Esta obra es el resultado de un trabajo académico-investigativo de tres docentes-investigadores, formadores de estudiantes-investigadores para un desarrollo económico y social de la costa Caribe basado en conocimiento, y de grupos dedicados a la investigación. Se tratan aspectos generales de técnicas básicas del Análisis Multivariado de Datos aplicadas a diferentes áreas del sistema de producción vacuno doble propósito; se contribuye a la divulgación de metodologías estadísticas de uso no tan tradicional.

Los autores hacen un análisis detallado de cada concepto para que sea fácil su entendimiento. La información recolectada de las diferentes fuentes de información fue organizada en bases de datos y exportados al programa estadístico R para su descripción y análisis. La Universidad de Sucre será reconocida en el área Estadística (Matemáticas aplicadas) por el compromiso en docencia, investigación, mayor cultura y divulgación con el uso de software libre (SL): R, Rstudio, Rcommander.

Felicitamos a los autores, a Ediciones Universidad Simón Bolívar y a la Universidad de Sucre por su apoyo fundamental.

ISBN 978-958-8930-53-4

EDICIONES UNIVERSIDAD SIMÓN BOLÍVAR
RECONOCIDO POR COLCIENCIAS 2014 - 2017
Universidad de Sucre
Visible para todos

Análisis Multivariado de Datos

Melba L. Vertel M. | Luz M. Botero A. | Jesús A. Cepeda C.

EDICIONES UNIVERSIDAD SIMÓN BOLÍVAR
RECONOCIDO POR COLCIENCIAS 2014 - 2017
Universidad de Sucre
Visible para todos

Análisis Multivariado de Datos

Aplicación: Sistema de producción doble propósito

Melba Liliana Vertel Morinson | Luz Mercedes Botero Arango
Jesús Antonio Cepeda Coronado

MELBA LILIANA VERTEL MORINSON

Licenciada en Matemáticas y Física, Universidad de Córdoba. Magíster en Ciencias-Estadística, UNAL. Investigadora Asociada (I), COLCIENCIAS. Líder, Grupo de Investigación Estadística y Modelamiento Matemático aplicado a Calidad Educativa. Docente Tiempo Completo Exclusiva, Categoría Titular, Área: Estadística, Universidad de Sucre.

LUZ MERCEDES BOTERO ARANGO

Zootecnista, Universidad de Antioquia. Magíster en Desarrollo Rural, Pontificia Universidad Javeriana. Investigadora Junior (IJ), COLCIENCIAS. Investigadora, Grupo de Investigación en Biodiversidad Tropical. Docente Tiempo Completo, Categoría Titular, Área: Doble Propósito, Universidad de Sucre.

JESÚS ANTONIO CEPEDA CORONADO

Matemático, UNAL. Especialista en Educación Matemática, Universidad Distrital Francisco José de Caldas. Investigador, Grupo de Investigación Estadística y Modelamiento Matemático aplicado a Calidad Educativa. Docente Tiempo Completo Exclusivo, Categoría Asociado, Área: Matemáticas Aplicadas, Universidad de Sucre.

EDICIONES UNIVERSIDAD SIMÓN BOLÍVAR

RECONOCIDO POR COLCIENCIAS 2014 - 2017

Análisis Multivariado de Datos

Aplicación: Sistema de producción doble propósito

Melba Liliana Vertel Morinson | Luz Mercedes Botero Arango
Jesús Antonio Cepeda Coronado

EDICIONES UNIVERSIDAD SIMÓN BOLÍVAR
BARRANQUILLA-CÚCUTA, COLOMBIA

EDICIONES UNIVERSIDAD SIMÓN BOLÍVAR
RECONOCIDO POR COLCIENCIAS 2014 - 2017

Análisis Multivariado de Datos

Aplicación: Sistema de producción doble propósito

Melba Liliana Vertel Morinson | Luz Mercedes Botero Arango
Jesús Antonio Cepeda Coronado

Vertel Morinson, Melba Liliana
Análisis multivariado de datos: aplicación: sistema de producción doble propósito / Melba Liliana Vertel Morinson, Luz Mercedes Botero Arango, Jesús Antonio Cepeda Coronado -- Barranquilla: Ediciones Universidad Simón Bolívar, 2016.

223 p.; 17 x 24 cm.
ISBN: 978-958-8930-53-4

1. Análisis multivariante 2. Estadística matemática 3. Análisis estadístico multivariante 4. Ganado vacuno – Investigaciones – Métodos estadísticos 5. Ganadería – Estadísticas 6. Producción animal – Investigaciones – Métodos estadísticos I. Botero Arango, Luz Mercedes II. Cepeda Coronado, Jesús Antonio III. Universidad de Sucre. Grupo de Investigación en Estadística y Modelamiento Matemático aplicado a Calidad Educativa IV. Tít.

519.535 V567 2016 SCDD 21 ed.

Universidad Simón Bolívar – Sistema de Bibliotecas

ANÁLISIS MULTIVARIADO DE DATOS
Aplicación: Sistema de producción doble propósito

ISBN: 978-958-8930-53-4

©Ediciones Universidad Simón Bolívar
Carrera 54 No. 59-102
http://publicaciones.unisimonbolivar.edu.co/edicionesUSB/
dptopublicaciones@unisimonbolivar.edu.co
Barranquilla - Cúcuta

Impresión
Editorial Mejoras
Calle 58 No. 70-30
info@editorialmejoras.co
www.editorialmejoras.co

A este libro se le aplicó
Patente de Invención No. 29069

Junio de 2016
Barranquilla

Printed and made in Colombia

AGRADECIMIENTOS

A Dios. A mi compañero de la fórmula mágica, Jesús Antonio Cepeda Coronado, prueba de que la sabiduría es genética. A mis hijos: Sebastián David, Jesús Manuel y Angie. A mis padres: Melba y Manuel.

Debo expresar un muy especial agradecimiento a la colega y amiga, Luz Mercedes Botero Arango, compañera de luchas en muchas batallas quijotescas, gracias al Altísimo hemos logrado llevar a feliz término nuestros proyectos de vida. Reconocerle en estas líneas que es 'una dura' en muchos temas, pero más en: Sistemas de producción vacuna doble propósito.

A los compañeros de trabajo por su amistad, respeto y cariño: Félix Eduardo Rozo, Juan Alberto Barbosa, Gilberto Carreño y Alfredo Fernández. A compañeros trabajadores como Ángela Tobios, Inés Arrieta, Carolina Navarro, Rocío Acosta, Ivonne Navarro, Kelly Salazar, Jhon Javier Parra, Jhon Pablo Martínez por su entrega al trabajo y por tratar cada día que nuestro barco, la Universidad de Sucre, no se hunda.

A Richard Chávez por su valiosa asistencia en el procesamiento del texto.

Al profesor Luis Felipe Echeverría Martínez, Licenciado en Artes Plásticas, por lograr que sus alumnos de la Universidad de Sucre (Grupo Obregón) puedan expresar a través de la pintura su querer. A mi hijo, Sebastián David, agra-

decerle por la donación de dos de sus pinturas para la portada de este libro.

Agradezco, a mis estudiantes sobresalientes de semilleros de investigación, que incursionaron en el aprendizaje del Análisis de Datos Multivariados bajo la escuela francesa. También, a estudiantes de pregrado y/o las Maestrías de Biología, y Ciencias Ambientales de la Universidad de Sucre, quienes a lo largo de los últimos diez años han colaborado con la revisión, correcciones y sugerencias, contribuyendo ampliamente a mejorar el manuscrito utilizado para producir artículos científicos de alto nivel y publicados en revistas indexadas. Expreso también un gran agradecimiento a Pedro Blanco, Eduar Bejarano y Ricardo Pérez por dejar que sus pupilos utilicen conocimientos estadísticos de alto nivel en sus procesos académico-investigativos.

Y como dice José Martí (1853-1895): *Triste cosa no tener amigos, pero más triste es no tener enemigos. Porque quien enemigos no tenga, es señal de que no tiene, ni talento que haga sombra, ni bienes que se le codicien, ni carácter que impresione, ni valor temido, ni honra de la que se murmure, ni ninguna cosa buena que se le envidie*.

Si A es el éxito en la vida, entonces A = X + Y + Z. Donde X es trabajo, Y es placer y Z es mantener la boca cerrada: Einstein.

Conclusión: La vida es multivariada.

Este trabajo hace parte de la contribución académica del grupo de investigación en Estadística y Modelamiento Matemático aplicado a Calidad Educativa adscrito a la Universidad de Sucre.

Melba Liliana

Con estas notas agradezco a mis padres, que desde el cielo siguen iluminando mi senda. A mis hijos, Ana Isabel y Jesús Manuel, razones de mi existir. A mi fiel amor, Melba Liliana, y a la Universidad de Sucre, que me acogió en su seno y con la que juntos hemos crecido.

Jesús Antonio

Agradezco a mi familia y alumnos por el tiempo que dejé de compartir con ellos. A los colegas por su apoyo desinteresado al aportar sus artículos. A Melba Liliana, que me invita a compartir sus empresas quijotescas y a Dios, por todas sus bendiciones.

Luz Mercedes

CONTENIDO

INTRODUCCIÓN 19

I. SISTEMA DE PRODUCCIÓN VACUNO DOBLE PROPÓSITO..... 25

1. MODELO DE PRODUCCIÓN REGIONAL PARA CONTRIBUIR CON LA SOBERANÍA ALIMENTARIA. 27
   - 1.1. Introducción 27
   - 1.2. Contexto del sistema SPVDP en cifras 27
   - 1.3. El sistema vacuno doble propósito en la seguridad alimentaria 31
   - 1.4. Características del sistema vacuno doble propósito..... 35
   - 1.5. REFLEXIÓN FINAL..... 45

II. COMPONENTES DE DIÁLOGO 47

2. LOS DATOS 49
   - 2.1. Introducción 49
   - 2.2. Ejemplos de tablas de datos. 51
   - 2.3. Ejemplos de variables pecuarias 56
   - 2.4. Análisis estadístico de los datos de una tabla con información pecuaria..... 58
   - 2.5. Planteamiento de la investigación: diseño experimental 61
3. REPRESENTACIÓN GRÁFICA DE LOS DATOS 62
   - 3.1. Introducción 62
   - 3.2. Análisis preliminar de datos..... 63

3.2.1. Estadísticas de resumen 64
3.2.2. Histogramas y diagramas de dispersión 64
3.2.2.1. Histogramas 65
3.2.2.2. Diagrama de dispersión 66
3.2.3. Perfiles 66
3.2.4. Diagrama de tallos y hojas 67
3.2.5. Diagrama de "bigotes": *boxplot* 68
3.2.6. Otras gráficas multivariadas 69
3.2.6.1. Rostros de Chernoff (1973) 70
3.2.6.2. Gráfico de estrellas (*stars plots*) 70
3.2.6.3. Curvas de Andrews. 71
3.3. Diagnóstico de normalidad. 71
3.3.1. Caso univariado 72
3.3.1.1. Métodos gráficos 72
3.3.2. Caso multivariado 77
3.4. Contrastes de normalidad 77
3.4.1. Prueba Ji-cuadrado 77
3.4.2. Kolmogorov-Smirnov 78
3.4.3. Shapiro-Wilk 79
3.4.4. Pruebas de asimetría y kurtosis 81
3.4.5. Posibles soluciones cuando se rechaza la hipótesis de normalidad 83
3.5. Transformaciones para conseguir datos normales 83
4. ANÁLISIS EN COMPONENTES PRINCIPALES (ACP) PONDERADO 84
4.1. Introducción 84
4.2. Objetivos del ACP ponderado 85
4.3. Solución del ACP, mejor plano de proyección. 85
4.4. Fórmulas del ACP(X,M,D) 86
4.5. Ayudas para la interpretación de las gráficas factoriales 87
4.6. ACP ponderados particulares para tablas de datos X 90
4.7. ACP ponderados particulares para relacionar dos tablas de datos (X, Y). 92

5. UTILIZACIÓN DEL ÁLGEBRA Y LA GEOMETRÍA EN ACP(X,M,D) . 94
5.1. Los datos 94
5.2. Representación de los n individuos en (R^p,Ip) 98
5.3. Inercia y contribución a la inercia 99
5.4. El objetivo del ACP y solución 102
5.5. Componentes principales 103
5.6. Representación aproximada de los individuos en el plano factorial 1-2 106
5.7. Calidad de representación 106
5.8. Representación de las p variables en (R^n,D) 108
5.9. El ACP en R^n, espacio de las variables 109
5.10. Representación aproximada de las variables sobre el plano factorial 1-2 109
5.11. Variables en el espacio de representación de los individuos 113
5.12. Individuos y variables suplementarias 114
5.13. Taller de Afianzamiento 117

III. TÉCNICAS MULTIVARIADAS BÁSICAS 129
6. ANÁLISIS EN COMPONENTES PRINCIPALES (ACP) 131
6.1. Dominio de aplicación. 131
6.2. Orígenes del Análisis en Componentes Principales 131
6.3. Fundamentos del método. 131
6.3.1. La nube de variables. 132
6.4. El Análisis en Componentes Principales (ACP) como ACP(X,M,D). 133
6.4.1. Análisis en Componentes Principales (ACP) de datos originales 133
6.5. Elementos suplementarios 134
6.6. Ejemplo de ACP centrado-ponderado 135
6.7. Guía para el análisis de un ACP centrado-normado 143
7. ANÁLISIS DE CORRESPONDENCIAS SIMPLES (ACS) 144
7.1. Análisis de la tabla T: Tabla de frecuencias relativas 147
7.2. Marginales filas y columnas. 148
7.3. Tablas de perfiles fila y columna 150

7.4. El Análisis de Correspondencias Simples, ACM (T) como un ACP (X, M, D) 155
7.5. Guía para el análisis. 162
8. ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES (ACM) 163
8.1. Dominio de aplicación 164
8.2. Fundamentos del método 164
8.3. Tabla disyuntiva completa y tabla de Burt 165
8.4. Análisis de correspondencias de la TDC 168
8.5. El Análisis de Correspondencias Múltiples, ACM (Q) como un ACP(X,M,D) 171
9. INVESTIGACIONES 177

APÉNDICE 195
APÉNDICE I. Software R 195
APÉNDICE II. Script en R para los resultados de los capítulos del libro 199

BIBLIOGRAFÍA 207

LISTA DE FIGURAS

Figura 1. SPVDP- Producción simultánea, armónica, eficiente y sostenible de carne y leche en una sola unidad económica "la vaca" 27
Figura 2. Ganado bovino de carne 30
Figura 3. Ganado bovino en el sistema de producción doble propósito 33
Figura 4. Algunas variables productivas y reproductivas del SPVDP 39
Figura 5. Unidad de producción en el SPVDP 40
Figura 6. Curvas de lactancia, según datos recolectados (PROD) para época 1- orden 3 y modelo matemático polinomial inverso (M7) 41
Figura 7. Esquema de una tabla de datos: X; yuxtaposición de tablas de datos 49
Figura 8. Ejemplos de tablas múltiples 50
Figura 9. Esquema de métodos factoriales 60
Figura 10. Histogramas de analfabetismo y NBI 65
Figura 11. Gráficos generados en el paquete agricolae 65
Figura 12. a. Plano cartesiano de las variables analfabetismo y NBI de la Tabla 7
b. Disperso-grama para los datos socioeconómicos de la Tabla 7 ... 66
Figura 13. Perfiles de la matriz de datos X: indicadores socioeconómicos 67
Figura 14. Diagrama de tallos y hojas para cada variable de la aplicación 68
Figura 15. *Boxplot* e indicadores numéricos de variables socioeconómicas 69

Figura 16. Rostros de Chernoff y gráfico de estrellas para individuos (departamentos) 70
Figura 17. Gráfica de una distribución normal y su relación con un *boxplot* ... 73
Figura 18. Histogramas a partir de datos generados de varias distribuciones .. 73
Figura 19. Histogramas con la densidad normal superpuesta 75
Figura 20. Gráfico cuantil-cuantil para verificar normalidad 76
Figura 21. Representación geométrica de una tabla común a una nube de puntos en el espacio de los individuos y en el espacio de las variables 84
Figura 22. Fotografías alusivas a un plano factorial 86
Figura 23. Yuxtaposición de dos tablas de datos 93
Figura 24. Marco de utilización del análisis canónico de las correlaciones, análisis sobre variables instrumentales y el análisis de co-inercia .. 93
Figura 25. Representación de los individuos 99
Figura 26. Calidad de representación del individuo y_i 106
Figura 27. Representación aproximada sobre el primer plano 1-2 107
Figura 28. Representación de las variables como vectores 109
Figura 29. Representación de las variables 111
Figura 30. Representación aproximada de las variables del ejemplo numérico 112
Figura 31. Representación simultánea en el primer plano factorial de individuos y variables 114
Figura 32. Diagrama de dispersión del ejemplo numérico 117
Figura 33. Nube de puntos centrados del ejemplo numérico 118
Figura 34. Elipse 122
Figura 35. Representación de los datos. Traslación y rotación 122
Figura 36. Plano factorial de los individuos del ejemplo numérico 123
Figura 37. Traslación de los individuos (centrado) del ejemplo numérico 126
Figura 38. *Boxplot* e indicadores numéricos de las variables socioeconómicas; rostros de Chernoff y gráfico de estrellas para departamentos en estudio 137
Figura 39. Plano factorial 1-2 del ACP: filas-individuos y columnas-variables 143

Figura 40. Histograma de los pesos de las filas 150
Figura 41. Histograma de los pesos de las columnas 150
Figura 42. Perfiles filas. Fincas vs. causas de pérdidas 152
Figura 43. Perfiles columna. Causas de pérdidas vs. Fincas 153
Figura 44. Gráficas de los perfiles fila y columna en conjunto 154
Figura 45. Análisis de correspondencias simples de una tabla de frecuencias T 156
Figura 46. Plano factorial 1-2 Fincas vs. causas de pérdidas 158
Figura 47. Subnubes en el primer plano 173
Figura 48. Plano factorial 1-2 del ACM: filas-individuos y columnas-variables 176
Figura 49. Una visión esquemática del funcionamiento de R 196

LISTA DE TABLAS

Tabla 1. Modelos no lineales de curvas de crecimiento: ecuación, estimación de parámetros (A, K, B y M), CME y coeficiente de determinación (R2) 43
Tabla 2. Valores hematológicos promedio en bovinos pertenecientes a fincas evaluadas..... 51
Tabla 3. Causas que afectan las pérdidas en un hato ganadero del sistema de producción doble propósito..... 52
Tabla 4. Valor nutritivo de algunas arbóreas (árboles forrajeros) tropicales a la edad de siete meses..... 53
Tabla 5. Algunas enfermedades infecciosas que afectan el ganado doble propósito en fincas seleccionadas 54
Tabla 6. Tabla de datos codificados y tabulados resultados de aplicar la encuesta 55
Tabla 7. Datos socioeconómicos de departamentos del Caribe colombiano..... 64
Tabla 8. Datos ordenados, cuantiles muestrales y cuantiles poblacionales..... 76
Tabla 9. Fórmulas del ACP(X,M,D)..... 87
Tabla 10. Contribución de las variables a la inercia total..... 98
Tabla 11. Contribución de los individuos a la inercia 100
Tabla 12. Correlaciones de las variables con las componentes principales..... 112
Tabla 13. Datos socioeconómicos de los departamentos del Caribe 136

Tabla 14. Varianza y contribución a la inercia de variables originales y estandarizadas 137
Tabla 15. Tabla de datos de indicadores socioeconómicos estandarizados..... 138
Tabla 16. Matriz de correlaciones..... 138
Tabla 17. Valores propios, inercia acumulada, porcentaje de inercia, y porcentaje de inercia acumulada de los ejes factoriales para el ACP centrado-normado..... 139
Tabla 18. Distancias al cuadrado, contribución de cada individuo a la inercia. 140
Tabla 19. Coordenadas y ayudas a la interpretación de los departamentos..... 140
Tabla 20. Coordenadas y ayudas a la interpretación de las variables..... 140
Tabla 21. Frecuencias relativas..... 147
Tabla 22. Notación y marginales filas de la aplicación en estudio 148
Tabla 23. Notación y marginales columnas de la aplicación en estudio 149
Tabla 24. Notación y valores de los perfiles fila para la aplicación en estudio..... 151
Tabla 25. Notación y valores de los perfiles columna para la aplicación en estudio..... 153
Tabla 26. Perfiles columna..... 153
Tabla 27. Valores propios de la aplicación Causas de pérdidas en ganado bovino..... 157
Tabla 28. Resultados de Encuesta de dinámica de ordeño 165
Tabla 29. Tabla disyuntiva completa del ejemplo 166
Tabla 30. Tabla de Burt (B) 168
Tabla 31. Valores propios, inercia acumulada, porcentaje de inercia, y porcentaje de inercia acumulada de los ejes factoriales 172
Tabla 32. Coordenadas y ayudas a la interpretación de individuos-vacas (filas) 174
Tabla 33. Coordenadas y ayudas a la interpretación de las categorías-variables (columnas) 174

INTRODUCCIÓN

Investigadores en un amplio rango de disciplinas han llegado, por razones científicas diferentes, a la conclusión de que los Sistemas de Producción Vacuna en Doble Propósito (SPVDP) desempeñan un papel primordial en el desarrollo socioeconómico de la región tropical (Aguilera *et al.*, 1999; Koppel *et al.*, 1999). Lo anterior, es evidente en Colombia, donde del total del rebaño lechero, solo el 9,5 % está conformado por razas especializadas (básicamente Holstein y Pardo Suizo), ubicadas principalmente en las zonas altas del trópico, en tanto que el resto del rebaño (90,5 %) está conformado por grupos raciales mestizos (57,6 %) y criollos (32,6 %), ubicados en las zonas medias y bajas del trópico (Arango, 1986; Cuadrado *et al.*, 2003; Silva *et al.*, 2011). En la región Caribe colombiana, donde se encuentra el mayor hato ganadero del país (aproximadamente el 32 % del total), la actividad del doble propósito es de gran importancia económica y de impacto social, representando en la mayoría de sus departamentos más de la mitad del Producto Interno Bruto Agropecuario (Encuesta Nacional Agropecuaria CCI-MADR, 2009).

El proceso de toma de decisiones en este sistema de producción debe estar basado en métodos cuantitativos, especialmente en aquellos que pertenecen a la ciencia de la estadística; usando frecuentemente en investigaciones agropecuarias para el análisis de los resultados técnicas descriptivas que no requieren de cálculos muy complicados (como por ejemplo: proporciones, porcentajes de aparición, medidas de tendencia central, dispersión, entre otros) o estadís-

tica inferencial (intervalos de confianza y prueba de hipótesis) para una o dos poblaciones (Canavos, 1987; Weimer, 1999); conformándose el investigador (aparentemente) con la simple descripción y comparación de grupos a través de análisis descriptivo univariado y bivariado. Estos tipos de análisis, si bien es necesario realizarlos, porque permiten un primer acercamiento a las características de la información trabajada, muchas veces, como se ha podido constatar en diferentes publicaciones, no son siempre los más adecuados para la solución de los problemas propuestos, ni para lograr los objetivos planteados en dichas investigaciones.

