508\ Lejer + 0{,}43983\,(L/VO/día) + 0{,}09386\ kgGB + 0{,}00844\ \%GB + \\ + 0{,}11387\ \%P + 0{,}06449\ VO + 0{,}07667\ VT + 0{,}72669\,VT/(haVO + VS) + \\ + 0{,}11217\,\%\mathrm{VR} - 0{,}05955\,\%VM + 0{,}23664\ kgPVEVA - 0{,}12377(haVO + VS) - \\ - 0{,}21760\ EGFC - 0{,}32054\ EGC$$

- Componente Principal 4: **Reposición del ganado**

$$CP4 = -0{,}0592\ Lejer + 0{,}28626\,(L/VO/día) - 0{,}04704\ kgGB + 0{,}18086\ \%GB + \\ + 0{,}27451\ \%P - 0{,}08848\ VO - 0{,}15611\ VT - 0{,}19598\,VT/(haVO + VS) + \\ + 0{,}56464\ \%VR + 0{,}63956\ \%VM - 0{,}01178\ kgPVEVA - 0{,}04488(haVO + VS) + \\ + 0{,}01609\ EGFC + 0{,}05818\ EGC$$

La primera componente principal agrupa aquellas variables que tienen que ver específicamente con la producción de la leche durante el ejercicio anual, que es un indicador productivo a nivel empresarial. La segunda agrupa a las variables que describen la calidad de la leche. La tercera tiene en cuenta a aquellas que miden la producción en litros por vaca, que es un indicador productivo a nivel individual y la cuarta tiene en cuenta la reposición y existencia del ganado por ejercicio.

En la Tabla 5 se observan los coeficientes estimados de cada variable para cada componente principal.

Tabla 5: Coeficientes estimados en cada componente principal.

VARIABLE	**CP1**	**CP2**	**CP3**	**CP4**
Lejer	0,39552	-0,05994	0,09508	-0,05927
L/VO/día	0,07013	-0,32348	0,43983	0,28626
kgGB	0,39825	0,00288	0,09386	-0,04704
%GB	0,05673	0,57878	0,00844	0,18086
%P	0,07240	0,52190	0,11387	0,27451
VO	0,39609	0,01106	0,06449	-0,08848

VT	0,36604	0,08179	0,07667	-0,15611
VT/(haVO+VS)	0,05940	0,17954	0,72669	-0,19598
%VR	0,10526	0,00587	0,11217	0,56464
%VM	0,04801	-0,16402	-0,05955	0,63956
kgPVEVA	0,00017	-0,45458	0,23664	0,01178
haVO+VS	0,38610	-0,05433	-0,12377	-0,04488
EGFC	0,35611	-0,05432	-0,21760	0,01609
EGC	0,29649	-0,07658	-0,32054	0,05818

Como se pudo observar, el análisis de componentes principales mostró que todas las variables incluidas participaron de la determinación de la variabilidad total de los datos.

Cada componente principal se interpreta de acuerdo a la magnitud de los coeficientes obtenidos en cada variable. La primera componente principal puede interpretarse de la siguiente manera:

Se obtendrá un valor alto de la componente principal 1, es decir, de la "producción de ejercicio" al obtener valores altos de las variables:

- litros de ejercicio,
- kilogramos de grasa butirosa,
- vacas en ordeñe,
- vacas totales,
- hectárea de vaca total,
- equivalente grano forrajes conservados
- equivalente grano concentrados

y valores bajos de las variables:

- litros por vacas en ordeñe por día,
- porcentaje de grasa butirosa,
- porcentaje de proteína,
- vacas totales por hectárea de vaca total,
- porcentaje de vacas de rechazo,
- porcentaje de vacas muertas
- kilogramos peso vivo equivalente a vaca adulta.

Las tres componentes principales restantes se interpretan en forma similar.

6. CONSIDERACIONES FINALES

El presente trabajo se pudo realizar gracias a que se contó con registros de observaciones productivas, reproductivas y de información general tomados en empresas tamberas de diferentes regiones del país y a lo largo de los años. De ahí la importancia de contar con bases de datos confiables para la realización de estudios relacionados con la producción tambera de la República Argentina.
En este sentido, las bases de datos de AACREA constituyen un material sumamente valioso para este tipo de análisis tal como se hizo referencia en la introducción. Por lo expuesto, las bases de datos de AACREA constituyen un material de significativa importancia en el país para el análisis de la producción tambera. Las dos variables más utilizadas para medirla, son los litros de leche y los kilos de grasa butirosa producidos por hectárea por vaca por año.

Se hace notar la capacidad de los gráficos radiales para sintetizar visualmente las diferencias entre las variables para cada variable respuesta. Se diferencian claramente los tres estratos productivos.

Con la técnica de Clúster se determinó que 14 variables incluidas en la base de datos se podían agrupar en 4 clústeres, de acuerdo al método jerárquico aglomerativo. El análisis de componentes principales mostró que todas las variables incluidas participaron de la determinación de la variabilidad total de los datos. Se necesitaron cuatro componentes principales para explicar el 76,07% de la variabilidad. Esto demuestra la complejidad de los sistemas biológicos, y en este caso más aún, dado que se trabaja con seres vivos interactuando con distintas regiones, años, etc. Si ambas técnicas son exploratorias y se requerirá de otras técnicas estadísticas para realizar un análisis inferencial de los datos, vale destacar la gran capacidad que tiene tanto el análisis de clúster como el análisis de componentes principales para describir la variabilidad de los datos de una manera adecuada, de tal modo que concuerda con lo esperado.

Por lo tanto se concluye que aplicando cualquiera de las dos técnicas estadísticas se llega a resultados similares, donde se identificó cuatro grupos importantes de variables: producción general del ejercicio, calidad de leche, producción en litros de leche y reposición del ganado.