La evolución progresiva y sostenida de la informática en los últimos años ha permitido que técnicas de Análisis Multivariado de Datos (ADM) sean más utilizadas en estudios científicos. Desde la conferencia-cursillo de Cabarcas & Pardo (2001), en el Simposio Internacional de Estadística organizado por la Universidad Nacional de Colombia, realizado en Santa Marta (Magdalena), Colombia, nos interesó el ADM bajo la escuela francesa, lo cual nos llevó al punto de hacer tesis de maestría en ADM (Vertel & Pardo, 2010) (Lebart *et al.*, 1995; Escofier & Pagès, 1988-1998; Chessel *et al.*, 2004). El interés en esta temática nos hizo aprender muchas herramientas necesarias en forma autónoma: es el caso del *software R* (*R Development Core Team* 2014), libre y gratis utilizado para la enseñanza y la investigación de la Estadística (Cabrera, 2002), las notas de Correa & Salazar (2000) fueron de gran utilidad para romper el hielo al R. Se tuvo contacto con el ADE4 (Análisis de Datos Ecológicos y Ambientales con procedimientos exploratorios euclidianos) (Thioulouse, *et al.*, 1997) a través de una conferencia-cursillo de Pardo y Ortiz (2004) en *Análisis multivariado de datos en R*, en el Simposio Internacional de Estadística. El grupo de Bioestadística de Lyon implementó ADE4 en R y tiene a disposición gran cantidad de información tanto didáctica como investigación en la página web: http://pbil.univ-lyon1.fr/ADE-4/ADE-4.html.

El Análisis Multivariado de Datos se constituye en una generalización de la estadística descriptiva univariada y bivariada. La interpretación de las representaciones gráficas del ADM requiere del conocimiento de la lógica de los

métodos y están siempre acompañadas de índices numéricos que complementan y enriquecen los análisis. Siendo el objetivo de estos métodos la descripción y exploración de la información no se requiere de modelos preestablecidos, ni de supuestos que muchas veces no se cumplen. Los métodos logran la presentación analógica de la información recurriendo a principios geométricos. La tabla de datos se representa, luego de una transformación adecuada, en un espacio de múltiples dimensiones: *nube de puntos*. En la representación geométrica la distancia entre puntos significa la diferencia entre los elementos considerados: si están cerca se parecen, si están lejos son muy diferentes (Cabarcas & Pardo, 2001). Desde un punto de vista teórico, el mayor aporte del ADM es la obtención de una visión geométrica de los individuos, de las variables y de los métodos estadísticos.

Los investigadores saben que la información pecuaria es esencialmente de naturaleza multivariada, aspectos básicos relacionados con la genética, nutrición, producción, reproducción, sanidad y manejo de vacunos de doble propósito. El SPVDP motiva la inclusión del Análisis de Datos Multivariados para la determinación de grupos homogéneos, en general, para alcanzar una comprensión más profunda acerca de los niveles de rentabilidad, costos, o eficiencia técnica o económica de una explotación vacuna.

En Sucre, la costa Caribe y Colombia, los trabajos de investigación de SPVDP en los que se aplican técnicas de análisis multivariados son muy pocos. Frecuentemente estos estudios consumen muchos recursos, tiempo de productores y de investigadores, por la gran cantidad de información que se necesita resumir y analizar. La determinación de tipologías que clasifiquen y caractericen las explotaciones ganaderas de acuerdo a sus sistemas de producción es importante (Smith *et al.*, 2002), ya que la existencia de una efectiva clasificación y caracterización podría hacer más eficiente la aplicación de algunas políticas gubernamentales. Esto es motivado por el hecho de que productores o explotaciones con diferentes características requieren instrumentos específicos que se adapten a sus particulares necesidades. De igual forma, una política de protección o de apertura hacia el comercio exterior podría tener un impacto

distinto sobre diferentes sistemas productivos. Así, un conocimiento adecuado de estos sistemas permitiría una orientación más precisa de las compensaciones que el Estado le otorgaría a explotaciones más perjudicadas.

Del Programa de Asistencia Técnica a la Comisión Regional de Competitividad del departamento de Sucre (Observatorio del Caribe, 2012) y de otras fuentes, se tomaron ejemplos de propuesta de proyectos donde se podría utilizar el Análisis Multivariado de Datos para realizar análisis e interpretación de resultados*:*

- ✓ Fortalecimiento de microempresas de lácteos y cárnicos en gestión organizativa, empresarial, comercial, desarrollo productivo y tecnológico.
- ✓ Fortalecimiento nacional de los programas de Zootecnia e Ingeniería Agroindustrial como centro de transferencia de tecnología de producción bovina sostenible.
- ✓ Fortalecimiento institucional de colegios técnicos agropecuarios, con el fin de ofrecer mercados técnicos agrícolas y pecuarios altamente calificados y con gran perfil humano y social.
- ✓ Fortalecimiento de programas de investigación, ajustes, transferencia y validación de tecnología integral en SPVDP. En especial: Calidad composicional e higiénico-sanitaria de leche cruda entregada en época seca-lluviosa.
- ✓ Caracterización y diagnóstico de la calidad higiénica, composicional y sanitaria del queso costeño.

Retomando todos estos aspectos, con base en la experiencia docente e investigativa de muchos años en el área estadística y en ciencias veterinarias, y con el producto de publicaciones, premios nacionales y consultas bibliográficas se ha querido elaborar un libro, enfocado en los fundamentos teórico-prácticos del ***Análisis multivariado de datos. Aplicación: sistema de producción doble propósito***, dirigido a estudiantes, técnicos, profesionales, investigadores del área agroindustrial y agropecuaria, que sirva de guía como camino de avance hacia la excelencia.

Se tuvo suficiente información bibliográfica de las técnicas estadísticas multivariadas a tratar. Principalmente escrita en lenguas extranjeras (inglés y francés). Con esto, se contribuye desde las matemáticas aplicadas (Estadística) a la divulgación de técnicas estadísticas de uso no tan tradicional en ciencias aplicadas (pecuaria).

Se contó con información bibliográfica de los sistemas de producción doble propósito aportada por diferentes fuentes: biblioteca digital y bases de datos de la Universidad Nacional por ser sus egresados y redes académicas.

Se tuvo el acompañamiento de la investigadora Luz Mercedes Botero en la asesoría técnica en el área del sistema de producción doble propósito y del investigador Jesús Cepeda en la asesoría matemática. Se contó con el acompañamiento de empresas ganaderas como USATI Ltda., y de revistas especializadas (Revista *MVZ*, categoría A1, Colciencias).

En este libro presentamos técnicas del Análisis Multivariado de Datos aplicadas a diferentes áreas del sistema de producción vacuno doble propósito, contribuyendo a la divulgación de metodologías estadísticas de uso no tan tradicional. La información recolectada de las diferentes fuentes de información fue organizada en bases de datos codificadas y tabuladas en hojas de cálculo. Los datos y variables de individuos en estudio fueron exportados al programa estadístico R (*R Development Core Team* 2014) para su descripción y análisis.

I. SISTEMA DE PRODUCCIÓN VACUNO DOBLE PROPÓSITO

1. MODELO DE PRODUCCIÓN REGIONAL PARA CONTRIBUIR CON LA SOBERANÍA ALIMENTARIA

1.1. Introducción

El Sistema Vacuno Doble Propósito (SPVDP) es aquel que produce simultáneamente carne y leche (Figura 1), donde los animales se alimentan fundamentalmente de gramíneas, leguminosas y arbóreas y en el que las vacas se ordeñan en forma manual una vez al día, requiriendo del apoyo del ternero durante toda la lactancia (ASODOBLE, 1992; Plasse, 1992; Magaña, 1995; Tatis & Botero, 2005; Román-Ponce *et al*., 2013). No se circunscribe a una raza, un cruce, o una especie; de acuerdo a la ubicación geográfica o nivel tecnológico los animales que conforman el hato pueden ser *Bos taurus* o *Bos indicus*, o los cruces resultantes entre ellos.

Figura 1. SPVDP- Producción simultánea, armónica, eficiente y sostenible de carne y leche en una sola unidad económica "la vaca"
Fuente: ASODOBLE (1992)

Las ganaderías del mundo son manejadas en el sistema doble propósito con explotaciones de ovejas o borregas (*Ovis aries*), cabras (*Caprae a. hircus*), búfalas (*Bubalus bubalis*), nak (*Bos mutus*), renas (*Rangifer tarandus*), entre otros mamíferos domésticos de interés zootécnico.

1.2. Contexto del sistema SPVDP en cifras

América Latina sumó una población total de 597,5 millones (en adelante mill) de habitantes en el 2013; según proyecciones para el 2017 la población será de 629,5 mill. Para CEPAL-Naciones Unidas (2011), en el año 2005 solo el

22,2 % de la población era rural y el grado de urbanización de la región era similar a la de los países industrializados; ellos estiman que el porcentaje de la población urbana se estabilizará en torno al 81 % para el año 2020. Pese a que el crecimiento del sector pecuario ha sido dinámico entre el 2003-2007, al crecer 14,2 %, la participación de la agricultura en el PIB de los diferentes países de Latinoamérica ha decrecido al 5,23 % en promedio (Díaz, 2002; CEPAL-Naciones Unidas, 2010; CEPAL-FAO-IICA, 2013).

La población de América Latina deriva el 11 % de sus calorías y el 26 % de su proteína de productos de origen animal. La leche y la carne participan con el 18,7 % del total de la alimentación diaria de la población; la leche representa 97 kg/habitante/año, mientras que todas las carnes contribuyen con 47,5 kg/habitante/año, la carne vacuna con 16,6 kg/año (Bourges, *et al*., 2001; FAO-Banco Mundial, 2001). A pesar de la persistencia de la pobreza urbana, los hábitos alimentarios cambian con la urbanización de la población hacia un mayor consumo de productos de origen animal; de ahí la importancia del desarrollo del sector agropecuario por la cantidad de alimentos que debe producir para una población que se incrementa y que exige mayor proteína animal en la dieta alimenticia.

- Recursos disponibles en América Latina

En conjunto, América del Sur y Centroamérica ocupan 2.050 millones de hectáreas (en adelante ha), que representan el 15,4 % de la extensión global mundial; el área agropecuaria en la región es de 708,7 millones de ha, que equivalen a 14,3 % de la superficie agropecuaria global; de esta el 77,8 % equivalente a 551 millones de ha, están cubiertas de praderas y pastos permanentes (FAO-Banco Mundial, 2001; CEPAL-FAO-IICA, 2013).

Las organizaciones mundiales que velan por la alimentación mundial consideran que las tierras potenciales para abastecer la mayor demanda de carne y leche se alberga en los trópicos templados, pero opinan que mientras África no tiene el desarrollo ni la tecnología para producir más, América Latina sí los tiene, sobre todo en el cono sur.

América Latina cuenta con aproximadamente el 27,9 % de la población mundial de vacunos, unas 387 mill de cabezas, albergados en países con énfasis exportador e importador, y en todos la producción de carne bovina es un rubro decisivo del sector pecuario, con cría realizada fundamentalmente en sistemas pastoriles sin subsidio alguno (Rearte, 2007; OCDE/FAO, 2013).

- Producción de leche y carne en América Latina

En el año 2011 América Latina produjo 81,1 mill de ton métricas de leche líquida de vaca, y 17,4 mill de ton de carne vacuna (CEPAL-FAO-IICA, 2013). Se registran avances sorprendentes en eficiencia de la producción de carne y leche que han contribuido a su aumento durante los últimos 10 años; el rendimiento en leche representa 22 % más, y en carne vacuna 7 % más, que superan con creces los avances logrados en Estados Unidos y el resto del mundo. Según Montero (2013), en los próximos 10 años se espera una mayor producción por vaca; en las zonas tropicales ese incremento dependerá de un mayor número de estos animales que ingresarán al sistema doble propósito.

De acuerdo a OCDE/FAO (2013), las proyecciones indican que la producción de leche continuará aumentando rápidamente en América Latina, donde pasará de 78,7 millones de ton en 2011, a 93,8 mill de ton en 2020, un alza del 20 %, que en razón del incremento en los precios de la energía y los granos, la producción basada en praderas fortalecerá sus ventajas comparativas frente a los sistemas de producción basados en cereales. Otro dato mencionado en este informe dice que actualmente la producción mundial de carnes se sitúa en 288 millones de ton, de las cuales la mayor proporción es carne porcina, seguida por la aviar, la bovina y la ovina.

Puricelli (2011) estima para el año 2014 una producción de carne vacuna a nivel mundial de 58,6 mill/ton, donde Brasil y Paraguay lideran la producción de carne de calidad. Los analistas internacionales consideran que en la siguiente década habrá aumentos en los precios de prácticamente todos los bienes básicos o materias primas (*commodities*) agrícolas. De ahí la importancia de que la ganadería nacional alcance mayores índices de productividad y sustentabilidad.

- La actividad ganadera en Colombia

En el país la actividad ganadera reviste mucha importancia para el desarrollo del campo. En la actualidad está considerada como un sector atractivo para la inversión, por su ubicación geográfica, que le confiere capacidad para alimentar el ganado con base en pastoreo durante todo el año. Colombia cuenta con el tercer hato vacuno de América Latina, con 23.400.000 de cabezas, por debajo de Brasil y Argentina; ocupando el duodécimo lugar en el mundo (FEDEGAN, 2013).

Nuestro país tiene un importante inventario de razas, destacándose el Brahman, Gyr, Holstein, Normando y Pardo Suizo, entre otras (Figura 2). Así mismo, posee un rico recurso zoogenético representado en nueve razas criollas o *Bos taurus* naturalizadas, patrimonio de la ganadería mundial: Costeño con Cuernos, Romo Sinuano, Hartón del Valle, San Martinero, Casanareño, Blanco Orejinegro, Caqueteño, Velásquez y Lucerna.

Figura 2. Ganado bovino de carne
Fotografías: Autores

La ganadería nacional se distribuye en 39,2 mill de ha de pastos y rastrojos y se ubica en los 29 departamentos del territorio nacional. Ese hato vacuno se maneja en tres sistemas de producción: cría y ceba con 13,7 mill; doble propósito con 8,2 mill y lechería especializada con 1,5 mill de reses. Anualmente se sacrifican en promedio 3,9 mill de cabezas, que representa una tasa de extracción de 16,6 % de la que se obtienen 885 miles de ton de carne en canal (Cámara Gremial de la Leche, 2011; Osorio, 2013a y FEDEGAN-FNG, 2013). En leche el sistema doble propósito produce 3.771 mill/litros/año correspondientes al 60 % de la leche total y, la lechería especializada, 2.514 mill/litros/año (40 %). Esta leche es producida por cerca de 400 mil ganaderos, una gran parte de ellos pequeños productores (FEDEGAN-FEP, 2010).

Según el Ministerio de Agricultura y Desarrollo Rural –MADR– (2009), el consumo de carne de res y productos lácteos corresponde al 18 % del gasto de alimentos y el 5 % del total del gasto familiar. El consumo promedio por habitante es de 145 litros/leche/año y 18,1 kg carne vacuna/habitante/año.

La ganadería colombiana equivale a 2,5 veces el sector avícola, 3,3 veces el sector cafetero, 3,2 veces el sector floricultor, 4,9 veces el sector porcícola, 5,7 veces el sector bananero, 9 veces el sector palmicultor; genera 950.000 empleos directos, el 7 % del total del país y más del 25 % del total del agro (FEDEGAN-FNG, 2012).

1.3. El sistema vacuno doble propósito en la seguridad alimentaria

En 1992, Dieter Plasse hablando de la ganadería en el mundo, aseveró: "(…) como todos los mamíferos, los rumiantes producen carne y leche, separarlos es un hecho desafortunado.

Tradicionalmente, los sistemas de producción tanto en Europa como en América Latina, han sido de doble propósito", como manejo de una unidad biológica, donde la vaca y el ternero están juntos durante la lactancia, aprovechando los beneficios del bienestar animal, la mejor nutrición por la leche residual que

toma la cría y la disminución de la mastitis cuando el ternero sella el esfínter del pezón con su saliva al mamar.

Los modelos de producción de doble propósito son especialmente versátiles en un ambiente económico que apunta a la reducción progresiva de los subsidios por parte del Estado. En Europa, desde finales de la década de los 90, los productores de leche solo reciben subsidios estatales si mantienen el ternero al pie de la vaca hasta el destete, en vez de sacrificarlo; es una forma de reducir la producción de leche y las existencias de lácteos sostenidas por el Estado, procurando producir carne para disminuir su importación. En el mundo la producción de carne y leche con el sello "natural" vienen adquiriendo sobreprecios. En el mundo, las razas de doble aptitud se vienen incrementando, al igual que los cruzamientos para los terneros de sacrificio (Maldini & Peixoto, 2011). Los modelos de doble propósito garantizan mayor rentabilidad (Ritchie *et al.*, 2013; Botero M., en Tatis & Botero, 2005) que la explotación exclusivamente lechera y vienen siendo una alternativa importante en la producción de carne.

Según la FAO (2010), hacer que la producción láctea a pequeña escala sea más competitiva, puede ser un arma poderosa para reducir la pobreza, elevar los niveles de nutrición y mejorar los medios de vida de la población rural en los países en desarrollo, lo que se logra con el sistema vacuno doble propósito que tiene espontánea acogida entre los productores rurales de América Latina, contribuyendo con la soberanía y seguridad alimentaria de la población rural.

- Procedencia de nuestro modelo mental

El modelo de desarrollo agrícola de los años 60 se basó en la Revolución Verde. Este modelo se fundamentó en la especialización para la producción de la mayor cantidad de alimentos en forma intensiva y un menor espacio de terreno, sin que el modelo se haya traducido en una disminución del hambre en el mundo.

El traducir el modelo a la ganadería vacuna derivó en dos sistemas de obten-

ción de la leche y la carne: sistema especializado de producción de carne, que originó los *feedlot* o engorde en corral, donde razas especializadas como la Angus reciben grandes cantidades de maíz (*Zea mays*) y soya (*Glycine max*), para lograr incrementar su peso en el menor tiempo posible. Y, los sistemas de lechería especializada con razas como la Holstein, donde la vaca después del parto es apartada de la cría, la cual se sacrifica en gran proporción, luego del parto la vaca se somete a estabulación, ordeño mecánico y con gran parte de la dieta que proviene de balanceados. Estos sistemas requieren poca mano de obra, altos niveles de innovación tecnológica, mucho capital y crédito.

Figura 3. Ganado bovino en el sistema de producción doble propósito
Fotografías: Autores

En contraposición, persistió el sistema doble propósito (Figura 3), que produce carne y leche con la unidad biológica representada en una cría, y leche producida durante una lactancia del ternero. El SPVDP se ubica en 20 departamentos colombianos: en las regiones Caribe y Andina, y en los departamentos de Meta y Caquetá. De acuerdo a un estudio elaborado con más

de 30 mil datos por ASODOBLE (Tatis & Botero, 2005), los promedios de producción del sistema doble propósito fueron: lactancia de 10,5 meses (rango 7-12 meses), 1.200 litros/leche por lactancia (rango 600-1500 lt); un ternero de 145 kg (rango 90-175 kg); intervalo entre parto de 14,5 meses, natalidad 74 %. Aunque los indicadores individuales sean bajos, la producción se alcanza con animales adaptados a las condiciones tropicales: alta temperatura y luminosidad, fuertes inviernos, extensas sequías, grandes distancias recorridas en busca de alimentos.

Según la GTZ-CIAT (1985), la dinámica del sector lácteo en Colombia era decreciente en la respuesta a las necesidades de la población hacia el siglo XXI. Sin embargo, en el año 2011 la producción alcanzó 6.285 mill de litros de leche, de los cuales 10 % se procesaron en finca como queso y suero, el 8 % fueron de autoconsumo, el 45 % en acopio formal de la industria de lácteos, y el 37 % se absorbieron por el sector informal en la venta de leche cruda y sus derivados (FEDEGAN, 2012). El asombroso avance productivo provino del ingreso de vacas al sistema de cría y al sistema doble propósito, permitió brindar seguridad alimentaria, pues junto con la carne vacuna, representaron más del 14 % de las proteínas en la dieta de los colombianos.

Los acontecimientos políticos y económicos de los últimos años han puesto de relieve la vulnerabilidad de la seguridad alimentaria mundial y las perturbaciones importantes en los mercados agrícolas globales y la economía mundial.

La crisis de los precios de los alimentos y de la economía redujo el poder adquisitivo de amplios segmentos de la población en muchos países en desarrollo, se redujo el acceso a los alimentos y se socavó la seguridad alimentaria. Este hecho y el deterioro ambiental, han llevado a replantear los sistemas intensivos de producción, dependientes de la energía fósil y los granos; a la par que se hace necesario mejorar las relaciones políticas entre países para apuntar a una redistribución de los excedentes económicos, mediante precios equitativos y con respeto por la soberanía de los países, de manera que se produzca de acuerdo a su cultura y vocación agropecuaria.

1.4. Características del sistema vacuno doble propósito

El sistema cuenta con una lógica que sobrepasa los mercados y los direccionamientos políticos y económicos; por siglos ha posibilitado la seguridad y soberanía alimentaria de vastos territorios.

Dada esta característica de sostenibilidad económica es importante realizar un inventario de los recursos a los cuales apela el productor: capital, tierra, mano de obra, infraestructura social y productiva, normatividad agropecuaria. Entre los productores del sistema doble propósito ha imperado una lógica de racionalización de la disponibilidad de estos recursos que les ha permitido persistir a través de los siglos, según ciertas estrategias connaturales en el sistema.

Veamos:

Flexible. El productor puede adaptarse a las fluctuaciones del mercado; si hay mayor demanda por leche puede realizar doble ordeño; si hay enlechada o se incrementa la demanda por carne puede suspender el ordeño, toda la leche la tomará la cría y se obtendrá un *baby beef*.

Mano de obra. Los sistemas especializados demandan poca mano de obra pues muchas tareas son ejecutadas mecánicamente.

En el sistema doble propósito se requiere mucho personal, aunque su nivel de cualificación no sea alto, en el caso de reducir este sistema –como muchos técnicos proponen– se ocasionará una pérdida de 200.000 empleos entre la población más pobre.

Al comparar con el sistema cría y ceba, el doble propósito requiere 55 % más de mano de obra por cada 100 reses.

Actualmente, un hombre ordeña a mano 25 vacas /130 litros/ en una jornada de cuatro horas equivalentes a 47.450 litros/año, lo que puede ser un indicador de baja eficiencia al compararlo con uno que manipula una máquina de ordeño

automatizada, pero en cualquier circunstancia no se puede obviar la consideración de lo que representa socialmente el desplazamiento de una población a la que no se le brinda otras alternativas de empleo. Según un estudio realizado por Osorio (2013b), en Brasil el promedio nacional de leche obtenida por un hombre es de 75.000 litros/año y las haciendas de alta tecnología alcanzan niveles superiores a 300.000 litros/hombre/año; entretanto en las lecherías especializadas de Colombia el promedio fue de 97.500 litros/hombre/año, con valores máximos de 190.000 litros/hombre/año. Las diferencias pueden explicarse básicamente por el tamaño del hato y el nivel tecnológico utilizado.

Tierras ganaderas. En el mundo las mejores tierras se deben destinar a la agricultura, la ganadería se debe desplazar a tierras de aptitud pastoril o de silvo-pastoreo.

En América Latina, el doble propósito se ubica en explotaciones con amplio rango de superficie (20-1.000 ha) y diferentes niveles de intensificación (Aranguren-Méndez *et al.,* 2007). Según CEGA (1998), más de las 2/3 partes del inventario de ganado bovino en Colombia corresponden a sistemas de pastoreo tradicional y extractivo, con condiciones de producción caracterizadas por una utilización extensiva de la tierra y el logro de muy bajos niveles de productividad.

La producción animal basada en pasturas se encuentra dividida entre los sistemas extensivos privados típicos de América del Norte, Australia y partes de América del Sur, y los sistemas de libre-acceso en África, Andes, Asia y Siberia, los cuales siguen siendo una opción para los productores tradicionales (Blench, 2001). Lo común en Colombia son los sistemas vacunos con base en pasturas, sin embargo, cabe la posibilidad de introducir tecnologías mínimas y de bajo costo como el pastoreo en rotación, los potreros con arbóreas, el manejo del heno –ensilaje–, bloques multi-nutricionales, permitiendo hacer un uso racional del recurso y liberar parte de las tierras que están dentro del negocio ganadero, pero que tienen vocación agrícola.

Alimentación. Colombia no es, ni será en corto plazo, autosuficiente en producción de granos para biocombustibles y alimentación humana, porcina, avícola. El país consume cada año aproximadamente 4,1 mill/ton/maíz amarillo, 85 % de los cuales procede de mercados externos.

El aumento de la producción ganadera en el mundo, entre 1993-2020, exigirá un incremento de 292 millones de ton/año de cereales para la alimentación animal; además requerirá productos del mar (harina de pescado) y derivados de la matanza (harinas de sangre y hueso) (Bourges *et al*., 2001). Pero los vacunos son eficientes en la utilización de alimentos ricos en fibra y estos deben ser sus alimentos básicos.

Un programa de alimentación animal se debe enfocar al mejoramiento continuo de las condiciones de los animales, que satisfaga sus requerimientos nutricionales (en cantidad y calidad) y les permita un buen desempeño: peso al nacimiento, peso al destete, producción de leche e intervalo entre partos, además de atender a la salud y al bienestar animal. Lo ideal es que el ganado coseche su propio alimento, porque es la forma natural y más económica de hacerlo. Sistemas rotacionales como el modelo Voisin facilitan cosechar el pasto en el punto óptimo, aumentan la producción por unidad de área e incrementan la persistencia de la pradera.

Los vacunos alimentados a libre voluntad consumen 70 %-85 % de gramíneas y 15-30 % de leguminosas y otras especies forrajeras (Botero M., 2011). Por lo tanto, hay ventajas con potreros biodiversos que ofrecen variedad de especies, lo que a su vez beneficia la fauna benéfica y reduce la presencia de plagas y parasitosis. Esto lo permite el sistema doble propósito cuando desde temprana edad los bovinos tienen acceso a bancos de arbóreas como: leucaena (*Leucaena leucocephala*), guácimo (*Guazuma ulmifolia*), totumo (*Crescentia cujete*), pipón o hueso de gallina (*Lonchocarpus santae-mantae*), tamarindo de monte (*Dialium guianense*), acacio panoramero (*Acacia spp*), guacamayo (*Piptadermia flova*), hobo (*Spondia mombi*) y uvito (*Cordia alba*).

Tal tipo de alimentación es lo que permite que en el sistema doble propósito sea posible homologar la producción –con solo pasto– de 5 litros de leche para venta por el valor de kilo del ternero destetado, de mínimo 150 kg. Además del necesario uso de sales mineralizadas con base en las deficiencias de los suelos de cada zona, los granos y tortas proteicas pueden ser un complemento en la dieta básica, siempre y cuando sea económica y socialmente viable.

Calidad de la carne y la leche. Alimentos producidos a base de gramíneas y otras plantas son más orgánicos; en el proceso se obtiene grasa polinsaturada (omega 3 y omega 6). Al utilizar pocos químicos de síntesis, hay mayor bienestar animal y la leche tiene mayores sólidos. Los ácidos linoleicos conjugados, ALC o CLA, por sus siglas en inglés, son una familia de por lo menos 28 isómeros del ácido linoleico, encontrados en los productos cárnicos y lácteos que provienen de bovinos, ovinos, caprinos y bufalinos, según investigaciones de Correa en Botero M. (2010).

De acuerdo con Contexto ganadero (2013),

> en las zonas del trópico alto colombiano, la leche tiene niveles elevados de ácido linoleico conjugado, que es considerado desde ya, un componente nutracéutico porque genera innumerables beneficios al consumidor final; datos preliminares de estudios realizados en nuestro país, sugieren que el lácteo de este trópico tiene mayores niveles de dicho ácido que el promedio mundial (Boletín del 2 de diciembre de 2013).

Genética. La ganadería colombiana de la región Caribe es hoy básicamente cebuina y la ganadería de carne y el doble propósito está influenciada por esa raza. Para Botero M. (2011), no es que el cebú sea malo, pero se debe apelar masivamente al vigor híbrido; continúa explicando que existe un genotipo animal adecuado para cada ambiente, por lo tanto es imposible formular un solo genotipo superior que encaje en los diversos ecosistemas del país, lo razonable y sostenible es adaptar el genotipo al ambiente, no lo contrario, buscar adaptar el ambiente al genotipo. El departamento técnico de ASODOBLE

(1992) concuerda con Koger *et al.* (1975) y Molinuevo (2003), cuando concluyeron: "la máxima producción por unidad de área de pastoreo se obtiene con animales de tamaño medio, producciones medias y alta fertilidad".

Por tanto, hay que reformular la selección orientada por animales de gran tamaño y alta producción, pues son factores que afectan en forma negativa la reproducción, característica que es la de mayor importancia económica en la empresa ganadera (Figura 4). En el SPVDP se hace uso del cruzamiento de razas buscando dos objetivos básicos: complementar sus características deseables y hacer uso del vigor híbrido, que incrementa la producción con un mínimo de insumos adicionales.

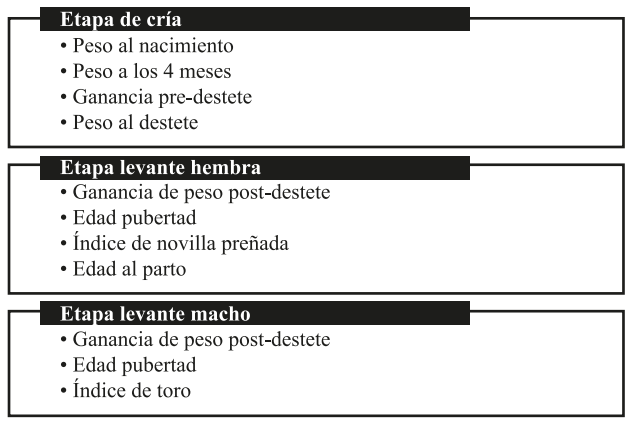


Figura 4. Algunas variables productivas y reproductivas del SPVDP.
Fuente: Elaboración propia

Cuando las condiciones ambientales y de manejo son óptimas, se puede tener mejoramiento de la producción avanzando hasta 3/4 de sangre europea, pero cuando las condiciones son más adversas no se debe superar el ½ sangre como lo expresó Madelena (1986). Lo ideal es tener un animal con 50 % *B. indicus* x 25 % *B. taurus* naturalizado x 25 % *B. taurus* especializado.

Eficiencia reproductiva. En todos los sistemas vacunos en Colombia la edad al primer parto es alta, cercana a los 40 meses de edad. De acuerdo con Botero M. (2010a), la eficiencia reproductiva es mayor en el doble propósito frente al

sistema de cría especializada, por el efecto benéfico en la presentación de celo de la vaca cuando se realiza el aparte del ternero.

Se reconoce que la nutrición y la lactancia tienen efectos sobre la reproducción (Figura 5) y existe abundante literatura internacional que vincula al amamantamiento, la presencia permanente del ternero (Williams *et al.*, 1990) y la subnutrición energética (Wettemann & Bossis, 2000; Hess *et al*., 2005) con el alargamiento de los días abiertos en bovinos (Tatis & Botero, 2005).

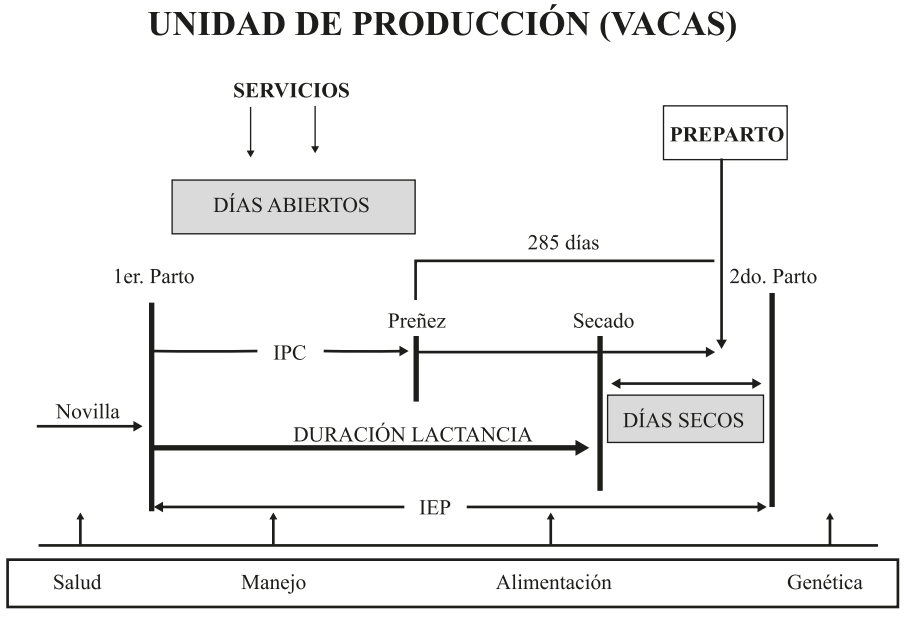


Figura 5. Unidad de producción en el SPVDP.
Fuente: USATI (2004)

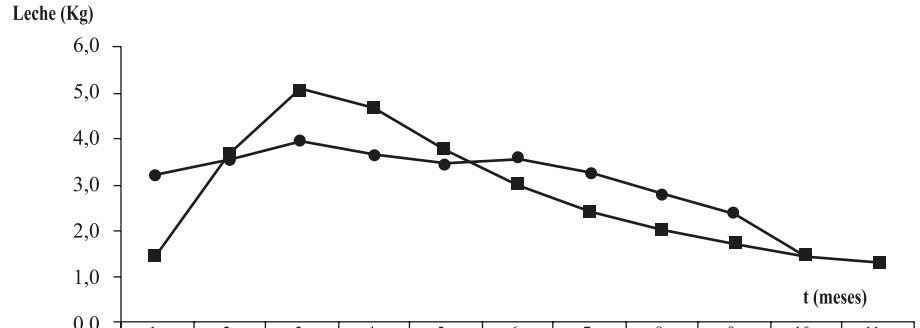
Producción de leche. En el 90 % de los hatos el ordeño se realiza de forma manual, con el ternero al lado como apoyo para que la vaca inicie el proceso de vaciado de la leche; cuando el ordeño finaliza, la vaca ha escondido entre 15-25 % de la leche total, lo que se conoce como leche residual que solo el ternero es capaz de extraerla.

Esta leche residual se pierde en las lecherías especializadas y se aprovecha

en las de doble propósito. Algunos investigadores y políticos sostienen que el SPVDP debe desaparecer por su supuesta baja producción.

Para Botero M. (2010), al hacer el siguiente análisis puede conducir a otras conclusiones: si en el país se producen 15.000.000 litros/día se debe contar con un hato de 500.000 vacas especializadas que produzcan 30 lit/día (hoy el promedio de producción está en 7 lit/vaca/día), y si se logra incrementar en un litro (1 lit/vaca/día) se aumenta la producción nacional en 500.000 lit/día. Mientras tanto, el ordeño de 5.000.000 vacas en el sistema doble propósito que producen 3 litros/día (hoy el promedio de producción está en 2,7 lt/vaca/día), con el incremento de la producción 1 lit/vaca/día, representará el aumento de 5.000.000 lit/día.

Botero y Vertel (2006), con información recolectada durante 10 años y después de utilizar los datos de las lactancias completas de 500 hembras vacunas mestizas, graficaron la curva de lactancia para el SPVDP (Figura 6).



	1	2	3	4	5	6	7	8	9	10	11
Prod	3,17	3,56	3,96	3,67	3,47	3,57	3,24	2,78	2,34	1,43	1,32
M7	1,41583	3,68596	5,07185	4,68933	3,76761	2,97944	2,40294	1,98758	1,6824	1,4521	1,27369

Figura 6. Curvas de lactancia, según datos recolectados (PROD) para época 1- orden 3 y modelo matemático polinomial inverso (M7).
Fuente: Botero y Vertel (2006)

Encontraron que el modelo polinomial inverso es el que mejor caracterizaba la curva de lactancia, por presentar los mayores valores para el estadístico Durbin-Watson y coeficiente de determinación (R^2). A partir de ella, concluyeron

que a medida que se aumenta el número de partos por hembra hay un incremento en la producción inicial de leche, y que la producción máxima inicial la tienen a partir del tercer parto.

A diferencia de las hembras con alto mestizaje *Bos indicus,* las vacas con alto mestizaje *Bos taurus* presentaron mayores producciones de leche, pero una menor persistencia. En los dos grupos genéticos el pico de producción está entre el tercero y cuarto mes, con una producción máxima que se incrementa paulatinamente a medida que aumenta el número de partos. Concluyeron que la curva de lactancia está influenciada por la raza o cruce de ellas, por la edad de la hembra, el año y la época de parto, que son responsables de mudanzas en el pico de producción, su persistencia y pendiente; sin embargo fue evidente el efecto del manejo zootécnico que recibía el hato en ordeño sobre las variables de la curva.

Producción de carne. Aunque los terneros del sistema cría tienen mayor peso al destete (gdp= 0,7 kg/día), esa ventaja se reduce al posdestete por el bajo desarrollo del rumen. En contravía, los provenientes del sistema doble propósito por efecto compensatorio igualan el peso a los 24 meses de edad. Sin embargo, el sistema cría tiene una tasa de crecimiento y desarrollo lento de sus crías, lo que afecta la cantidad de carne para la venta, la edad al primer servicio de las hembras, el porcentaje de fertilidad del hato y la eficiencia reproductiva en general, tanto en machos como en hembras.

Para determinar la curva de crecimiento de crías del sistema doble propósito, Botero & Vertel (2007) utilizaron 52 crías macho vacuno, que fueron pesadas en forma individual mes a mes después del ordeño de sus madres, sirviendo de apoyo y aprovechando la leche residual. Se obtuvieron 468 datos que fueron ajustados a modelos no lineales (Botero *et al.*, 2014) para estimar el crecimiento del animal y los parámetros de la curva (Tabla 1). Se determinó que en la ganancia de peso mes a mes hay un incremento superior a 1 kg/día/cría desde el nacimiento hasta el mes de vida, cuando cría y vaca permanecen juntas gran parte del día.

Tabla 1. Modelos no lineales de curvas de crecimiento: ecuación, estimación de parámetros (A, K, B y M), CME y coeficiente de determinación (R^2)

Modelos no lineales	Ecuación	*A*	*K*	*B*	*M*	*CME*	R^2
Gompertz	$Y = A * \exp^{(-b * \exp^{-Kt})}$	312,28	0,87	0,04	-	191,96	0,75
Brody	$Y = A * (1 - b * exp^{-K})$	181,44	1,47	0,15	-	192,23	0,75
Logístico	$Y = A * (1 + b * exp^{-K})^{-1}$	154,79	2,57	0,26	-	192,75	0,75
Bertalanffy	$Y = A * (1 - b * exp^{-Kt})^{3}$	194,93	-0,37	0,12	-	192,53	0,75
Richards	$Y = A * (1 - b * exp^{-Kt})^{M}$	203,53	-0,46	0,11	2,6	192,57	0,73

Fuente: Freitas (2005)

A partir de este momento, las ganancias de peso son poco significativas (promedio 0,32 kg/día) lo cual lleva al pobre desempeño del peso final al destete. Se considera que la tasa deseable de ganancia de peso en crías es de 0,5 kg/día en el doble propósito (Tatis & Botero, 2005), lo cual demuestra el potencial productivo que tienen ante una oferta ambiental adecuada es alto.

Esta es una de las debilidades del sistema, por lo tanto se hace necesario diseñar estrategias para mejorar el peso del ternero al destete, pero no incrementando el suministro de leche, sino potenciando desde temprana edad el hecho de ser rumiante, y aportando mayor cantidad de proteína a partir de bancos de arbóreas forrajeras, leguminosas y no leguminosas, donde el ternero pueda estar y consumir en las horas de la tarde, después de ser apartado de su madre.

Viabilidad de las crías. La muerte de terneros es responsable de pérdidas económicas considerables: puede variar entre 5-50 %, y está asociada a un gran número de factores, pero los principales son el tiempo transcurrido desde el nacimiento hasta la primera ingesta de calostro, el tamaño del rebaño, y el personal a cargo de la crianza.

El incremento de más de 50 % de sangre europea y el hacinamiento aumentan la mortalidad, la cual en hatos de pequeños productores de doble propósito es menor, debido a la observación constante, a mañana y tarde, que se tiene de la cría durante el ordeño y encierro de los animales, lo cual posibilita una atención a tiempo.

Según Tomas Preston (1976) una mortalidad menor a 5 % en condiciones de trópico bajo, es un porcentaje aceptable, e intentar reducirlo supone el gasto de muchos insumos y cargar en el hato animales no adaptados.

Rentabilidad. Los sistemas de producción bovina requieren de la implementación de estrategias que apliquen principios biológicos, matemáticos y económicos para optimizar la productividad.

Para esto, es necesaria la constante observación, la permanente toma de datos y el acertado análisis de todos los eventos que ocurren dentro de la empresa, para decidir adecuadamente.

Para obtener el costo de producir un litro de leche en el sistema doble propósito, Botero y Rodríguez (2006) diseñaron una metodología para calcularlo en la región Caribe, como herramienta de medición de la eficiencia del sistema. La metodología consideró la estructura de costos, incluyó el costo del arrendamiento de la tierra, el valor de los kilogramos de carne que deja de ganar el ternero durante la etapa de lactancia respecto al sistema cría donde no se ordeña la vaca.

Al aplicar la metodología encontraron que el costo medio anual de producir leche fue US$ 0,11/litro/leche producida (48,38 % costos fijos y 51,62 % costos variables) y que existe una relación de inversa proporcionalidad entre los costos y el volumen de producción y venta de leche, con un coeficiente de determinación de 82,34 %. El 47,36 % del precio de venta de la leche equivalió al costo de producirla. El punto de equilibrio en unidades producidas e ingresos es de 29,47 % (60 litros/día), respecto a la producción diaria de leche.

Martínez (2013) reconoce la inquietud existente entre los productores de lechería especializada y afirma:

> la producción de leche aquí (trópico alto) es muy costosa, y ese es el Talón de Aquiles en competitividad. El futuro del sector, en un entorno de apertura

> comercial, depende de las medidas que se tomen para reducir los costos de producción lechera y en fomentar su competitividad (p.1).

Investigación. Debe destacarse el papel que jugó la ciencia, la investigación y la tecnología en el desarrollo del cerrado brasileño. “La gran conclusión es que la tierra no influyó en el desarrollo de los cerrados, sino la tecnología e investigación. Todos los productores tuvieron la posibilidad de usar la tecnología”. Sin embargo, en Colombia muchos investigadores se han dedicado a desdeñar el sistema, faltando mucho por investigar para determinar su importancia y rumbo a seguir. El Estado y las entidades privadas deben destinar más recursos a la investigación del sistema doble propósito (Elves, 2013).

Algunos de los obstáculos más importantes al aumentar el aporte de la ganadería a la seguridad alimentaria y la lucha contra la pobreza en América Latina se relaciona con la falta de acceso a tecnología, crédito, recursos de tierra, mercados, información y capacitación.

Los productores deberán recibir la suficiente educación, capacitación y transferencia de tecnología apropiada, de tal forma que sean más receptivos y se pueda generar confianza hacia las instituciones públicas, la academia y los profesionales del sector.

Para zanjar la brecha que se ha forjado entre los productores y la institucionalidad, es necesario tener en cuenta en la formulación de políticas para el sector, la necesidad de concertar con los productores ya que ellos poseen un “saber” que han construido a través del método de error-éxito resultado del diario vivir con los diferentes factores que impactan la producción. Es un conocimiento que quizás no les ha permitido insertarse al mercado global, pero que en todo caso sí les ha reportado contribuciones a su seguridad alimentaria.

1.5. REFLEXIÓN FINAL

Es necesario que en Colombia se comprenda la necesidad de trabajar de forma asociativa, y que se vea el impacto sobre todas las cadenas productivas.

Para la FAO (2010), la ganadería seguirá contribuyendo a la seguridad y soberanía alimentaria, la reducción de la pobreza y el crecimiento general de la región.

No obstante, advierte que se requieren diversas medidas para que los pequeños productores mejoren sus niveles de vida y puedan contribuir con la seguridad alimentaria de un país. Estas son: mejora en la infraestructura y desarrollo de sistemas confiables de transporte y *marketing* entre las zonas rurales y los mercados; mejor acceso a sistemas de comunicación e información para apoyar la toma de decisiones; mayor acceso a crédito, a nuevas tecnologías y nuevos insumos de producción; implantación de servicios ampliados de extensión agraria para proporcionar capacitación y asistencia técnica urgente en crianza, producción, *marketing* y gestión ganadera.

II. COMPONENTES DE DIÁLOGO

2. LOS DATOS

La naturaleza de las filas y columnas de una *tabla de datos* junto con los objetivos del estudio determinan los métodos ADM a utilizar.

2.1. Introducción

En *Ciencias Agropecuarias y Agroindustriales* se busca documentar con datos los fenómenos que están siendo observados sobre poblaciones, muestras o grupos. La medición de varias características de un mismo individuo, unidad de observación o unidad experimental, ya sea en forma simultánea o con ciertos intervalos de tiempo, genera una serie de datos que deben ser analizados con técnicas multivariadas.

Los datos originales de las características evaluadas pueden ser de tipo cuantitativo o cualitativo. De cualquier manera, la información sobre las unidades de observación se transforma en *tablas de datos.* Una *tabla de datos* (Figura 7) generalmente tiene *filas* que representan los *individuos*, y *columnas* que representan las *variables,* las cuales pueden ser continuas o nominales según la *escala de medición* (nominal, ordinal, intervalo o razón).

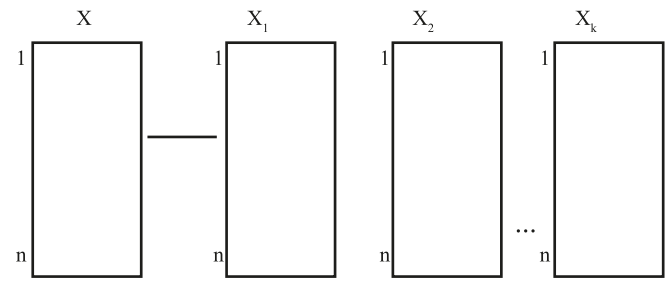


Figura 7. Esquema de una tabla de datos: **X;** yuxtaposición de tablas de datos
Fuente: Escofier & Pagés (1988-1998)

Los datos de una tabla pueden ser de naturaleza muy diversa:

- Tablas de *individuos x variables continuas*.
- Tablas de *frecuencias* (tablas de contingencia, ausencia-presencia, conteos, porcentajes).

- Tablas de *individuos x variables cualitativas*.
- Tablas *múltiples* (Figura 8).

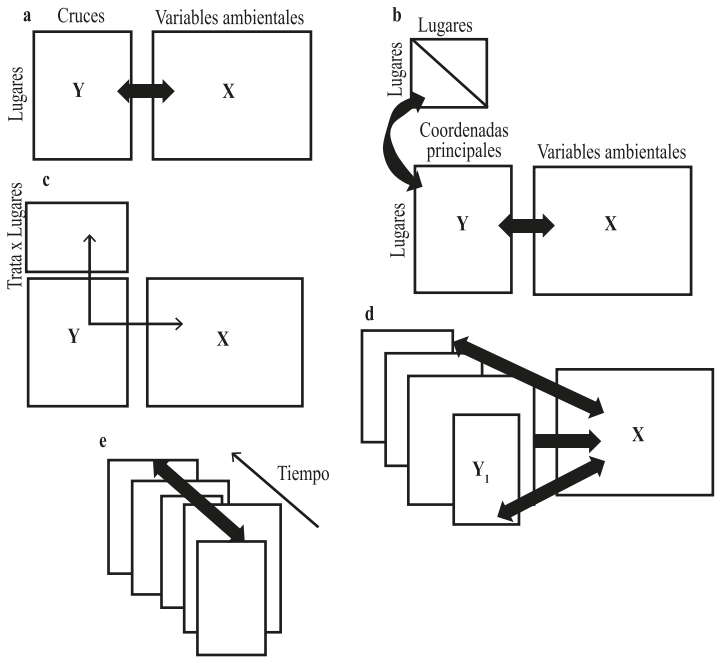


Figura 8. Ejemplos de tablas múltiples
Fuente: Dray & Chessel (2003); Dray *et al.* (2003)

El individuo puede ser una parcela de experimentación, una finca, un animal, una planta, una porción de terreno, y las características serán una serie de atributos, mediciones, evaluaciones, estimaciones, tratamientos o propiedades correspondientes a esos individuos (unidades experimentales).

Lo que interesa a los investigadores, en primera instancia, es hacer una lectura de la información contenida en la *tabla de datos* y entonces son pertinentes las siguientes preguntas:

- ✓ ¿Cómo abordar la información que hay en una tabla de datos?
- ✓ ¿Cómo leer una tabla de datos?, es decir, ¿qué información importante hay en una tabla de datos en relación con los objetivos del estudio?
- ✓ ¿Cómo obtener un mensaje que pueda ser luego verbalizado por el investigador y que sirva para la comunicación de los resultados?

Para que el cerebro humano pueda captar lo más importante de la información de una tabla hay que consentir perder información para ganar en significación (Cabarcas & Pardo, 2001). Este órgano entiende mejor la información en forma análoga, es decir, en forma gráfica, en lugar de la información digital o el conjunto de cifras de una tabla (Ver capítulo 3. Representación gráfica de datos multivariantes). Los métodos estadísticos exploratorios multidimensionales hacen posible la interrelación entre múltiples variables.

2.2. Ejemplos de tablas de datos

A continuación, se presentan ejemplos hipotéticos del área pecuaria de *tablas de datos* de baja dimensión (pocas filas, pocas columnas para entender más adelante la lógica de los métodos).

Ejemplo 1. Tabla de datos cuantitativos

El Análisis en Componentes Principales (en adelante ACP, Capítulo 6) permite analizar la información de una tabla de tipo "individuos x variables cuantitativas". Se tiene por costumbre escribir en las filas a los "individuos", los que representan las unidades estadísticas en un análisis. Por ejemplo, al estudiar la hemoparasitosis en la productividad bovina (Alfaro *et al.*, 1999), se evalúan valores hematológicos promedio en bovinos pertenecientes a fincas muestreadas (Tabla 2).

Existe variabilidad entre fincas y dentro fincas, lo cual es necesario considerar para los tratamientos preventivos y curativos. Estos son los indicadores hematológicos correctos para todas las fincas.

Tabla 2. Valores hematológicos promedio en bovinos pertenecientes a fincas evaluadas

Finca (Individuos)	Hematocritos (%)	Hemoglobina (%)	Glóbulos rojos (mm^3)
Muri	36,23	11,90	5969231
Sunsun	31,04	10,34	5177272
El diamante	29,12	9,68	4829167
Bienfresca	29,97	9,68	4932353
Altamira	38,11	12,66	6294737
Queregua	36,50	12,16	6046667
La cañada	36,12	12,01	5970833

Valores de referencia: Hematocritos 30-44 %, Hemoglobina: 9-14 %, Glóbulos rojos: 5-6 x 10 mm^3
Fuente: Alfaro *et al.* (1999)

También se puede estudiar la composición química de tipos de pastos para determinar su influencia en la producción de carne (Cuadrado *et al.*, 2005), caracterizar condición bovino-métrica (Medina, 2005), evaluar el efecto de la administración por vía parenteral de un compuesto de sulfato de cobre sobre hepatología, hemoquímica y bioactividad del líquido ruminal (García *et al.*, 2005), caracterizar funcionalidad reproductiva de toros por grupos de edad x cruce genético (Vejarano *et al*., 2005).

Ejemplo 2. Tabla de frecuencias

Inicialmente el Análisis de Correspondencias (Capítulos 7 y 8) fue concebido para analizar, describir y representar gráficamente el cruce de dos variables cualitativas. Su objeto de estudio era, y es, la Tabla de Contingencia (TC). Este análisis se ha ido extendiendo al estudio y descripción de tablas de números positivos (*tablas de frecuencias*) de muy diversos estilos: conteos (Tabla 3), porcentajes (Tabla 4), presencia (Tabla 5).

Caso 1. Datos para conteos (frecuencias absolutas)

La sobrevivencia es uno de los pilares fundamentales de la eficiencia productiva. Una alta mortalidad se traduce en bajo número de hembras de reemplazo y pocos machos para la venta (Alfaro *et al.,* 2004, Alfaro *et al.,* 2006, Pérez-Hernández *et al.,* 2003).

El objetivo del estudio puede ser identificar *principales causas de pérdidas* en el primer año de vida (Tabla 3).

Tabla 3. Causas que afectan las pérdidas en un hato ganadero del sistema de producción doble propósito

Causas / Individuo	A-Accidentes	B-Desconocidas	C-Enfermedades infecciosas, gastrointestinales y parasitarias	D-Deficiencias nutricionales	E-Problemas genéticos
Finca 1	15	54	231	149	0
Finca 2	30	29	126	51	1
Finca 3	12	51	533	125	0

Fuente: Adaptado de varios autores

Caso 2. Datos para porcentajes (frecuencias relativas)

El conocimiento de la importancia de los arboles nos han llevado a cambiar el control de malezas, conservando las plantas arbóreas más gustosas y productivas (Tatis & Botero, 2005). ¿Recomendarían los expertos según los datos recolectados (Tabla 4), seguir estudiando la producción y gustosidad de las arbóreas para escoger las superiores?

Tabla 4. Valor nutritivo de algunas arbóreas (árboles forrajeros) tropicales a la edad de siete meses

Parámetros / Especies	Materia seca: MS (%)	Proteína cruda: PC (%)	Digestibilidad *in vitro*: DIG (%)
Gliricidia maculata	22,9	25,7	68,9
Leucaena leucocephala	24,5	33,3	68,6
Erythrina glauca	22,6	20,8	57,2
Albizia lebbeck	37,4	21,6	60,8
Diphysa robinioides	33,8	22,9	66,3
Crescentia cujete	28,9	15,0	47,6
Guazuma ulfimolia	28,1	18,3	61,5
Samanea saman	41,4	21,4	38,4
Enterolobium cyclocarpum	36,7	26,2	57,2

Fuente: Shneichel y Sebert. ICA-GTZ (1990)

Caso 3. Datos presencia-ausencia

El objetivo del estudio es identificar enfermedades infecciosas más frecuentes en las fincas (por causas directas o indirectas). Sí – presencia, No – Ausencia (Tabla 5).

Se pueden analizar aspectos relacionados con:

- El impacto de las enfermedades animales en países en desarrollo (reducción en el rendimiento del producto, baja en la eficiencia reproductiva, disminución en la vida productiva y en la calidad de productos o servicios);
- Los limitantes de la producción ganadera en el país (manejo de la alimentación, salud animal y mejoramiento genético);
- Las principales enfermedades limitantes de la producción pecuaria en las Américas: fiebre aftosa, rabia, brucelosis, tuberculosis, estomatitis vesicu-

lar, leucemia, carbón bacteridiano, rinotraqueitis infecciosa (IBR), paratuberculosis, tricomoniasis, vibriosis leptospirosis, colibacilosis, mastitis (Páez *et al*., 2002), pasteurelosis, salmonelosis.

Tabla 5. Algunas enfermedades infecciosas que afectan el ganado doble propósito en fincas seleccionadas

Causas / Individuo	Rabia	Aftosa	Brucelosis	Mastitis	Leptospirosis
Finca 1	0	0	0	1	1
Finca 2	1	0	0	1	1
Finca 3	1	1	1	0	0
Finca 4	0	1	0	1	1
Finca 5	0	1	1	1	1
Finca 6	0	1	0	0	0

Fuente: Adaptado de varios autores

Ejemplo 3. Tabla de datos cualitativos

Las "encuestas" se organizan en torno a "unidades temáticas" que resultan del cuadro conceptual del estudio. Las "unidades temáticas" de una encuesta determinan la estrategia de observación, pero también la estrategia de análisis de datos. Por medio de "encuestas", se elaboran tablas que resumen las *p* características observadas sobre *n* unidades de observación. La construcción del objeto de estudio se hace mediante el tratamiento de la información contenida en esas tablas (datos cualitativos).

El Análisis de Correspondencias Múltiples (Capítulo 8) es un instrumento adaptado al tratamiento estadístico de los datos producidos por vía de "encuestas". Este método de análisis estadístico responde a una doble exigencia: objetividad en el proceso de reducción y de exploración de lo observado, y tratamiento de la información con el nivel de síntesis adecuado al cuadro conceptual utilizado (Crivisqui, 1993).

Como parte de la caracterización higiénico-sanitaria y microbiológica de sistemas de producción bovina, se realiza encuesta estructurada a productores de alguna(s) provincia(s) específica(s), ejemplos: Herrero *et al*. (1998); Páez *et al.* (2002) y Botero *et al.* (2012).

Se presenta parte de una encuesta correspondiente a la dinámica del ordeño para diferentes fincas en estudio.

DINÁMICA DEL ORDEÑO						
En el ordeño utiliza ternero (OrduTer)	SÍ		NO			
Sitio de ordeño (SitOrd)	Corral		Potrero			
Lava el sitio de ordeño (LaSitOrd)	SÍ		NO			
EN EL ORDEÑO						
Se lava las manos antes del ordeño (LMaAOrd)	SÍ		NO			
Se lava las manos entre vacas (LMaEVa)	SÍ		NO			
Realiza lavado de pezones (LaPezo)	SÍ		NO		A veces	
Fuente de agua utilizada (FuAgUt)	Pozo		Acueducto		Corriente	
Realiza secado de pezones (SecPez)	SÍ		NO		A veces	

Fuente: Adaptado de varios autores

Los datos registrados para las fincas escogidas se presentan en la Tabla 6.

Tabla 6. Tabla de datos codificados y tabulados resultados de aplicar la encuesta

Individuo	OrduTer	SitOrd	LaSitOrd	LMaAOrd	LMaEVa	LaPezo	FuAgUt	SecPez
Finca 1	Sí	Corral	No	No	No	No	Pozo	Sí
Finca 2	Sí	Potrero	Sí	No	No	Sí	Pozo	Sí
Finca 3	Sí	Potrero	Sí	Sí	No	No	Corriente	Sí
Finca 4	No	Corral	Sí	Sí	Sí	No	Pozo	Sí
Finca 5	Sí	Potrero	Sí	No	Sí	Sí	Pozo	Sí
Finca 6	No	Corral	No	Sí	No	Sí	Pozo	No

Fuente: Adaptado de varios autores

Se busca una caracterización de la calidad higiénico-sanitaria de la leche de ganado vacuno entregada en época seca por productores pecuarios.

Otros ejemplos donde se podría usar este tipo de análisis son: Smith & colaboradores (2002), caracterizar algunas variables cualitativas relacionadas a los sistemas productivos lecheros de Chile; Herrera *et al.* (2008), estudiar la dinámica de los hemoparásitos en bovinos teniendo en cuenta aspectos importantes como el municipio de procedencia de los animales, el sistema de explo-

tación a partir de la raza del animal, época climática-año y tipo de parásito; caracteriza el efecto de factores no genéticos y grupo racial, sobre la Vida Útil (VU) de vacas doble propósito (Murcia & Martínez, 2013).

2.3. Ejemplos de variables pecuarias

Los registros son básicos e imprescindibles en el manejo de una empresa agropecuaria, pues permiten identificar a tiempo los aciertos, desaciertos y oportunidades de mejora, por lo que son una herramienta en la proyección y en la toma de decisiones de una empresa ganadera. Existen diversas variables de alimentación, composicionales, productivas, reproductivas, sanitarias, etc., que permiten obtener información necesaria sobre la productividad económica de la empresa ganadera por fechas (meses o años), por hatos, por épocas climáticas, por cruces, por tipos de alimentación.

Composicional

Acidez (mL de NaOH 0.1N%)
Crioscopia (°H)
Cloruros (%p/v)
Densidad (g/mL a 15°)
Grasa (%)
Sólidos totales (%); Sólidos no grasos (%)
Producción (Kg)
Tiempo de Reducción del Azul de Metileno (TRAM)
Tratamientos: Clase I. Leche fría con 4 h.; Clase II. Leche fría con >2-4 h.; Leche caliente 30° a 2 h
Meses: Enero, febrero, marzo, abril, mayo, junio, julio, agosto septiembre, octubre, noviembre, diciembre
Departamentos: Córdoba, Sucre, Bolívar, Cesar, Magdalena, Atlántico, La Guajira.

Reproductivas

Natalidad o parición real (%);
Natalidad o parición estimada (%)
Intervalo entre partos (IEP)

Número de vientres paridos
Número de servicios por preñez o concepción
Preñez al primer servicio (%)
Abortos o pérdidas prenatales (%)

Productivas

Tasa de sobrevivencia por categoría (ejemplo: número animales vivos mayores de 1 año, número animales muertos mayores de 1 año)
Número de terneros nacidos año contable, Número de terneros muertos año contable
Número de terneros destetados
Destete (%)
Peso al nacimiento, Peso al 1, 2, 3, 4, 5, 6, 7, 8, 9° mes de nacimiento
Peso al destete
Número de vientres en el hato, Tasa de desecho, Mortalidad anual de vientres (%)
Tasa de desecho o descarte anual de vientres
Tasa anual de mortalidad
Vida útil o productiva de vientres
Edad al primer parto, Vida útil o productiva
Edad al sacrificio
Cantidad de leche producida por lactancia (kg)
IEP (días)
Producción de leche por día de intervalo entre partos
Peso del ternero al destete
Producción de carne por día de IEP
Número promedio de vacas lactantes por mes
Número de ordeñadores necesarios en ordeño manual
Total de hectáreas en pastoreo (ha/UA), Producción de materia seca (MS) por hectárea
Peso de una Unidad Animal (UA)
Capacidad de carga animal

Días de descanso por potrero, Días de ocupación por grupo, Número de grupos de animales
Número de pastoreos por potrero/año
Número de días de descanso de la pastura, Número de grupos de animales
Número de potreros/finca
Número de hectáreas efectivas de pastoreo/finca
Área por potrero
Cantidad de heces producidas /hato
Cantidad de orina producida /hato
Relación entre toros o detectores de celo a vientres aptos

Variables cuantitativas

Estacionalidad (días), Grado de confinamiento (0-24 meses), Producción/vaca, Praderas artificiales/Praderas totales, Carga animal, Índice de mecanización, Índice de construcciones, Producción/mano de obra.

Variables cualitativas

Época de concentración de partos
Dedicación del propietario al rubro lechero
Nivel de estudios del propietario
Nivel de estudios de los ordeñadores
Método de encaste de las vacas lecheras
Sistema de frío para la leche
Empleo de algún sistema de control lechero
Número de ordeños diarios en invierno
Uso de la terapia de secado.

2.4. Análisis estadístico de los datos de una tabla con información pecuaria

Los investigadores saben que la información de los fenómenos agropecuarios es esencialmente de naturaleza multivariada, y no la utilizan por falta de conocimiento o de manejo de software. Por eso el objetivo de este texto es hacer una introducción al lenguaje, lógica y aplicación de los métodos mul-

tivariados, para que los investigadores puedan recurrir a ellos y recurran a la implementación en un software libre y gratis al que toda la comunidad académico-científica puede acceder.

Lebart *et al*. (1995) han acuñado para estos métodos el nombre de exploratorios multidimensionales, pero se usó mucho en el pasado el de *análisis de datos* y es sinónimo de *estadística descriptiva multivariada* o *análisis multivariado de datos.* Se constituyen en una generalización de la estadística descriptiva univariada y bivariada, pero la presencia de más variables o dimensiones la hace más compleja. La interpretación de las representaciones gráficas requiere del conocimiento de la lógica de los métodos y están siempre acompañadas de índices numéricos que complementan y enriquecen los análisis. En otras palabras, la utilización de estos métodos requiere de un entrenamiento para su implementación e interpretación y hace prácticamente indispensable el trabajo interdisciplinario en la investigación. La utilización adecuada de estos métodos es posible cuando una sola persona juega el doble papel: de investigador social y estadístico (Cabarcas & Pardo, 2001). Los métodos logran la presentación analógica de la información recurriendo a principios geométricos.

La tabla de datos se representa, luego de una transformación adecuada, en un espacio de múltiples dimensiones: *nube de puntos* (Figura 9). En la representación geométrica la distancia entre puntos significa la diferencia entre los elementos considerados: si están cerca se parecen, si están lejos son muy diferentes. Sobre una tabla de datos son posibles dos representaciones complementarias: *la nube de los puntos fila* y *la nube de los puntos columna*.

Para ubicar un punto en el plano se requieren dos coordenadas, en el espacio tres coordenadas y en un espacio abstracto de p dimensiones, p coordenadas. El conjunto de las coordenadas necesarias para un punto se denomina *vector*. En una tabla de n filas y p columnas, se tiene una nube de n puntos filas en donde cada fila está representada mediante un vector de p coordenadas y una nube de p puntos columnas con cada punto representado por un vector de n coordenadas.

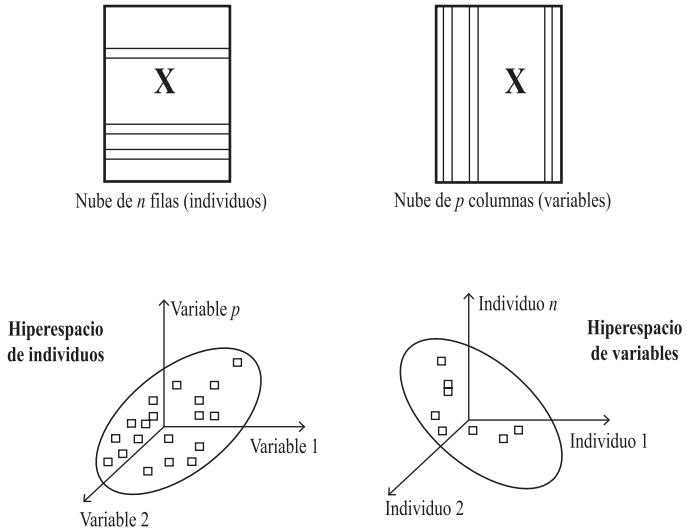


Figura 9. Esquema de métodos factoriales
Fuente: Dray y Dufour (2007)

Las nubes de puntos construidas son abstractas pues no podemos ver espacios de más de tres dimensiones, en realidad, en nuestros documentos vemos bien dos dimensiones (planos). Pero la geometría abstracta de esas representaciones multidimensionales cumple con las mismas propiedades de la geometría plana y del espacio.

Se recurre entonces a proyecciones sobre planos y a agrupamientos de puntos cercanos, para observar lo más importante de esas representaciones multidimensionales. La lectura utilizando proyecciones es el principio de los *métodos factoriales*, en cuyo caso la pérdida de la información se manifiesta en forma de errores de proyección. En los métodos factoriales se busca el plano para el cual los errores de proyección son en conjunto los menores posibles: *primer plano factorial*.

La naturaleza de las filas y columnas de una tabla de datos junto con los objetivos del estudio determinan los métodos multivariados a utilizar. Dentro de los métodos factoriales el más útil en ciencias aplicadas es el Análisis de

Correspondencias Múltiples (ACM), ya que es el adecuado para la lectura de tablas de "individuos" por variables cualitativas (nominales u ordinales). El ACM es una generalización del Análisis de Correspondencias Simples (ACS), utilizado para la lectura de tablas de frecuencias (conteos, porcentajes, presencia-ausencia, tablas de contingencia). El ACS se puede ver como la aplicación simultánea de dos análisis en componentes principales (ACP). Aún bajo el supuesto de que un investigador esté interesado en la aplicación del ACM es necesario conocer previamente los otros dos métodos. En la mayoría de las aplicaciones se utilizan métodos de clasificación que dan lecturas complementarias a los métodos factoriales, de la tabla de datos.

2.5. Planeamiento de la investigación: diseño experimental

Las investigaciones pecuarias muchas veces tienen en su planeación y desarrollo la implementación del diseño experimental. En este tipo de investigaciones se recoge información de muchas variables, que hace imprescindible su relación para elaborar planes de prevención, control y mejoramiento.

Entonces se tiene un camino a seguir cuando sea así, primero diseño experimental para la investigación planteada en el área del conocimiento y después análisis multivariado de datos para la relación o caracterización del fenómeno.

La importancia de los pasos iniciales. El planeamiento es la fase más importante de un trabajo científico, pues es donde se hace la justificación de los tratamientos que se pretende comparar. Una revisión bibliográfica de los trabajos en el área, es muy importante pues permite situar el problema en el contexto general. En el Capítulo 8 se muestran investigaciones publicadas en libros y revistas científicas, en los que se utilizó en el planeamiento de la investigación el diseño experimental. Para el análisis estadístico se realiza un análisis individual de las variables con diseño experimental y posteriormente un análisis multivariado de las variables. En esta última (Análisis Multivariado) se hace uso del conocimiento y la programación de los métodos factoriales objeto de este libro.

El planeamiento debe contener:

a) Introducción. Enfatizar la importancia del estudio, situación o problema. Relación de preguntas y respuestas enfatizando las hipótesis que estarán a prueba. Evitar la ambición excesiva (resolver todos los problemas en un solo ensayo).
b) Relación de los objetivos principales. Relación de los objetivos a ser estudiados: generales y específicos.
c) Descripción de la metodología y del material. Tratamientos, diseño, repeticiones, parcelas o unidades experimentales, material experimental a ser utilizado, tamaño de la parcela, medidas a ser efectuadas, método de muestreo, software.
   Esquema del análisis estadístico: *Estadística descriptiva univariada* (tablas de frecuencia, gráficas e indicadores numéricos); *Estadística inferencial* para el diseño experimental elegido (Esquema del análisis de varianza con los grados de libertad, así como las pruebas que serán utilizadas); *Estadística multivariada* (relacionar variables de factores de estudio).
d) Conclusiones que se esperan obtener con el trabajo.

3. REPRESENTACIÓN GRÁFICA DE LOS DATOS

3.1. Introducción

En este capítulo veremos la descripción de datos multivariantes, estudiando su representación gráfica y las posibles transformaciones de las variables que conduzcan a una interpretación más simple de los datos.

También introduciremos un análisis inicial de la homogeneidad de la muestra, mediante el estudio de los posibles valores atípicos, debido a errores de medida, u otras causas de heterogeneidad.

El objeto y materia prima del trabajo estadístico está contenido en los datos, los cuales suministran información referente a un objeto, en un tiempo determinado.

Resultan entonces tres componentes estadísticos: de un lado están *los objetos*,

con los que se intenta desarrollar algún estudio, por otro lado, *las características o atributos* inherentes a los primeros y finalmente, *el momento u ocasión* en que están inscritos los dos primeros (objeto y variable).

Se puede concebir entonces una colección de información sobre un objeto con un atributo en un tiempo. Un punto corresponde al valor del atributo j-ésimo, para i-ésimo individuo, en el instante *t*. En general, los procedimientos estadísticos consideran constantes o fijos algunos de los tres componentes señalados.

Cuando se fija el tiempo y se recolecta información (características o atributos) sobre un conjunto de individuos estudiados por la mayoría de las *técnicas de análisis multivariado*, a veces se les llama estudios *transversales*. Cuando se estudian las variables en diferentes tiempos, se ocupan los métodos de series cronológicas (estudios longitudinales).

3.2. Análisis preliminar de datos

En esta sección estudiaremos algunas funciones útiles para el análisis inicial de los datos, tales como *summary()*, *plot()*, *hist()*, entre otras.

Para la explicación de las definiciones utilizaremos los datos para un **Perfil socioeconómico de los departamentos caribeños**, los cuales contribuyen en un alto porcentaje a la producción de la ganadería doble propósito. Es necesario fomentar la organización entre productores pecuarios mediante esquemas que propicien su integración a la industria. Se requiere que el productor sea proveedor constante tanto de leche como de carne de calidad y que sea beneficiario del valor agregado generado en el procesamiento (Pérez-Hernández *et al.*, 2003).

En este breve informe se presentan algunos indicadores económicos y sociales de la región Caribe colombiana, y para efectos de este documento, dicha región está conformada por los departamentos de Atlántico, Bolívar, Cesar, Sucre, Córdoba, Magdalena, La Guajira y el archipiélago de San Andrés y Providencia.

Datos de observación. En el ejemplo de aplicación se tienen siete departamentos del Caribe como individuos, los cuales están descritos en % por variables socioeconómicas registradas en el Censo de 2005 (Departamento Nacional de Planeación, 2014): Necesidades Básicas Insatisfechas (NBI), Analfabetismo, Desempleo y Producto Interno Bruto (PIB) *per cápita*. El archipiélago de San Andrés y Providencia no se tuvo en cuenta por presentar indicadores socioeconómicos atípicos (Tabla 7).

Tabla 7. Datos socioeconómicos de departamentos del Caribe colombiano

	Atlántico	Bolívar	Cesar	Córdoba	La Guajira	Magdalena	Sucre
NBI	16,1	30,0	35,7	35,8	37,5	30,6	42,4
Analfabetismo	4,5	9,5	13,7	15,8	14,4	13,4	15,3
Desempleo	12,8	9,8	6,2	13,6	5,9	7,1	6,2
PIB per cápita	1,67	8,59	1,36	1,20	3,02	0,75	0,80

Fuente: Vertel (2012)

3.2.1. Estadísticas de resumen

Primero estudiaremos la función *summary()*, la cual entrega las estadísticas de resumen básicas (Díaz & Morales, 2012).

Nótese que todas las variables son numéricas.

```
summary(datos)
      NBI+            analfab         desempleo        PIB perc++
 Min.   :16.10   Min.   : 4.50   Min.   : 5.9   Min.   :0.750
 1st Qu.:30.30   1st Qu.:11.45   1st Qu.: 6.2   1st Qu.:1.000
 Median :35.70   Median :13.70   Median : 7.1   Median :1.360
 Mean   :32.59   Mean   :12.37   Mean   : 8.8   Mean   :2.484
 3rd Qu.:36.65   3rd Qu.:14.85   3rd Qu.:11.3   3rd Qu.:2.345
 Max.   :42.40   Max.   :15.80   Max.   :13.6   Max.   :8.590
```

+: Las variables en R deben ser codificadas hasta con 8 caracteres alfanuméricos.
++: No se recomienda tildar o utilizar la letra ñ.
Fuente: Elaboración propia

3.2.2. Histogramas y diagramas de dispersión

El primer paso de cualquier análisis de datos multivariado es representar gráficamente las variables individualmente, mediante un histograma o un diagrama de caja (*boxplot*). Estas representaciones gráficas son muy útiles para detectar asimetrías, heterogeneidad, datos atípicos, etc.

3.2.2.1. Histogramas

Si queremos un histograma de NBI o analfabetismo, usamos la función *hist()*. Igual, se puede hacer para cada una de las variables cuantitativas (Figura 10).

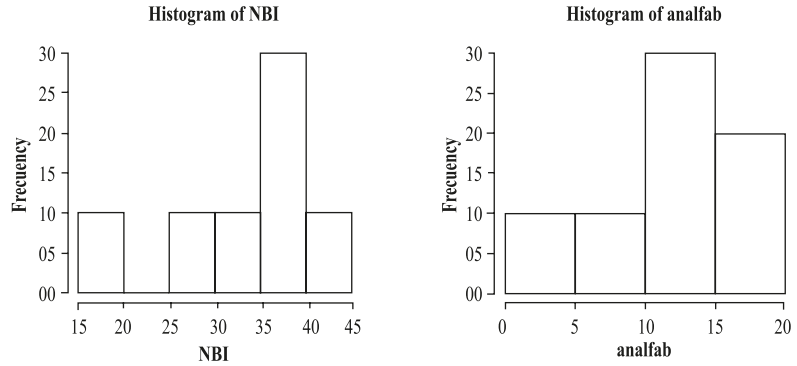


Figura 10. Histogramas de analfabetismo y NBI.
Fuente: Elaboración propia

Hay otro paquete llamado *agricolae* (Mendiburu, 2010) que se puede instalar en el software R (2014). En primer lugar, el investigador se formula una serie de preguntas, que espera tener respuesta al iniciar, conducir y culminar el experimento que genera una gran cantidad de variables (cualitativas, cuantitativas). Ofrece *estadística descriptiva* para datos agrupados (resumen numérico y gráfico de cada variable numérica) y *diseño experimental* (Kuelh, 2001; Vertel, 2005). Por ejemplo: en la Figura 11, se muestran gráficos generados en el paquete *agricolae*.

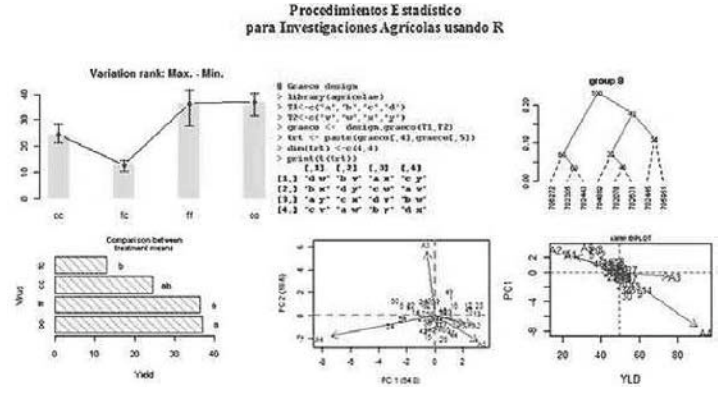


Figura 11. Gráficos generados en el paquete *agricolae*
Fuente: http://tarwi.lamolina.edu.pe/~fmendiburu

3.2.2.2. Diagrama de dispersión

Cuando se dispone de dos variables en un plano cartesiano es relativamente sencillo, en estadística este tipo de gráficos se les llama *diagramas de dispersión*. Es tal vez el más antiguo de los gráficos multivariados. Está limitado a la presentación de dos variables, aunque se pueden realizar modificaciones de tal forma que nos permita incluir más. Tomaremos las variables *analfabetismo* y *NBI* para ver su relación a través de la Figura 12a.

En la Figura 12b se han hecho dispersogramas (Díaz & Morales, 2012) por pares de variables. Además, en estas gráficas se puede advertir la posible asociación lineal entre pares de variables.

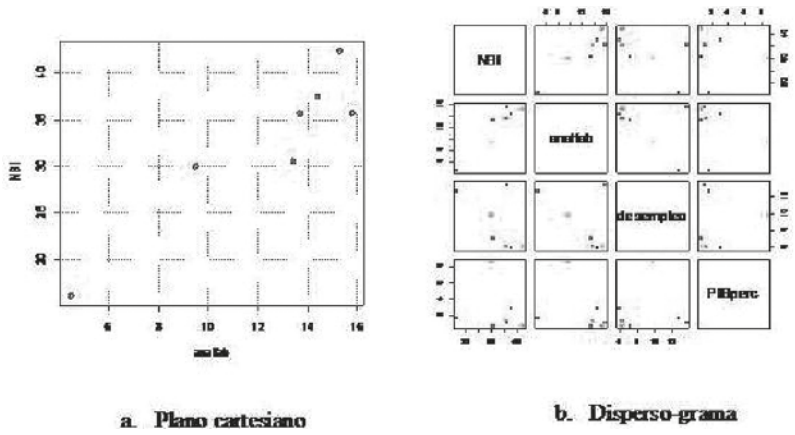


Figura 12. a. Plano cartesiano de las variables analfabetismo y NBI de la Tabla 7; b. Disperso-grama para los datos de variables socioeconómicos de la Tabla 7
Fuente: Elaboración propia

En R obtenemos el diagrama de dispersión (plano cartesiano) con la función plot: plot(x, ...) plot(x,y,xlim=range(x),ylim=range(y),type="p", main,xlab,ylab,...) plot(y ~ x, ...)	En R obtenemos el disperso-grama con la función pairs: pairs(datos, col = rainbow(5), pch = 15) ó plot(datos, col = rainbow(5), pch = 15)

3.2.3. Perfiles

Los perfiles se presentan a manera de histogramas, donde cada barra corresponde a una variable y su altura al valor de la misma: *barplot* (). A veces en

lugar de barras se construye una línea poligonal. Cada diagrama corresponde a un conjunto. La Figura 13 muestra los perfiles para los datos de la matriz del ejemplo de *indicadores socioeconómicos*.

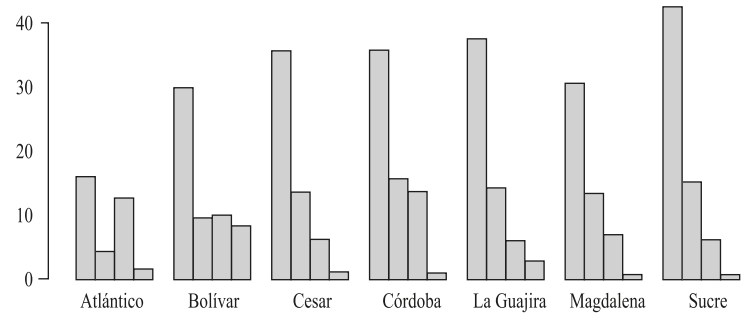


Figura 13. Perfiles de la matriz de datos **X:** indicadores socioeconómicos
Fuente: Elaboración propia

En R: barplot(t(datos),beside=TRUE, col=rainbow(5)).

3.2.4. Diagrama de tallos y hojas

Un *diagrama de tallos y hojas* (Weimer, 1999) ofrece una forma novedosa y rápida de exhibir información numérica, sirve para ordenar datos y calcular medidas de posición. Un **punto de posición** es un punto tal que un cierto porcentaje de datos cae antes que él. El procedimiento para construirlo es el siguiente:

1. Redondear los datos convenientemente en dos o tres dígitos significativos.
2. Disponer los datos en una tabla con dos columnas como sigue:
   a) Para datos con dos dígitos, escribir en la columna izquierda los dígitos de las decenas, este es el tallo, y a la derecha, después de una línea o dos puntos, las unidades, que son las hojas. Así por ejemplo 58, 58 se escribe 5|8 o 5:8.
   b) Para datos con tres dígitos, el tallo estará formado por los dígitos de las centenas y las decenas, las cuales se escriben en la columna izquierda, separados de las unidades (hojas). Por ejemplo, 236 se escribe 23|6 o 23:6.
   c) Cada tallo define una clase, y se escribe una sola vez. El número de hojas representa la frecuencia de dicha clase.

A continuación en la Figura 14, se muestran diagramas de tallos y hojas de las variables socioeconómicas de los departamentos costeños (Tabla 7): NBI, analfabetismo, desempleo y PIB *per cápita*. En R, la función es *stem*().

NBI	Analfabetismo	Desempleo	PIB *per cápita*
1 | 6	0 |	4 | 9	0 | 88247
2 |	0 | 5	6 | 221	2 | 0
3 | 01668	1 | 0344	8 | 8	4 |
4 | 2	1 | 56	10 |	6 |
		12 | 86	8 | 6

Figura 14. Diagrama de tallos y hojas para cada variable de la aplicación
Fuente: Elaboración propia

3.2.5. Diagrama de "bigotes": boxplot

Un diagrama (Figura 15) de estos consiste en una caja y guiones o segmentos (Weimer, 1999). El extremo inferior de la caja es el primer cuartil () y el superior, el tercer cuartil ().

Los segmentos o bigotes se extienden desde la parte superior de la caja a valores adyacentes; es decir, la observación más pequeña y la observación más alta que se encuentran dentro de la región definida por el límite inferior y el límite superior. Las observaciones atípicas son puntos fuera de los límites inferior y superior, los cuales son señalados con estrellas (*).

En la práctica, no esperaríamos una muestra de datos perfectamente simétricos, donde el lugar ocupado por la mediana es un buen indicador de la simetría. Si la distribución está sesgada a la izquierda, la mediana queda a la derecha del centro de la caja, y si está segada a la derecha, la mediana está a la izquierda del centro de la caja. Se pueden construir estos diagramas para variables conjuntamente. Este tipo de gráficas facilitan la lectura sobre localización, variabilidad, simetría, presencia de observaciones atípicas e incluso asociación entre variables, en un conjunto de datos.

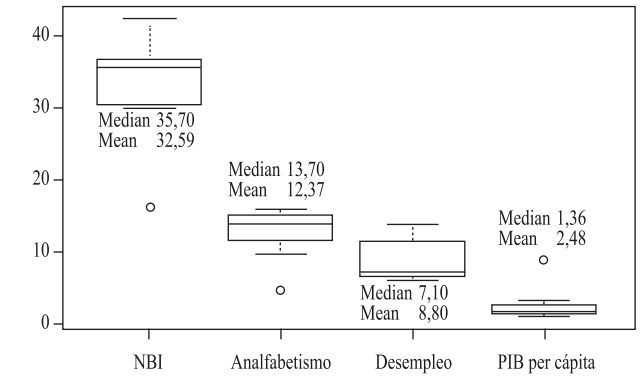


Figura 15. *Boxplot* e indicadores numéricos de variables socioeconómicas
Fuente: Elaboración propia

La investigación de observaciones atípicas revela a menudo información útil y es bastante posible que una de ellas sea la *joya entre las piedras* en lugar de la *piedra entre las joyas*.

Estas observaciones pueden afectar tanto la media como la desviación estándar del conjunto de datos, distorsionando así el centro y la variabilidad; no hay consenso entre los investigadores sobre lo que constituye una observación atípica en un conjunto de datos.

Una de las dos reglas prácticas siguientes es de uso típico para detectar observaciones aberrantes en un conjunto de datos.

Regla 1. El tamaño de muestra es mayor a 10, la distribución de frecuencia tiene forma de campana y el puntaje *z* para la medida dista más de tres desviaciones estándar de la media.

Regla 2. La medida cae más de tres (rango intercuartil) debajo del cuartil menor, o más de tres IQR arriba del cuartil superior.

3.2.6. Otras gráficas multivariadas

Finalmente, hay gráficas muy populares como las caras de Chernoff (1973) y de estrellas (Figura 16), donde se asocia a cada variable o bien un rasgo de una cara (facilidad con que distinguimos facciones) o bien parte de una estrella.

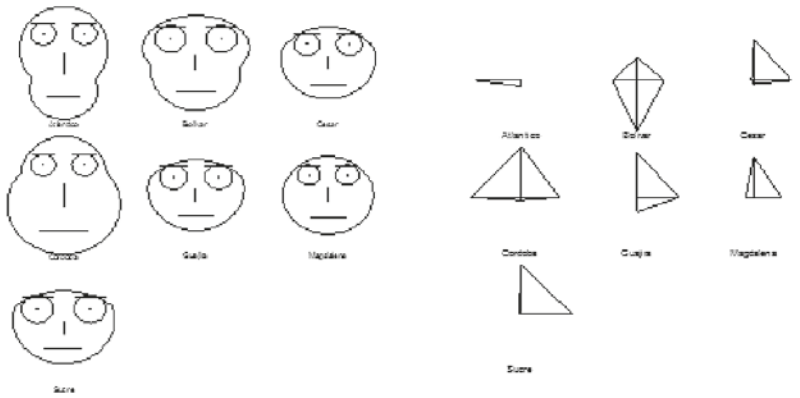


Figura 16. Rostros de *Chernoff* y gráfico de *estrellas* para individuos (departamentos)
Fuente: Elaboración propia

3.2.6.1. Rostros de Chernoff (1973)

Asocia a cada variable una característica del rostro; tal como la longitud de la nariz, tamaño de los ojos, forma de los ojos, ancho de la boca, entre otros. La Figura 16 presenta siete objetos mediante siete rostros.

En R: Necesita la librería: *library(TeachingDemos)* y la función es *faces2()*.

*3.2.6.2. Gráfico de estrellas (*stars plots*)*

Supóngase un conjunto de datos multivariados ordenados matricialmente, de manera que las filas corresponden a las observaciones y *p* columnas, una por cada variable.

En dos dimensiones se pueden construir círculos (uno por cada observación multivariada) de un radio prefijado, con *p* rayos igualmente espaciados emanando del centro de cada círculo. Las longitudes de los rayos son proporcionales a los valores de las variables en cada observación. Los extremos de los rayos pueden conectarse con segmentos de líneas rectas para formar una estrella. Con cada observación representada por una estrella, estas pueden ser agrupadas según sus similitudes. Es conveniente estandarizar las observaciones, caso en el cual pueden resultar valores negativos (Peña, 2002; Díaz & Morales, 2012).

3.2.6.3. Curvas de Andrews

Un gráfico de Andrews está basado en una transformación de Fourier del conjunto de datos multivariados. Básicamente una transformación de Fourier es una representación funcional alternante de senos y cosenos, de cada observación. La transformación se define como:

$$f(t) = \frac{x_1}{2} + x_2\, sen(t) + x_3 cos(t) + x_4 sen(2t) + x_5 cos(2t) + \cdots \qquad (1)$$

Cada variable de cada observación es representada por una componente individual en la suma de la transformada de Fourier. Tradicionalmente, *t* varía entre $-\pi$ y π para permitir una adecuada representación de los datos. La magnitud de cada variable de un sujeto particular afecta la frecuencia, la amplitud y la periodicidad de *f*, dando una representación única para cada sujeto.

3.3. Diagnóstico de normalidad

La mayoría de procedimientos estadísticos que se usan habitualmente suponen que los datos observados proceden de una población con distribución normal.

Una razón de ello es que muchas variables asociadas a fenómenos naturales y sociales siguen aproximadamente esta distribución, otra razón (quizás la más importante) del uso extendido del supuesto de normalidad es la facilidad y *elegancia* con que se obtienen los estimadores y los procedimientos para la inferencia y prueba de hipótesis.

Aunque muchas de las técnicas estadísticas son poco sensibles a la violación del supuesto de normalidad (en general, este supuesto puede obviarse cuando se cuenta con un tamaño de muestra grande –resultados asintóticos–), es recomendable contrastar siempre si se puede asumir o no una distribución normal. El diagnóstico del supuesto de normalidad incluye desde la simple exploración visual de los datos hasta técnicas estadísticas sofisticadas que ayudan a decidir si es razonable suponer la normalidad del conjunto de datos en cuestión.

En el análisis multivariado de datos no se requiere de modelos preestablecidos, ni de supuestos (normalidad, por ejemplo). Aunque si los datos para la descripción multivariada provienen de un experimento controlado donde aplican diseño experimental y contiene información de muchas variables, el análisis estadístico de cada variable lleva intrínseco el supuesto de normalidad.

En este capítulo se hace una revisión de la literatura y se resumen las técnicas más extendidas para diagnosticar normalidad. El documento se divide en dos grandes secciones: *caso univariado* y *caso multivariado*. En cada caso se estudian procedimientos gráficos y las estadísticas para realizar la prueba formal del supuesto de normalidad. El aporte significativo al documento de Díaz y Morales (2012) es la revisión y recopilación de las diferentes opciones (funciones y librerías) que tiene R (*R Development Core Team* 2014) para la realización de diferentes gráficas y pruebas, todos los comandos se transcriben en el documento y pueden ser ejecutados por el lector.

3.3.1. Caso univariado

Sabemos que si un vector aleatorio es normal *p*-variables, entonces cada una de las variables aleatorias componentes es normal univariada. Por tanto, si por lo menos uno de los componentes de un vector aleatorio no es normal, podemos asegurar que este no es normal multivariado, de ahí la importancia de conocer las técnicas usadas para diagnosticar normalidad en el caso univariado.

3.3.1.1. Métodos gráficos

Para el diagnóstico de normalidad en el caso univariado se han desarrollado varias estrategias gráficas, que de manera visual alertan sobre la normalidad o no de los datos. En este capítulo se le hace más énfasis al manejo del software R.

Histogramas. Es una herramienta sencilla de implementar, porque todos los paquetes estadísticos tienen programas elaborados. Si los datos son normales, el histograma debería mostrar la bien conocida forma acampanada (Figura 17).

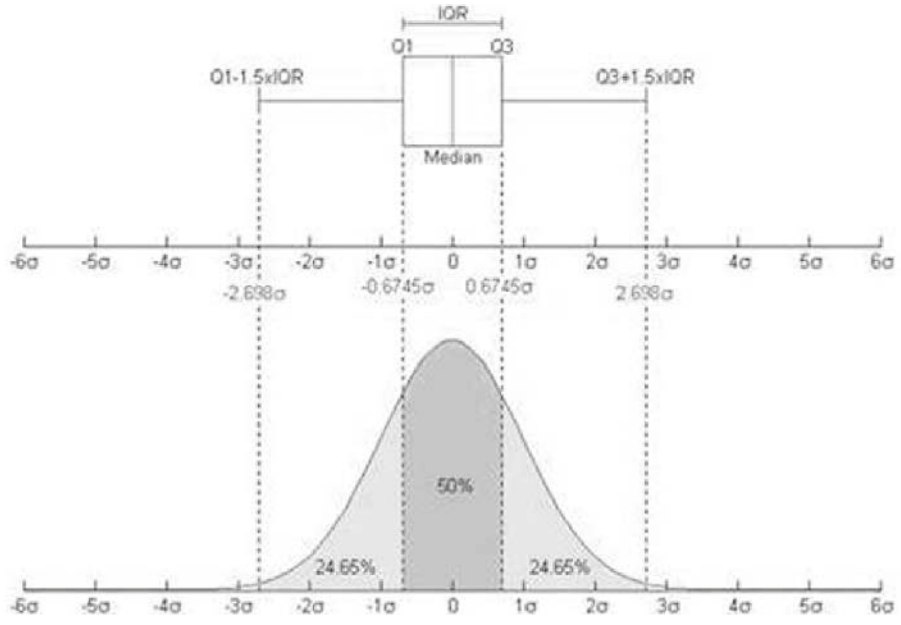


Figura 17. Gráfico de una distribución normal y su relación con un *boxplot*
Fuente: Elaboración propia

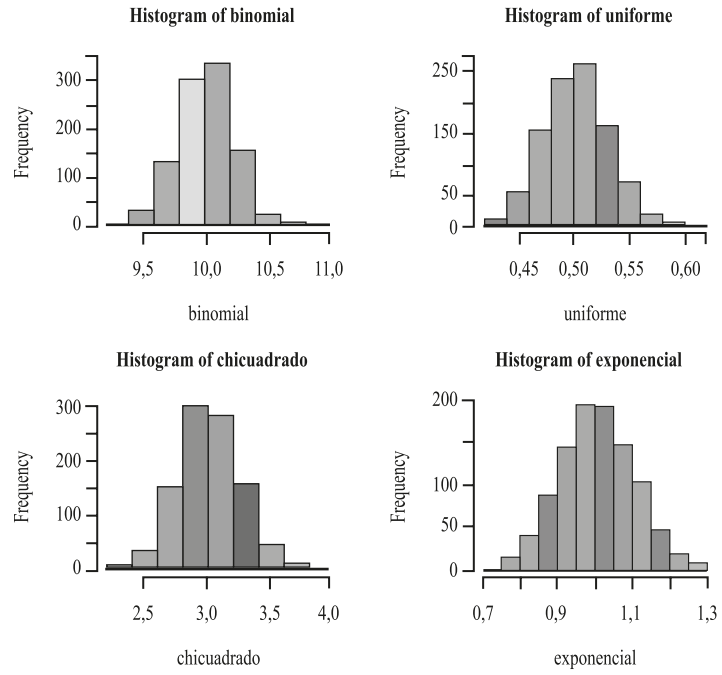


Figura 18. Histogramas a partir de datos generados de varias distribuciones
Fuente: Elaboración propia

En la Figura 18 se muestran cuatro histogramas elaborados a partir de datos simulados; ¿podrían señalar los que provienen de datos normales?

```
# programa para verificar el TLC con las siguientes distribuciones #
binomial(20,0.5), uniforme, chi- cuadrada y una exponencial
win.graph(); par(mfrow=c(2,2)); aux<-matrix(0,100,1000); par(m-
frow=c(2,2)); aux<-matrix(0,100,1000); muestras<- matrix(rbi-
nom(aux,20,0.5),100,1000); binomial<-apply(muestras,2,mean); hist(bi-
nomial,col=5:8); muestras<-matrix(runif(aux),100,1000)
uniforme<-apply(muestras,2,mean); hist(uniforme,col=3:7); mues-
tras<-matrix(rchisq(aux,3),100,1000); chicuadrado<-apply(muestras,2,-
mean); hist(chicuadrado,col=6:1);
muestras<-matrix(rexp(aux),100,1000); exponencial<-apply(muestras,2,-
mean); hist(exponencial,col=3:5); par(oma=c(1,1,1,1),new=T,font=2,-
cex=1)
```

Fuente: Adaptado de varios autores

El siguiente código genera el gráfico que se muestra en la figura 19 a partir de los datos simulados. Para unos datos en particular, solo tienen que introducir en el vector x de la forma $x < -c(x_1,x_2,...,x_n)$.

```
# Histogramas simulados (figura 19)
x<-rnorm(100,mean=10,sd=1) #introduzca sus datos en x
#el histograma
histo<-hist(x,prob=T,main="",ylim=range(hist(x)$density))
# La densidad
z<-pretty(histo$breaks,n=50)
y<-dnorm(z,mean=mean(x),sd=sd(x))
lines(z,y,lty=3,lwd=2,col="blue")
#para más detalles de la función hist
#digite en la consola ?hist
```

Fuente: Adaptado de varios autores

Gráfico Q-Q. Otro procedimiento gráfico para verificar normalidad univariada es el gráfico de probabilidad normal.

Este es un gráfico de los cuantiles empíricos contra los cuantiles teóricos (de ahí el nombre Q-Q plot) de la distribución normal estándar. Cuando los puntos en el gráfico de probabilidad normal quedan cerca de una línea recta, el supuesto de normalidad es razonable. El patrón de la desviación de los puntos de una línea recta indica la naturaleza de la separación de la normalidad tales como asimetría, apuntamiento, datos extremos o múltiples modas.

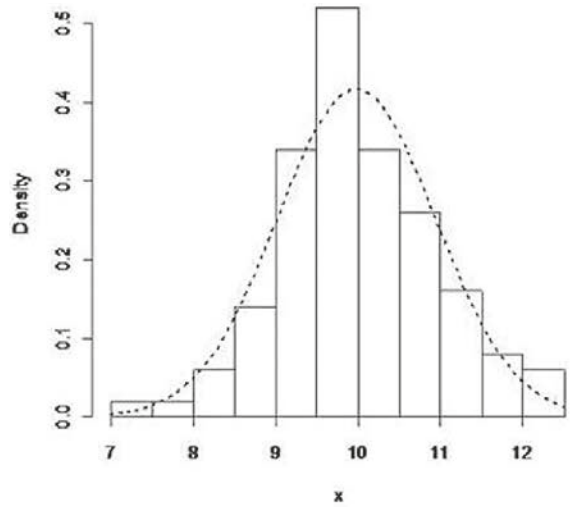


Figura 19. Histograma con la densidad normal superpuesta
Fuente: Adaptado de varios autores

La mayoría de los paquetes estadísticos facilitan la elaboración de este gráfico, sin embargo aquí se ilustrará paso a paso su construcción y luego se darán los comandos de R para generarlo.

1. Ordene las observaciones y denote los valores ordenados por $y_{(1)},y_{(2)},y_{(n)}$; de esta forma $y_{(1)} \leq y_{(2)} \leq ... \leq y_{(n)}$, entonces el punto $y_{(i)}$ es cuantil muestral *i/n*. A menudo la fracción *i/n* se cambia por $\left(i-\frac{1}{2}\right)/n$ como una corrección por continuidad, de esta forma $y_{(i)}$ se designa como el $\left(i-\frac{1}{2}\right)/n$ cuantil muestral.
2. Calcule los cuantiles poblacionales $q_1,q_2,...,q_n$, donde q_i es el valor para el cual la probabilidad de obtener una observación menor o igual que él es igual a $\left(i-\frac{1}{2}\right)/n$, es decir, q_i es tal que: $P(Z < q_i) = \frac{i-\frac{1}{2}}{n}$ con Z normal estándar.
3. Grafique los pares y examine la linealidad de los puntos.

Para ilustrar este procedimiento, y todos los de esta sección, usaremos los datos 1.38 1.40 1.42 1.54 1.30 1.55 1.50 1.60 1.41 1.34. En la Tabla 8 se muestran los resultados de los cálculos. Con el código de R que está a continuación se obtiene el gráfico cuantil-cuantil que se muestra en la Figura 20.

Tabla 8. Datos ordenados, cuantiles muestrales y cuantiles poblacionales

$y_{(i)}$	$\left(i-\frac{1}{2}\right)/10$	q_i
1.30	0.05	-1.645
1.34	0.15	-1.036
1.38	0.25	-0.674
1.40	0.35	-0.385
1.41	0.45	-0.126
1.42	0.55	0.126
1.50	0.65	0.385
1.54	0.75	0.674
1.55	0.85	1.036
1.60	0.95	1.645

Fuente: Adaptado de varios autores

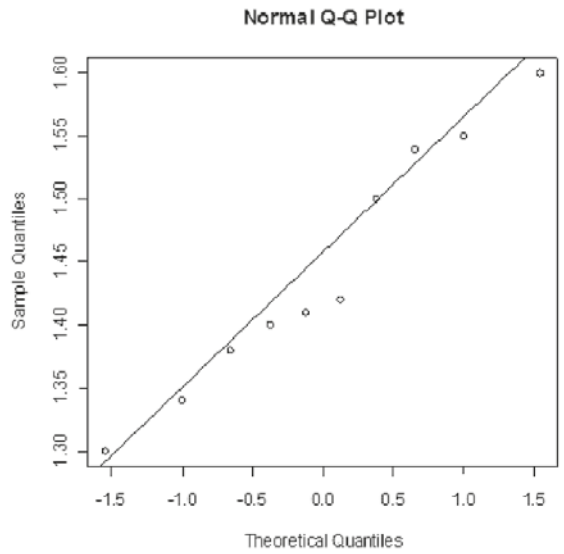


Figura 20. Gráfico cuantil-cuantil para verificar normalidad
Fuente: Adaptado de varios autores

3.3.2. Caso multivariado

La normalidad multivariada implica la normalidad de las distribuciones marginales, pero la normalidad de las marginales no garantiza que la distribución conjunta sea normal, a menos que las variables sean no correlacionadas, situación que es poco común. Así que las pruebas univariadas individualmente sirven para demostrar que no hay normalidad, cuando al menos una de ellas reporta no normalidad. Como en el caso univariado, se estudian *procedimientos gráficos* (gráfico del tipo *QxQ*, gráfico por pares) y *contrastes de multinormalidad* (Prueba basada en la distancia de Mahalanobis, Prueba de Mardia, Prueba de Shapiro-Wilk multivariada, prueba ε).

3.4. Contrastes de normalidad

En la literatura estadística se han propuesto varios procedimientos analíticos para probar la normalidad de datos univariados, aquí revisaremos la Prueba Ji-cuadrado de bondad de ajuste, la Prueba de Kolmogorov-Smirnov, la Prueba de Shapiro-Wilk y las pruebas basadas en asimetría y kurtosis.

3.4.1. Prueba Ji-cuadrado

Esta prueba es útil para probar el ajuste de un conjunto de datos a cualquier distribución. Se basa en el estadístico

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (2)$$

Donde O_i: frecuencias observadas en las k clases $[x_0,x_1),\ldots,[x_{k-1},x_k)$ y E_i son las frecuencias esperadas según el modelo probabilístico propuesto, para el caso normal se tiene $E_i = np_i$ con $p_i = P(x_{i-1} \leq X \leq x_i)$.

La estadística χ_0^2 se distribuye aproximadamente como una Ji-cuadrado con $k - r - 1$ grados de libertad, donde r es el número de parámetros que se estiman, en el caso de la normal $r = 2$ porque se estima la media y la varianza.

Con los siguientes comandos de R se realiza esta prueba, tenga en cuenta que hay que instalar la librería *nortest* (Gross, 2012).

```
y<-c(1.38,1.40,1.42,1.54,1.30,1.55,1.50,1.60,1.41,1.34);y
library(nortest)
pearson.test(y)
```

Fuente: Adaptado de varios autores

La salida se muestra a continuación; el p-valor indica que no hay suficiente evidencia para rechazar la hipótesis de normalidad. Tenga en cuenta, sin embargo, que esta prueba es poco potente con tamaños de muestra pequeños.

```
Pearson chi-square normality test
data: y
P = 3.2, p-value = 0.3618
```

Fuente: Adaptado de varios autores

3.4.2. Kolmogorov-Smirnov

Se asume que tenemos una muestra aleatoria $X_1, X_2, \ldots, X_n$ de alguna distribución continua con función de distribución acumulada $F(.)$.

Denotamos la función de distribución acumulada empírica por:

$$F_N(x) = \frac{1}{N}(número\ de\ obs\ \leq x) \qquad (3)$$

La prueba de *Kolmogorov-Smirnov* se utiliza para probar $H_0: F(x) = F_0(x)$ para todo contra $H_1: F(x) \neq F_0(x)$ para algún x, donde F_0 a una distribución $N(\mu,\sigma^2)$. El estadístico Kolmogorov-Smirnov es:

$$D_N = sup_x|F_N(x) - F_0(x)| \qquad (4)$$

Y es grande si los datos no son constantes con H_0. La distribución asintótica de D_N bajo H_0 cierta es

$$\lim_{x\to\infty} P\{\sqrt{N}D_N \leq x\} = Q(x)$$

con

$$Q(z) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp\{-2k^2z^2\}$$

Para cada $z > 0$. $Q(z)$ es la función de distribución acumulada de una distribución continua conocida como la *distribución Kolmogorov*. En general, los parámetros μ y σ^2 son desconocidos y se pueden reemplazar por su contraparte muestral.

Con el siguiente código de R se prueba, usando Kolmogorov-Smirnov, si los datos del ejemplo son normales con media $\mu = 1.4$ y $\sigma = 0.1$.

```
Prueba K-S
y<-c(1.38,1.40,1.42,1.54,1.30,1.55,1.50,1.60,1.41,1.34);y
ks.test(y,"pnorm",1.4,0.1)
```

Fuente: Adaptado de varios autores

La salida de la función *ks.test()* se muestra a continuación, el p-valor mayor que 0,05 indica que no hay evidencia para rechazar la hipótesis que los datos pertenecen a la distribución indicada.

```
 One-sample Kolmogorov-Smirnov test
data: y
D = 0.2413, p-value = 0.5285
alternative hypothesis: two-sided
```

Fuente: Adaptado de varios autores

3.4.3. Shapiro-Wilk

Esta prueba se basa en la comparación de los valores muestrales ordenados con su localización esperada bajo la hipótesis nula de normalidad. Aunque esta prueba es menos conocida es la que se recomienda para contrastar el ajuste de nuestros datos a una distribución normal, sobre todo cuando la muestra es pequeña ($n<30$). En esta prueba no es necesario calcular ni la media ni la varianza de la muestra para incluirlas en la hipótesis, pero requiere dos tipos

de tablas para su aplicación (Shapiro & Wilk, 1965). Los autores han proporcionado tablas para $n<50$.

Las pruebas de normalidad tienen como hipótesis:
H^o: La muestra proviene de una población con distribución Normal.
H^a: La muestra no proviene de una población con distribución Normal.

La prueba consta de los siguientes pasos:

1. Se ordenan los datos de menor a mayor: $y_1 \leq y_2 \leq ... \leq y_n$
2. Encontrar la Suma de Cuadrados Total: $SC_{Total} = \sum_{i,j} y_{ij}^2 - F.C$
3. Si n es par, $n = 2k$, y se calcula: $b = \sum a_{n-i+1}(y_{n-i+1} - y_i)$. Si n es impar, $n = 2k + 1$, se omite la mediana y se calcula b.

Si $n = 20$, $n = 20 = 2k$, resulta $k = 10$ y $b = a_{20}(y_{20} - y_1) + ... + a_{11}(y_{11} - y_{10})$.

Los coeficientes a_j se consiguen en tablas.

Se calcula la estadística:

$$W_{calc} = \frac{b^2}{SC_{total}} \qquad (5)$$

La regla de decisión dice: si $W < W_{\alpha,n}$ se rechaza H^o. Los valores pequeños de W indican falta de normalidad en la población.

Los comandos en R para realizar esta prueba son los siguientes:

```
y<-c(1.38,1.40,1.42,1.54,1.30,1.55,1.50,1.60,1.41,1.34);y
shapiro.test(y)
```

Fuente: Adaptado de varios autores

La salida de la función es la siguiente, como p-valor es mayor que 0.05 no rechazamos la hipótesis nula H_o*:Los datos son normales*.


```
 Shapiro-Wilk normality test
data: y
W = 0.9519, p-value = 0.6911
```

Fuente: Adaptado de varios autores

3.4.4. Pruebas de asimetría y kurtosis

Esta es una prueba clásica basada en las siguientes medidas de asimetría y kurtosis.

$$\sqrt{b_1} = \frac{\sqrt{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}{[\sum_{i=1}^{n}(y_i - \bar{y})^2]^{3/2}} \qquad (6)$$

$$b_2 = \frac{n\sum_{i=1}^{n}(y_i - \bar{y})^4}{[\sum_{i=1}^{n}(y_i - \bar{y})^2]^2} \qquad (7)$$

Estos son estimadores de los coeficientes de asimetría y kurtosis poblacionales $\sqrt{\beta_1}$ y β_2 respectivamente. Cuando la población es normal se tiene que $\sqrt{\beta_1} = 0$ y $\beta_2 = 3$.

La prueba de normalidad basada en la asimetría se lleva a cabo comparando $\sqrt{b_1}$ con valores tabulados o alternativamente, cuando $n \geq 8$, la función g definida por:

$$g(\sqrt{b_1}) = \delta sin^{-1}\frac{\sqrt{b_1}}{\lambda} \qquad (8)$$

Tiene aproximadamente una distribución normal estándar, donde $sin^{-1}(x) = ln(x + \sqrt{x^2 + 1})$ y los valores de λ y δ se obtienen de tablas.

Si los datos son normales, b_2 tiene, de manera asintótica, distribución $N(3,24/n)$ y por tanto tenemos una prueba de normalidad basada en la kurtosis: se rechaza la hipótesis de normalidad si

$$\frac{|b_2 - 3|}{\sqrt{24/n}} > Z_{\alpha/2}$$

Una prueba que usa simultáneamente la asimetría y la kurtosis se basa en la estadística

$$X^2 = \frac{nb_1}{6} + \frac{n(b_2 - 3)^2}{24} \qquad (9)$$

Bajo normalidad y asintóticamente, esta se distribuye como una Ji-Cuadrado con dos grados de libertad. Rechazamos la hipótesis de normalidad si $X^2 > \chi^2_2(\alpha)$.

Con el siguiente código de R se realizan las dos primeras pruebas, tenga en cuenta que la librería *moments* (Komsta & Novomestky, 2012), no pertenece al paquete básico de R así que hay que instalarlo manualmente.

```
y<-c(1.38,1.40,1.42,1.54,1.30,1.55,1.50,1.60,1.41,1.34);y
library(moments)
agostino.test(y)
anscobe.test(y)
```

Fuente: Adaptado de varios autores

La salida de los comandos se muestra a continuación (respectivamente)

```
 D'Agostino skewness test
data: y
skew = 0.1886, z = 0.2221, p-value = 0.8242
alternative hypothesis: data have a skewness
```

```
 Anscombe-Glynn kurtosis test
data: y
kurt = 1.8164, z = -0.9013, p-value = 0.3675
alternative hypothesis: kurtosis is not equal to 3
```

Fuente: Adaptado de varios autores

En ambos casos no se rechaza la hipótesis de normalidad.

Con el siguiente código en R se implementa la prueba ómnibus.


```
y<-c(1.38,1.40,1.42,1.54,1.30,1.55,1.50,1.60,1.41,1.34);y
library(moments)
b1<-skewness(y)^2;b1
b2<-kurtosis(y);b2
n<-length(y);n
X2<-((n*b1)/6)+(n*(b2-3)^2/24);X2
pvalor<-pchisq(X2,2,lower.tail=FALSE);pvalor
cat("X2=",X2,"p_valor=",pvalor,"\n")
```

Fuente: Adaptado de varios autores

Para los datos de nuestro ejemplo la prueba arroja:

X2= 0.6429844 p_valor= 0.7250663

Con esto se concluye que no hay evidencia para rechazar la hipótesis que los datos vienen de una distribución normal.

3.4.5. Posibles soluciones cuando se rechaza la hipótesis de normalidad

Si rechazamos o dudamos de la normalidad de nuestros datos, existen varias soluciones posibles:

- ✓ Si la distribución es más apuntada que la normal (mayor parte de los valores agrupados en torno de la media y colas más largas en los extremos), se debe investigar la presencia de heterogeneidad en los datos y de posibles valores atípicos o errores en los datos. La solución puede ser emplear pruebas no paramétricas.
- ✓ Si la distribución es *unimodal* y asimétrica, la solución más simple y efectiva suele ser utilizar una transformación para convertir los datos en normales.
- ✓ Cuando la distribución no es *unimodal* hay que investigar la presencia de heterogeneidad, ya que en estos casos la utilización de transformaciones no es adecuada y los métodos no paramétricos pueden también no serlo.

3.5. Transformaciones para conseguir datos normales

La utilización de transformaciones para lograr que los datos se ajusten a una distribución normal es en muchas ocasiones la solución más natural, ya que

existen gran cantidad de parámetros biológicos que tienen una distribución asimétrica, y que se convierten en aproximadamente simétricas al transformarlas mediante el **logaritmo**. Tenemos problemas con la transformación logarítmica ln(x) si la variable puede tomar el valor 0, por lo que en esos casos, o incluso si existen valores muy pequeños, será adecuado emplear la transformación ln(x+1). Cuando la desviación típica de los datos es proporcional a la media o cuando el efecto de los factores es multiplicativo, en lugar de aditivo, está indicado el uso de la transformación logarítmica. Otra transformación posible es $\sqrt{\ }$ que es aplicable cuando las varianzas son proporcionales a la media, lo que ocurre a menudo cuando los datos provienen de una distribución de Poisson (recuentos). Otra transformación habitualmente empleada es 1/x, que también precisa que sumemos una cantidad a cada valor si existen ceros.

4. ANÁLISIS EN COMPONENTES PRINCIPALES (ACP) PONDERADO

4.1. Introducción

El marco teórico general que permite definir métodos factoriales particulares es el análisis en componentes principales ponderado (Figura 21). Este análisis se presenta como diagrama de dualidad en Dray & Chessel (2003); y se encuentra como análisis factorial general en Escofier & Pagès (1988, 1998), Crivisqui (1993) y Fine (1996).

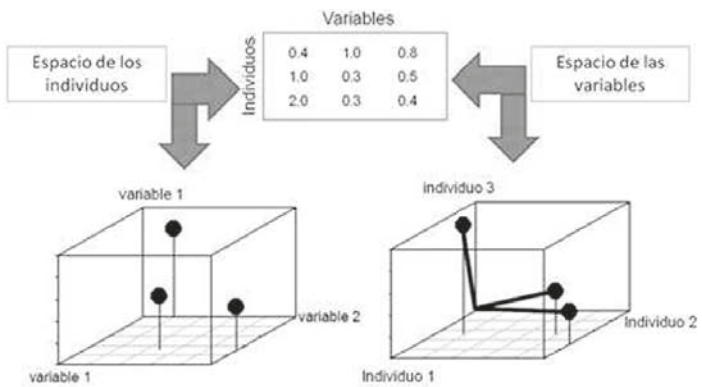


Figura 21. Representación geométrica de una tabla común a una nube de puntos en el espacio de los individuos y en el espacio de las variables
Fuente: Dray & Chessel (2003)

Se utiliza la notación ***ACP(X,M,D)*** para indicar el análisis en componentes principales con ponderaciones tanto en las filas ("individuos") como en las columnas (variables), donde: ***X*** es la matriz que contiene los datos a analizar (transformados). La reducción de la Matriz de Datos no modifica la evaluación de la relación entre dos variables cualesquiera de la tabla, y hace que la evaluación de la semejanza entre dos individuos cualesquiera de la tabla sea independiente de la escala de medida de las variables; ***M*** es la matriz diagonal de pesos de las columnas (variables), y ***D*** es la matriz diagonal de pesos de las filas ("individuos").

4.2. Objetivos del ACP ponderado

Los objetivos más importantes de todo análisis en componentes principales ponderado son:

- Generar nuevas variables que pueden expresar la información contenida en el conjunto original de datos.
- Reducir la dimensionalidad del problema que se está estudiando, como paso previo para futuros análisis.
- Eliminar, cuando sea posible, algunas de las variables originales si ellas aportan poca información.

4.3. Solución del ACP, mejor plano de proyección

Toda la información de interés para la comparación de los individuos está representada geométricamente en un espacio de muchas dimensiones ($\mathbf{R}^k$), describir esa representación, de alguna manera, es el objetivo del ACP. Para lograrlo hay que observar las proyecciones de la nube de individuos sobre ejes y planos. En primer lugar, se busca la dirección de un primer eje que conserve mejor las distancias originales de los puntos, que es igual a conservar las distancias de los puntos al centro de gravedad.

Este es un problema de optimización que se resuelve por el método de mínimos cuadrados (Rao, 1964) y que da como resultado que la dirección de ese eje sea la misma del vector propio asociado al valor propio más grande de la matriz **X'X**. Luego, se busca un segundo eje perpendicular al anterior que

recoja lo más posible la dispersión remanente y así sucesivamente. La solución es diagonalizar la matriz **X'X**, y obtener los valores propios ordenados de mayor a menor y los vectores propios asociados. Geométricamente, este resultado corresponde a una rotación del sistema de ejes (p.103). El primer eje tiene la dirección más alargada de la nube o sea la de mayor dispersión (corresponde a la dirección de mayor inercia).

Podemos leer en un documento solo dos dimensiones y por eso necesitamos el mejor plano para aproximarnos a la lectura. El plano conformado por los dos primeros nuevos ejes, denominados factoriales, es la mejor fotografía de la nube de puntos (Figura 22). Poner el sistema de referencia en el centro de gravedad, conlleva a concentrarse en la *dispersión* (la forma) de la nube de puntos.

Figura 22. Fotografías alusivas a un plano factorial
Fuente: Adaptado de varios autores

La nube de puntos posee una inercia que se puede descomponer en cada uno de los ejes de la representación, en planos y en subespacios de menor dimensión. Cada eje contribuye a la inercia con la varianza de la variable que representa, es decir corresponde a la contribución de la variable a la inercia total (Cabarcas & Pardo, 2001).

4.4. Fórmulas del *ACP(X,M,D)*

Un método de análisis en componentes principales ponderado específico queda completamente determinado definiendo las matrices: ***X***, que se obtiene de

los datos mediante la transformación adecuada, es un producto escalar de ***M***, métrica en el espacio de las filas y pesos en el espacio de las columnas, y ***D*** es un producto escalar de R^I, pesos en el espacio de las filas y métrica en el espacio de las columnas.

Las principales fórmulas del ***ACP(X,M,D)*** se resumen en la Tabla 9, de donde se pueden derivar las de un método particular una vez se han establecido las tres matrices ***(X,M,D)***.

Tabla 9. Fórmulas del ***ACP(X,M,D)***

Espacio	R^K	R^I
Nube	N_I	N_K
Coordenadas	Filas de **X**	Columnas de **D**
Pesos	Diagonal de **D**	Diagonal de **M**
Métrica	**M**	**D**
Inercia	Traza **(X´DXM)**	Traza **(XMX´D)**
Valor propio	λ_s	λ_s
Vector propio	$\boldsymbol{u}_S$	$\boldsymbol{v}_S$
Coordenadas factoriales	$F_S = \boldsymbol{XMu}_s = \sqrt{\lambda_s} V_s$	$Gs = \boldsymbol{X'Dv}_s = \sqrt{\lambda_s}\boldsymbol{u}_s$
Fórmulas de transición	$F_S = \frac{\mathbf{1}}{\sqrt{\lambda_s}} \boldsymbol{XMG_s}$ $F_s = (i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k=1}^{K} x_{ik} m_k G_s(k)$	$G_s = \frac{\mathbf{1}}{\sqrt{\lambda_s}} \boldsymbol{X'D} F_s$ $G_s = (k) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} x_{ik} d_i F_s(i)$

Fuente: Escofier & Pagès (1988-1998)

El análisis en componentes principales ponderado (Dray & Chessel, 2003; Pardo *et al.*, 2012) asociado a la tripleta **(X,M,D)** consiste en determinar valores y vectores propios asociados a las diferentes posibilidades de los productos matriciales **X'MXD**, **DXMX'**, **XMX'D** y **MX'DX**. En la práctica, se realiza una única diagonalización y a través de fórmulas de transición se encuentran los componentes, co-factores y ejes principales.

4.5. Ayudas para la interpretación de las gráficas factoriales

Se definen indicadores numéricos que son ayudas adicionales (Vertel & Par-

do, 2010) a las del ACP (contribuciones a la inercia de los ejes, calidades de representación individuos y variables, distancias al cuadrado).

Antes de interpretar los resultados obtenidos del ACP ponderado, se debe definir cada uno de los ejes factoriales. Para ello, es importante conocer cuáles modalidades de las variables en estudio han contribuido en la elaboración de cada uno de los ejes, es decir, el peso (contribución absoluta) que tiene cada modalidad en la definición de cada uno de los ejes. La suma de todas las contribuciones absolutas tanto para las frecuencias activas como para los individuos en cada eje factorial será igual a 100 %.

Definidas las contribuciones absolutas, se calculan las contribuciones relativas (cosenos al cuadrado), estas proveen información de cuánto de la inercia de una modalidad está explicada por el eje factorial. La suma de todos los cosenos al cuadrado será igual a 1. Es importante entender las diferencias entre las dos clases de contribuciones: La contribución absoluta de las modalidades al eje sirve primeramente como una guía para su interpretación, mientras que las contribuciones relativas indican qué tan bien una modalidad es descrita por el eje. Usualmente, una alta contribución de los puntos a las dimensiones implica también una alta contribución relativa. Debido a que ambos valores son siempre positivos, ayuda revisar las coordenadas 1 y observar en qué dirección del eje se encuentra cada una de las modalidades del estudio.

4.5.1. Calidad de la representación

Un indicador de esa calidad es el coseno que es una relación entre las magnitudes de la proyección y del vector original. Sin embargo, se utiliza coseno cuadrado, ya que para un punto, la suma de los cosenos cuadrados sobre todos los ejes factoriales es 1 y su coseno cuadrado en un subespacio se obtiene sumando sus cosenos cuadrados sobre los ejes factoriales que lo generan. Sobre un eje *s* el coseno cuadrado de un punto-fila **i** es:

$$\boldsymbol{cos_s^2(i)} = \frac{\boldsymbol{F_s^2(i)}}{\boldsymbol{\|i\|^2}} \qquad (10)$$

Las coordenadas de un vector fila **i** están en la fila **i** de la matriz. El valor del coseno al cuadrado coincide con la relación de contribuciones del individuo **i** a la inercia: *contribución a la inercia proyectada sobre el eje s/ contribución a la inercia* total y se llama también contribución relativa. La calidad de representación ayuda a evitar lecturas erróneas de puntos mal representados.

4.5.2. Contribución absoluta

Otro aspecto que ayuda a la interpretación es identificar las filas que más contribuyen a la inercia del *eje s*.

Este indicador se obtiene dividiendo la inercia proyectada del punto-***fila i*** sobre la inercia total del eje, que es igual al valor propio. Se suele expresar en porcentaje y recibe el nombre de contribución absoluta de la fila **i** al eje *s*:

$$\boldsymbol{con_s^2(i)} = \frac{\boldsymbol{p_i F_s^2(i)}}{\boldsymbol{\lambda_s}} \qquad (11)$$

4.5.3. Distancias al cuadrado

El parecido o diferencia de los individuos según las variables que los están caracterizando se traduce en la analogía geométrica en un sistema de distancias.

El ACP utiliza la distancia euclidiana canónica, es decir la que aprendimos desde la geometría elemental: la distancia al cuadrado entre dos puntos es la suma de las diferencias al cuadrado entre las coordenadas, o sea:

$$\boldsymbol{d^2(i,l)} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{(x_{ij} - x_{lj})^2} \qquad (12)$$

El parecido o diferencia de los individuos se traduce en cercanía o lejanía geométrica y entonces la lectura de las gráficas permite observar lo más importante de la tabla de datos.

Cada individuo contribuye con el cuadrado de su distancia al centro de gra-

vedad dividida por *n* (total de individuos). *Los individuos más alejados del centro* tienen más importancia en el análisis.

4.6. ACP ponderados particulares para tablas de datos X

Existen diferentes posibilidades de análisis en componentes principales ponderados ***ACP(X, M, D)*** para analizar tablas de datos (Noy-Meir, 1973; Noy-Meir *et al.,* 1975). Consideremos una tabla ***X*** conteniendo las cantidades de ***p*** variables (columnas) en ***n*** condiciones (filas).

A manera de ejemplo, en cada caso se procede a la elaboración del gráfico del plano factorial conformado por la combinación de los ejes 1 y 2 (previamente descrito), el cual representa tanto las filas-individuos como las columnas-variables. En su construcción se toman en cuenta las contribuciones tanto absolutas como relativas de las categorías que contribuyen significativamente a su formación.

Todas las fórmulas se pueden derivar del análisis en componentes principales ponderado ***ACP(X, M, D)*** (ver Tabla 9).

4.6.1. Análisis en Componentes Principales (ACP) de datos originales

$\boldsymbol{X} = [y_{ij}]$ $\qquad \boldsymbol{M} = \boldsymbol{I}_p$ $\qquad \boldsymbol{D} = \frac{1}{n}\boldsymbol{I}_n$

(Fine, 1996; Fernández, 2002; Vertel & Pardo, 2010)

4.6.2. Análisis en Componentes Principales (ACP) centrado por columnas

$\boldsymbol{X} = [y_{ij} - \overline{y}_j]$ $\qquad \boldsymbol{M} = \boldsymbol{I}_p$ $\qquad \boldsymbol{D} = \frac{1}{n}\boldsymbol{I}_n$

Con: $\overline{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$ (media ponderada por columnas)

4.6.3. Análisis en Componentes Principales (ACP) centrado por filas

$\boldsymbol{X} = [y_{ij} - \overline{y}_i]$ $\qquad \boldsymbol{M} = \boldsymbol{I}_p$ $\qquad \boldsymbol{D} = \frac{1}{n}\boldsymbol{I}_n$

Con: $\overline{y}_i = \frac{1}{p}\sum_{j=1}^{p} y_{ij}$ (media ponderada por filas)

4.6.4. Análisis en Componentes Principales (ACP) normado-centrado por columnas

$\boldsymbol{X} = \left[\frac{y_{ij} - \overline{y}_i}{\sigma_j}\right]$ $\qquad \boldsymbol{M} = \boldsymbol{I}_p$ $\qquad \boldsymbol{D} = \frac{1}{n}\boldsymbol{I}_n$

Con: $\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \overline{y}_i)}$ (desviación estándar)

4.6.5. Análisis de Correspondencias Simples (ACS) de la tabla de frecuencias T

La tabla de datos ***X*** del ***ACP(X, M, D)*** tiene término general $p_{ij} = \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}}$, donde: $\boldsymbol{D} = \boldsymbol{D_I} = diag(f_i)$ y $\boldsymbol{M} = \boldsymbol{D_J} = diag(f_j)$ (Greenacre, 2007; Vertel & Pardo, 2010). El ACS de la tabla ***T*** es el $\boldsymbol{ACP(P, D_J, D_I)}$.

4.6.6. Análisis de Correspondencias Múltiples (ACM) de la tabla [Q]

Para una tabla de datos **[*Q*]** formada por variables cualitativas (Tenenhaus &Young, 1985; Chessel, 1992, p.32).

Notamos ***Y***: tabla disyuntiva completa y $\boldsymbol{D_m}$: la matriz diagonal de frecuencias de las modalidades ***m***. El término general de la tabla ***Y*** es y_{im} (y_{im} es igual a 1 si la categoría ***m*** es observada en una fila ***i***).

El ACM de la tabla **[*Q*]** es el $\boldsymbol{ACP(YD_m^{-1} - 1_{nm}, \frac{1}{p}D_m, \frac{1}{n}I_n)}$. Donde: $\boldsymbol{1_{nm}}$ es la matriz de ***n*** filas, y ***m*** (modalidades) columnas donde los términos valen 1, ***p*** es el número de variables.

$\boldsymbol{X} = \left[\frac{y_{im}n}{n_m} - 1\right]$ $\qquad \boldsymbol{M} = \frac{1}{p}\boldsymbol{D_m}$ $\qquad \boldsymbol{D} = \frac{1}{n}\boldsymbol{I}_n$

Para detenerse en la práctica del ACP, entender en un lenguaje sencillo cómo se consigue el plano factorial de individuos y variables, como también las ayudas a la interpretación se muestra con una aplicación de dos variables y ocho individuos en el Capítulo 5 (p.104) de este documento, donde se explica paso a paso cada uno de los términos básicos del ACP. Para esto se utilizan conocimientos básicos de geometría elemental y álgebra lineal.

Las técnicas de análisis de datos multivariadas básicas (análisis en componentes principales, análisis de correspondencias simples y análisis de correspondencias múltiples) serán objeto de estudio en la parte III de este libro, implementadas sobre un ACP ponderado.

Se mostrará con aplicaciones del área pecuaria la lógica, utilización e implementación de cada una de ellas. Los resultados numéricos y gráficos serán generados en el *software* estadístico R (*Development Core Team* 2014) y los paquetes *ade4* (Chessel *et al.*, 2004; Dray & Dufour, 2007) *y FactoClass* (Pardo & Del Campo, 2007).

4.7. ACP ponderados particulares para relacionar dos tablas de datos (X, Y)

Hay tres estrategias principales para la combinación de dos o más tablas de datos, se superponen, debido a los muchos conjuntos de parámetros, y se da una amplia gama de prácticas. Solo hay que arrastrar estos tres principios para tomar una decisión o para construir asociaciones originales que puedan ser necesarias; se hace énfasis para el acoplamiento de dos tablas (Figura 23).

El acoplamiento de dos matrices de datos es una operación fundamental en ciencias pecuarias. Hay una gran cantidad de literatura sobre el tema. Entre las bases históricas deben recordar algunas operaciones básicas.

Se busca identificar y cuantificar la asociación entre dos o más grupos de variables, utilizando la teoría del ACP ponderado ***ACP(X,M,D)*** (Chessel, 1992).

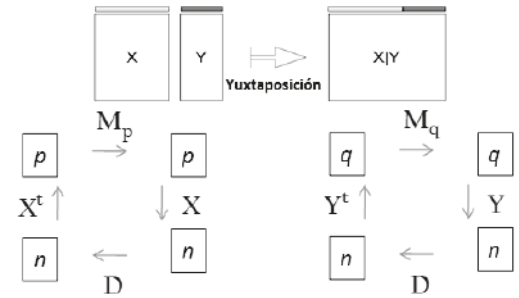


Figura 23. Yuxtaposición de dos tablas de datos
Fuente: Chessel (1992)

Al analizar dos tablas con información de parámetros del área pecuaria, ambas deben tener las mismas filas, simplemente se juntan y forman una nueva tabla que apela a un análisis más simple (Figura 24).

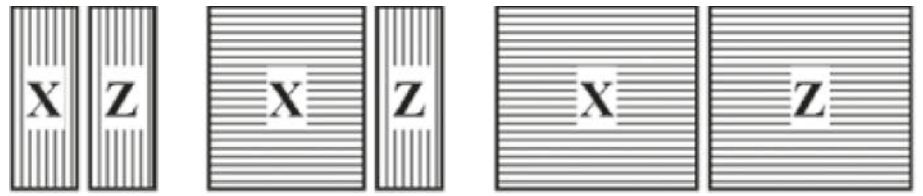


Figura 24. Marco de utilización del análisis canónico de las correlaciones, análisis sobre variables instrumentales y el análisis de co-inercia
Fuente: Adaptado de varios autores

Todas las fórmulas se pueden derivar de las correspondientes del ***ACP(X,M,D)*** (ver Tabla 9).

Presentamos algunos ejemplos a continuación, los cuales serán explicados con mayor detalle en el libro nivel avanzado.

4.7.1. El Análisis en Componentes Principales con Variables Instrumentales (ACPVI), de la tabla de datos [X Y] es el ACP(Y'DX,M,R)

Donde: $\boldsymbol{D = \frac{1}{n} I_n}$, $\mathbf{M = I_p}$, $\mathbf{R = (Y'DY)^{-1}}$. La tripleta anterior, se puede escribir ***ACP(P_yX, M, D)*** con: $\boldsymbol{P_y = Y_o (Y'_o DY_o)^{-1} Y'_o D}$. (ACPVI, Rao, 1964; Lebreton *et al*., 1991).

4.7.2. El Análisis Canónico de Correspondencias (ACC) de la tabla [T Z], ACP($\hat{Y}$, D_J, D_I)

(ACC, Ter-Braak, 1986), también llamado análisis factorial de correspondencias de variables instrumentales (AFCVI, Chessel *et al.*, 1987; Lebreton *et al.,* 1988; Faye *et al*., 1997; Vertel & Pardo, 2010).

4.7.3. El Análisis Factorial Múltiple (AFM) de la tabla [T Z], ACP $([P\ Z_O], \frac{1}{\lambda_1}(I_k, D_j), D_I)$

(AFM, Abdessemed & Escofier, 1992; Escofier & Pagès, 1988-1998; Vertel & Pardo, 2010).

4.7.4. El Análisis Canónico de Correlación (ANCOR) de la tabla de datos [X Y] es el ACP(Y'DX,(X^t DX)$^{-1}$, (Y^t DY)$^{-1}$) (ANCOR, Hotteling, 1933)

El análisis de COINERCIA analiza la tripleta (Y^tDX, Q, R), (COINERCIA, Dolédec & Chessel, 1994).

5. UTILIZACIÓN DEL ÁLGEBRA Y LA GEOMETRÍA EN ACP ($\mathbb{X}$, $\mathbb{M}$, $\mathbb{D}$)

5.1. Los datos

Matriz de datos: *n* individuos *p* variables reales

$$\mathbb{X} = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & & & & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} & \dots & x_{np} \end{pmatrix} \rightarrow w_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \epsilon \quad \mathbb{R}^p \qquad i - ésino\ indiciduo, i = 1, \dots, n$$

Se representa como puntos en $\mathbb{R}^p$ $(J = p)$, imagen que se denomina nube de filas.

$$x_j = \begin{pmatrix} x_{ij} \\ \vdots \\ x_{nj} \end{pmatrix} \epsilon \quad \mathbb{R}^p \qquad j - ésino\ indiciduo,\ \ j = 1, \dots, p$$

Las columnas de **X** representan a las variables, cada una se puede ver como un vector en $\mathbb{R}^n$. Los vectores constituyen la nube de columnas.

Las filas representan "individuos" y las columnas, "variables continuas", entonces los vectores fila representados en $\mathbb{R}^J$ constituyen la "nube de individuos" y los vectores columnas en $\mathbb{R}^I$, la nube de variables.

Sea $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ la media de x_j

El punto medio o "centro de gravedad" de la nube de los *n* individuos tiene dos formas de mirar:

$$g = \frac{1}{I}\sum_{i=1}^{n} y_i = \frac{1}{I}\mathbb{Y}'1_I, 1_I\ vector\ de\ I\ unos\ (Pardo, 2009)$$

$$g = \frac{1}{n}\sum_{i=1}^{n} W_i = \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_j \\ \vdots \\ \bar{X}_p \end{pmatrix} \qquad (Fine, 1996)$$

Matriz de datos

$$\mathbb{Y} = \begin{bmatrix} 18 & 9 \\ 13 & 9 \\ 15 & 5 \\ 9 & 7 \\ 11 & 3 \\ 5 & 5 \\ 7 & 1 \\ 2 & 1 \end{bmatrix} \qquad \begin{cases} W_1 = \begin{pmatrix} 18 \\ 9 \end{pmatrix} & W_2 = \begin{pmatrix} 13 \\ 9 \end{pmatrix} & W_3 = \begin{pmatrix} 15 \\ 5 \end{pmatrix} & W_4 = \begin{pmatrix} 9 \\ 7 \end{pmatrix} \\ W_5 = \begin{pmatrix} 11 \\ 3 \end{pmatrix} & W_6 = \begin{pmatrix} 5 \\ 5 \end{pmatrix} & W_7 = \begin{pmatrix} 7 \\ 1 \end{pmatrix} & W_8 = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \end{cases}$$

8 individuos dentro de $\mathbb{R}^2$

Centro de gravedad

2 variables de $\mathbb{R}^8$

$$x_1 = \begin{pmatrix} 18 \\ 13 \\ 15 \\ 9 \\ 11 \\ 5 \\ 7 \\ 2 \end{pmatrix} \qquad x_2 \begin{pmatrix} 9 \\ 9 \\ 5 \\ 7 \\ 3 \\ 5 \\ 1 \\ 1 \end{pmatrix}$$

Punto medio:

$$\bar{g} = \frac{1}{8}\sum_{i=1}^{g} W_i = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 10 \\ 5 \end{pmatrix} \qquad \bar{x}_1 = 10 \qquad \bar{x}_2 = 5$$

$$\bar{g} = \frac{1}{I}\mathbb{Y}'1_I = \frac{1}{8}\begin{bmatrix} 18 & 13 & 15 & 9 & 11 & 5 & 7 & 2 \\ 9 & 9 & 5 & 7 & 3 & 5 & 1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \frac{1}{8}\begin{bmatrix} 18+13+\cdots+2 \\ 9+9+\ldots+1 \end{bmatrix} = \begin{pmatrix} 10 \\ 5 \end{pmatrix}$$

Matriz de los datos centrados

Sea $Y = (y_{ij})$ $\qquad y_c = y_{ij} - y_j$

Al restar a cada vector individuo el centro de gravedad $\bar{g}$ se obtiene la matriz centrada $\mathbb{Y}_c$:

$$\mathbb{Y}_c = \begin{bmatrix} 18 & 9 \\ 13 & 9 \\ 15 & 5 \\ 9 & 7 \\ 11 & 3 \\ 5 & 5 \\ 7 & 1 \\ 2 & 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}[10 \qquad 5] = \begin{bmatrix} 18-10 & 9-5 \\ \vdots & \vdots \\ 9-10 & 7-5 \\ \vdots & \vdots \\ 2-10 & 1-10 \end{bmatrix} = \begin{bmatrix} 8 & 4 \\ 3 & 4 \\ 5 & 0 \\ -1 & 2 \\ 1 & -2 \\ -5 & 0 \\ -3 & -4 \\ -8 & -4 \end{bmatrix}$$

Matriz de las covarianzas

Sea

$$V(x_j) = \frac{1}{n}\sum_{i=1}^{nn} (Y_{ij})^2 \qquad (varianza\ de\ x_j) \quad y$$

$$cov(x_j, x_k) = \frac{1}{n}\sum_{i=1}^{nn} Y_{ij}Y_{ik} \qquad (covarianza\ de\ x_j\ y\ x_k$$

$$V = \begin{pmatrix} V(x_1) & & cov(x_j, x_k) \\ & V(x_i) & \\ cov(x_k, x_j) & & \end{pmatrix}$$

La traza de **V** es la suma de los elementos diagonales (suma de las varianzas de las variables:

$$I = t_r(V) = \sum_{j=1}^{p} V(x_j) = \sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n} y_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} y_{ij}^2 \qquad (13)$$

El concepto de inercia es central tanto a los métodos multivariados de datos como de clasificación; es la generalización del concepto de varianza al caso de varias variables.

Poner el sistema de referencia en el centro de gravedad conlleva olvidarse de la porción de la nube de puntos y concentrarse en su dispersión, es decir, su forma. La nube de puntos posee una inercia que se puede descompensar en cada uno de los ejes de la representación en planos y en subespacio de mayor dimensión. Cada eje contribuye a la inercia con la varianza de la variable que representa, es decir corresponde a la contribución de la variable a la inercia total. Se ve que la variable y_1 contribuye más a la inercia (73,33 %) y la variable y_2 es la que menos contribuye (26,67 %). Con una mayor dispersión de la nube en el eje horizontal ["1(y_1)"] (Tabla 10).

Tabla 10. Contribución de las variables a la inercia total

Variable	y_1	y_2	Total
Varianza	24,75	9,00	33,75
% Inercia	73,33	26,67	100,00

Fuente: Elaboración propia

Nota:

Las *n* filas de $\mathbb{X}$ son los vectores de $\mathbb{R}^P$, mientras que las *p* columnas de $\mathbb{X}$ son los vectores de $\mathbb{R}^n$.

$\mathbb{M}$ es la matriz de un producto escalas de $\mathbb{R}^P$, es una matriz cuadrática simétrica que define la función:*

$\mathbb{D}$ es la matriz de un producto escalas de $\mathbb{R}^n$, es una matriz cuadrática simétrica que define la función:*

$$*: (x,y) = \langle \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}, \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \rangle \in \mathbb{R}^p x \mathbb{R}^p \rightarrow X' \mathbb{M} \mathbb{Y} = \langle x|y \rangle_{\mathbb{M}} \in \mathbb{R}$$

$$(x,y) = \langle \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \rangle \in \mathbb{R}^n x \mathbb{R}^n \rightarrow X' \mathbb{D} \mathbb{Y} = \langle x|y \rangle_{\mathbb{D}} \in \mathbb{R}$$

5.2. Representación de los *n* individuos en ($\mathbb{R}^p$, I_p)

$\mathbb{R}^P$ es dotado de la métrica euclidiana clásica. La base canónica ($e_1, e_2 ..., e_p$) es una base ortogonal. Al eje engendrado por e_j corresponde la variable x_j. Representación de los individuos (Figura 28).

Representación de los individuos 'centrados'

El centrado en $\mathbb{R}^P$ es equivalente al cambio del origen del referencial posicionándolo en *g* (Figura 25). Al eje engendrado por e_j corresponde ahora la variable centrada y_j.

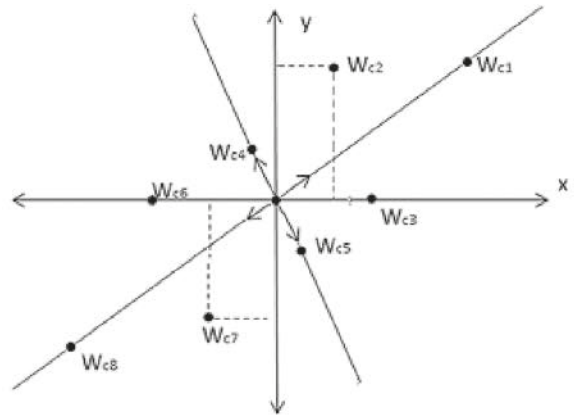


Figura 25. Representación de los individuos
Fuente: Adaptación de varios autores

5.3. Inercia y contribución a la inercia

La inercia de la nube de los individuos de $\mathbb{R}^P$ con respecto al centro de gravedad es definido por:

$$I = \frac{1}{n} \sum_{i=1}^{n} \|W_{ci}\|^2 \qquad (14)$$

(Media de los cuadrados de las distancias de los puntos individuos al centro de gravedad)

Puesto que tenemos $\|\mathbb{W}_{ci}\|^2 = \sum_{j=1}^{P} (y_{ij})^2$, decidimos:

$$I = \sum_{j=1}^{P} V(x_j) = t_r(V) = \frac{1}{8} \sum_{i=1}^{8} \|W_{ci}\|^2 \qquad (15)$$

Inercias:

$\|W_{c1}\|^2 = 8^2 + 4^2 = 64 + 16 = 80$; $\|W_{c2}\|^2 = 3 + 4^2 = 9 + 16 = 25$
$\|W_{c3}\|^2 = 5^2 + 0^2 = 25 + 0 = 25$; $\|W_{c4}\|^2 = (-1)^2 + 2^2 = 1 + 4 = 5$
$\|W_{c5}\|^2 = (1)^2 + (-2)^2 = 1 + 4 = 5$; $\|W_{c6}\|^2 = (-5)^2 + 0^2 = 25 + 0 = 25$
$\|W_{c7}\|^2 = (-3)^2 + (-4)^2 = 9 + 16 = 25$; $\|W_{c8}\|^2 = (-8)^2 + (-4)^2 = 64 + 16 = 80$

$$I = \frac{1}{8}(80 + 25 + 25 + 5 + 5 + 25 + 25 + 80) = \frac{270}{8} = 33.75$$

$$I = V(x_1) + V(x_2) = 33.75 = 24.75 + 9.0$$

$$V = \begin{pmatrix} 24.75 & 10.5 \\ 10.5 & 9.0 \end{pmatrix} \qquad V(x_1) = \frac{\sum_{j=1}^{n} x_{ij}}{n}$$

La contribución de los individuos i_o a la inercia I es: $\frac{\frac{1}{n}\|W_{ci_0}\|^2}{inercia}$

La contribución a la inercia de los individuos 1, 2, 3, 4, 5, 6, 7, 8 es respectivamente:

$$con_{\propto}(1) = \frac{\frac{1}{8}\|W_{ci}\|^2}{inercia} = \frac{\frac{1}{8}(80)}{33.75} = \frac{10}{33.75} = 0.2962 = 29.62\% = con_{\propto}(8)$$

$$con_{\propto}(2) = \frac{\frac{1}{8}\|W_{c1}\|^2}{inercia} = \frac{\frac{1}{8}(25)}{33.75} = 0.0925 = 9.25\% = con_{\propto}(3) = con_{\propto}(6) = con_{\propto}(7)$$

$$con_{\propto}(4) = con_{\propto}(5) = \frac{\frac{1}{8}(5)}{33.75} = 0.0185 = 1.85\%$$

Cada individuo contribuye con el cuadrado de su distancia al centro de gravedad dividiendo por total de individuos. Los individuos más alejados del punto tienen más importancia en el análisis.

Tabla 11. Contribución de los individuos a la inercia

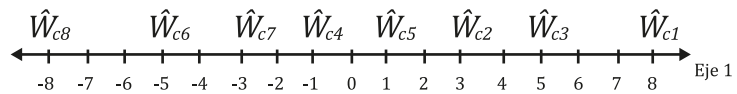
identificación	1	2	3	4	5	6	7	8	Total
Inercia	10	3.125	3.125	0.625	0.625	3.125	3.125	10	33.75
% inercia	0.2962	0.0925	0.0925	0.0185	0.0185	0.0925	0.0925	0.2962	1.00

$$\frac{d_i^2}{n} = con$$

Fuente: Fine (1996)

✓ Si proyectamos los puntos sobre el primer eje, obtenemos:

$$\begin{matrix} i & x_1 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{matrix}\begin{bmatrix} 8 \\ 3 \\ 5 \\ -1 \\ 1 \\ -5 \\ -3 \\ -8 \end{bmatrix}$$



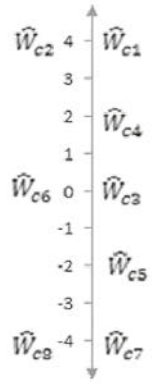
La inercia de la nube proyectada es:

$$\hat{I} = \frac{1}{8}[(8)^2 + (3)^2 + (5)^2 + (-1)^2 + (1)^2 + (-5)^2 + (-3)^2 + (-8)^2]$$

$$\hat{I} = \frac{1}{8}(64 + 9 + 25 + 1 + 1 + 25 + 9 + 64) = 24.75$$

✓ Si proyectamos los puntos sobre el segundo eje, observamos:

$$\begin{matrix} i & x_1 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{matrix}\begin{bmatrix} 4 \\ 4 \\ 0 \\ 2 \\ -2 \\ 0 \\ -4 \\ -4 \end{bmatrix}$$



La inercia de la nube proyectada es:

$$\hat{I} = \frac{1}{8}[(4)^2 + (4)^2 + (0)^2 + (2)^2 + (-2)^2 + (0)^2 + (-4)^2 + (-4)^2] = \frac{72}{8} = 9 = V(x_2)$$

5.4. El objetivo del ACP y solución

• El problema

El objetivo del ACP es la búsqueda de un subespacio de pequeña dimensión *s(s=1,2)* "lo más cercano posible" de la nube de puntos (Fine, 1996). Se necesita definir una distancia entre la nube de puntos *N* y un subespacio E_s.

Siendo $\widehat{W}_{ci}$ la proyección de W_{ci} sobre E_s, podemos proponer la definición:

$$d^2(N,E_s) = \frac{1}{n}\sum_{i=1}^{n} d^2(W_{ci},\widehat{W}_{ci}) = \frac{1}{n}\sum_{i=1}^{n} d^2 \|W_{ci}-\widehat{W}_{ci}\|^2 \qquad (16)$$

Pero, el teorema de Pitágoras permite escribir:

$$\|W_{ci}\|^2 = \|\widehat{W}_{ci}\|^2 + \|\widehat{W}_{ci}-W_{ci}\|^2 \qquad (17)$$

Pues, si $\hat{I}_s$ es la inercia de la nube proyectada sobre E_s, obtenemos:

$$\hat{I}_s = \frac{1}{n}\sum_{i=1}^{n}\|\widehat{W}_{ci}\|^2 \; y \qquad d^2(N,E_s) = I - \hat{I}_s \qquad (18)$$

Como *I* es la inercia de la nube en el espacio $\mathbb{R}^P$, *I* es una cantidad fija; finalmente es equivalente buscar un subespacio E_s que minimice la distancia de la nube a E_s a buscar un subespacio E_s que maximice la inercia de la nube proyectada sobre este subespacio.

Se puede demostrar que el subespacio de dimensión *S*, solución del problema, es incluido en el subespacio de dimensión (*s* + 1), solución del problema.

• Diagonalización de V

La solución utilizada la "diagonalización de *V*". Existen *p* números reales positivos $\lambda_1 \geq \ldots > \lambda_p \geq 0$ y *p* vectores asociados $(u_1, u_2, \ldots, u_p)$ que forman una nueva base ortogonal de $\mathbb{R}^P$ y que verifican:

$$Vu_k = \lambda_k u_k, \quad k = 1,2,\ldots,p$$

Sea $\Lambda = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & 0 & \lambda_p \end{pmatrix}$ pues tenemos $t_r = (\Lambda) = t_r = (V)$

$\lambda_1, \lambda_2, \ldots, \lambda_p$ son autovalores de V; $u_1, u_2, \ldots, u_p$ son los autovectores asociados de V.

• Solución del problema

El sub-espacio de dimensión *s*, solución del problema, es el que se ha generado por $\{u_1, u_2, \ldots, u_p\}$ y la inercia de la nube proyectada es: $\hat{I} = \lambda_1 + \lambda_2 + \cdots + \lambda_s$.

Aplicación numérica

Diagonalización de V_i: $\begin{cases} \lambda_1 = 30 & u_1 = (-0.89, -0.45) \\ \lambda_2 = 3.75 & u_2 = (0.45, -0.89) \end{cases}$

$$\Lambda = \begin{pmatrix} 30 & 0 \\ 0 & 3.75 \end{pmatrix}$$

Hubiéramos podido elegir $(-u_1)$ en lugar de u_1, o también $(-u_2)$ en lugar de u_2.

5.5. Componentes principales

El número de valores propios estrictamente positivos es igual al rango de *Λ(o de V, o de Y)*, suponemos generalmente que tenemos $p < n$ y que las *p* variables centradas y linealmente independientes (Fine, 1996); pues el rango es igual a *p*. Eso es lo que suponemos a continuación.

Sea:

$$u_k = \begin{pmatrix} u_{ik} \\ \vdots \\ u_{jk} \\ \vdots \\ u_{pk} \end{pmatrix} \; y \; c_k \sum_{j=1}^{p} u_{jk} y_k, \quad K = 1,\ldots,p \qquad (19)$$

Definimos, así p nuevas variables, combinaciones lineales de las p previas variables centradas y llamadas COMPONENTES PRINCIPALES.

$$c_{ik}\sum_{j=1}^{p} U_{jk}Y_{ik} \qquad k=1,\dots,p \qquad i=1,\dots,n$$

Se verifica que $C_k = 0$, $v(C_k) = \lambda_k$ y $cov(C_k,c_j) = 0$

Si $k \neq j$; dicho de otra manera los componentes principales son centrados, no correlacionados y sus varianzas son los valores propios; pues, la matriz de covarianzas de los componentes principales es Λ.

Sea $\underset{nxp}{C} = (c_{ik})$

Obtenemos las coordenadas del i-ésimo individuo en el nuevo sistema de ejes (origen en g y base $(u_1, u_2,\dots, u_p)$ en la i-ésima fila de $C:(C_{i1},\dots,C_{ip})$.

Aplicación numérica

Recordemos:

$$\begin{cases}\lambda_1 = 30 & u_1 = (-0.89,-0.45)\\ \lambda_2 = 3.75 & u_2 = (0.45,-0.89)\end{cases}$$

✓ $C_1 = (-0.89)X_1 + (-0.45)X_2$ Primera componente principal.

✓ $C_2 = (0.45)X_1 + (-0.89)X_2$ Segunda componente principal.

$$C = Xu = \begin{bmatrix}8&4\\3&4\\5&0\\-1&2\\1&-2\\-5&0\\-3&-4\\-8&-4\end{bmatrix}\begin{bmatrix}-0.89&0.45\\-0.45&-0.89\end{bmatrix} = \begin{bmatrix}-8.94&0\\-4.47&-2.236\\-4.47&2.236\\0&-2.236\\0&2.236\\4.47&-2.236\\4.47&2.236\\8.94&0\end{bmatrix}$$

Se verifica que:

✓ $\bar{c}_1 = 0, \quad \bar{c}_2 = 0$

$$\bar{c}_1 = \frac{(-8.94)+(-4.47)+(-4.47)+(0)+(0)+(4.47)+(4.47)+(8.94)}{8} = 0$$

$$\bar{c}_2 = \frac{(0)+(-2.236)+(2.236)+(-2.236)+(2.236)+(-2.236)+(2.236)+(0)}{8} = 0$$

✓ $V(C_1) = 0, \quad V(C_2) = 0$

$$V(C_1) = \frac{\sum(C_{i1}-\bar{C}_1)^2}{n} = \frac{240}{8} = 30.0 = \lambda_1$$

$$V(C_2) = \frac{\sum(C_{i2}-\bar{C}_2)^2}{n} = \frac{30}{8} = 3.75 = \lambda_2$$

✓ $Cov(C_1, C_2) = 0$

$$Cov(C_1,C_2) = \frac{1}{n}\sum_{i=1}^{n} C_{i1}C_{i2} = \frac{1}{8}[(0)+\cdots+(0)] = 0$$

✓ La matriz de las covarianzas de C_1, C_2 es Λ

$$\Lambda = \begin{pmatrix}V(C_1) & Cov(C_1,C_2)\\ Cov(C_1,C_2) & V(C_2)\end{pmatrix} = \begin{pmatrix}30.0 & 0\\ 0 & 3.75\end{pmatrix}$$

✓ Obtenemos las nubes de coordenadas de los individuos con respecto a la nueva base en las filas de C.

$$\begin{matrix}&Eje\ 1 & Eje\ 2\end{matrix}$$

$$C = X_u = \begin{bmatrix}-8.94&0\\-4.47&-2.236\\-4.47&2.236\\0&-2.236\\0&2.236\\4.47&-2.236\\4.47&2.236\\8.94&0\end{bmatrix}$$

El ACP completo corresponde a un cambio de referencial.

- El referencial inicial era $(0, e_1,\dots, e_p)$
- El nuevo referencial es $(g, u_1, u_2,\dots, u_p)$

5.6. Representación aproximada de los individuos en el plano factorial 1-2

El primer plano principal para g y su base es (μ_1, μ_2), los dos primeros autovectores de V.

Si $\widehat{W}_{ci}$ designa la proyección ortogonal sobre ese plano del individuo W_{ci}, las coordenadas de $\widehat{W}_{ci}$, con respecto a la base (μ_1, μ_2), son (C_{i1}, C_{i2}) y tenemos:

$$\|\widehat{W}_{ci}\|^2 = (C_{i1})^2 + (C_{i2})^2$$

Se reconoce la propiedad prevista en la solución ACP:

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n}\|\widehat{W}_{ci}\|^2 = \frac{1}{n}\sum_{i=1}^{n}(C_{i1})^2 + \frac{1}{n}\sum_{i=1}^{n}(C_{i2})^2 \Rightarrow$$

$$\hat{I} = V(C_1) + V(C_2) = \lambda_1 + \lambda_2 \qquad (20)$$

5.7. Calidad de representación

La calidad global de la representación se mide mediante:

$$\frac{\hat{I}}{I} = \frac{\lambda_1 + \lambda_2}{t_r(V)} \qquad (21)$$

La calidad de representación del individuo y_i (Figura 26) se mide (en aplicación del teorema de las tres rectas perpendiculares) mediante:

$$cos^2(W_{ci}, \widehat{W}_{ci}) = \frac{|\widehat{W}_{ci}|^2}{|W_{ci}|^2} = \frac{(C_{i1})^2}{|W_{ci}|^2} + \frac{(C_{i2})^2}{|W_{ci}|^2} \qquad (22)$$

$$cos^2(W_{ci}, \widehat{W}_{ci}) = cos^2(W_{ci}, \mu_1) + cos^2(W_{ci}, \mu_2)$$

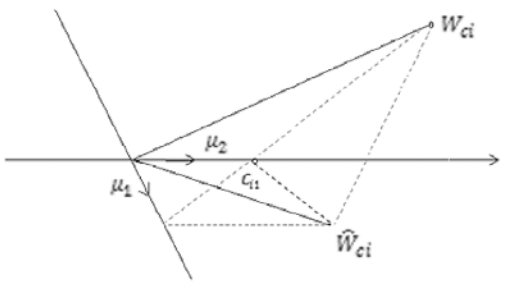


Figura 26. Calidad de representación del individuo y_i
Fuente: Adaptado de varios autores

✓ Representación aproximada sobre el primer plano principal (cuando $s = 2$)

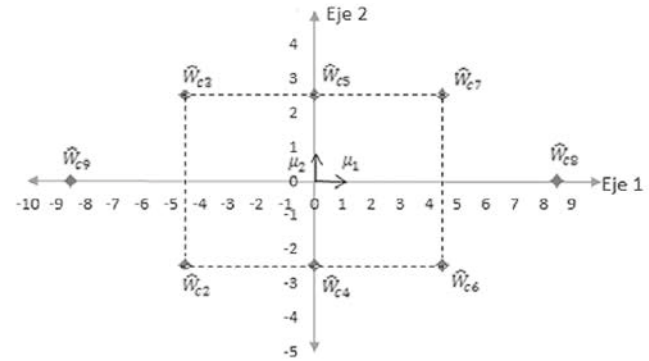


Figura 27. Representación aproximada sobre el primer plano 1-2
Fuente: Elaboración propia

Calidad global:

$$\frac{\hat{I}}{I} = \frac{30.0 + 3.75}{33.75} = 100\ \%$$

• Calidad de representación de los individuos

	Total	Eje 1	Eje 2
$Para\ W_{c1} = \frac{\|\widehat{W}_{c1}\|^2}{\|W_{c1}\|^2} = \frac{(-8.94)^2+(0)^2}{80}$	100 %	100 %	0 %
$Para\ W_{c2} = \frac{\|\widehat{W}_{c2}\|^2}{\|W_{c2}\|^2} = \frac{(-4.47)^2+(-2,236)^2}{25}$	100 %	80 %	20 %
$Para\ W_{c3} = \frac{\|\widehat{W}_{c3}\|^2}{\|W_{c3}\|^2} = \frac{(-4.47)^2+(2.236)^2}{25}$	100 %	80 %	20 %
$Para\ W_{c4} = \frac{\|\widehat{W}_{c4}\|^2}{\|W_{c4}\|^2} = \frac{(0)^2+(-2.236)^2}{5}$	100 %	0 %	100 %
$Para\ W_{c5} = \frac{\|\widehat{W}_{c5}\|^2}{\|W_{c5}\|^2} = \frac{(0)^2+(2.236)^2}{5}$	100 %	0 %	100 %
$Para\ W_{c6} = \frac{\|\widehat{W}_{c6}\|^2}{\|W_{c6}\|^2} = \frac{(4.47)^2+(-2,236)^2}{25}$	100 %	80 %	20 %
$Para\ W_{c7} = \frac{\|\widehat{W}_{c7}\|^2}{\|W_{c7}\|^2} = \frac{(4.47)^2+(2.236)^2}{25}$	100 %	80 %	20 %
$Para\ W_{c8} = \frac{\|\widehat{W}_{c8}\|^2}{\|W_{c8}\|^2} = \frac{(8.94)^2+(0)^2}{80}$	100 %	100 %	0 %

5.8. Representación de las variables en ($\mathbb{R}^n$, D)

Se dota $\mathbb{R}^n$ de la métrica de los pesos de los individuos.

Si todos los individuos tienen el mismo peso, se obtiene:

$$D = \frac{1}{n} = I_n = \begin{pmatrix} 1/n & \vdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1/n \end{pmatrix}$$

Y el producto escalar de dos variables x_j y x_k es entonces:

$$\langle x_j, x_k \rangle = \frac{1}{n} \sum_{i=1}^{n} x_{ij}\, , x_{ik} \qquad (23)$$

Sea: $y_j = x_j - \bar{x}_j$ (variables x_j centrada) y

$z_j = \frac{x_j - \bar{x}_j}{\sigma_j}$ Donde: $\sigma_j = \sqrt{V(x_j)}$ (variables x_j centrada reducida).

Sea $\perp$ = (1,1,…,1), el vector $\mathbb{R}^n$ cuyos elementos son todos iguales a 1.

Interpretación geométrica de los índices estadísticos

- Media $\langle x_j, \perp \rangle = \frac{1}{n} \sum_{I=1}^{n} x_{ij} = \bar{x}_j$
- Covarianza $\langle y_j, y_k \rangle = \frac{1}{n} \sum_{i=1}^{n} y_{ij}\, y_{ik} = cov(x_j, x_k)$
- Varianza $\|y_j\|^2 = \langle y_j, y_j \rangle = \frac{1}{n} \sum_{i=1}^{n} (y_{ij})^2 = V(\langle x_j \rangle)$
- Desviación estándar $\|y_j\| = \sigma_j$
- Correlación $cos(y_j, y_k) = \frac{\langle y_j, y_k \rangle}{\|y_j\| \|y_k\|} = \frac{cov(x_j, x_k)}{\sigma_j\, \sigma_k} = \rho(x_j, x_k)$

$1^{\perp}$ es el subespacio de $\mathbb{R}^n$ de dimensión n-1, ortogonal al vector $\perp$; S es la esfera de $1^{\perp}$ de centro 0 y rayo 1. Obtenemos la media al proyectar la variable sobre el eje dirigido por el vector 1. Obtenemos la variable centrada al proyectar la

variable sobre el subespacio ortogonal a 1. (Dentro de $\mathbb{R}^n$, centrar las variables equivale a proyectarlas sobre el subespacio ortogonal a 1.). La extremidad de la variable centrada reducida se ubica sobre la esfera unidad. La correlación de las variables es el coseno del ángulo que forma las variables centradas. Las variables están representadas por vectores de $\mathbb{R}^n$. El Ángulo que forman los vectores dentro de $1^{\perp}$ indica la correlación de esas variables (Figura 28).

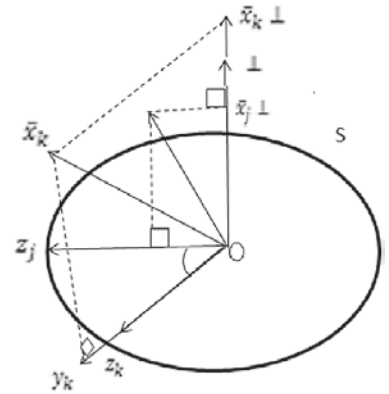


Figura 28. Representación de las variables como vectores
Fuente: Adaptado de varios autores

5.9. El ACP en $\mathbb{R}^n$, espacio de las variables

Propiedad

La primera componente principal C_1 es la combinación lineal de las variables centradas $y_j, j = 1,...,p$ que tiene la varianza máxima; la segunda componente principal C_2 es la combinación lineal de las variables centradas $y_j, j = 1,...,p$, no correlacionada con C_1 y que tiene la varianza máxima, etc.

5.10. Representación aproximada de las variables sobre el plano factorial 1-2

Las componentes principales $(C_1, C_2,..., C_p)$ forman una base ortogonal del subespacio de $\mathbb{R}^n$ de dimensión p generado por las p variables $y_1, y_2,..., y_p$.

Sea $f_k = \frac{C_k}{\sqrt{\lambda_k}}$, entonces $(f_1, f_2,..., f_p)$ es una base ortogonal de este subespacio y se obtiene:

$$z_j = \sum_{k=1}^{p} \rho(x_j, C_k)\ f_k \qquad j = 1, \dots p \tag{24}$$

Esas variables centradas reducidas son vectores cuyas extremidades se ubican sobre *S*, la esfera unidad de $1^{\perp}$.

Sea $\hat{z}_j$ la proyección de z_j sobre el primer plano principal (generado por f_1 y f_2), obtenemos:

$$\hat{z}_j = \sum_{k=1}^{2} \rho(x_j, C_k)\ f_k \qquad j = 1, \dots p \tag{25}$$

Las coordenadas de las variables centradas reducidas sobre el primer plano principal son las correlaciones de las variables con las componentes principales.

• Calidad de representación de las variables

Si dibujamos el círculo unidad sobre el primer plano factorial (principal), podemos medir virtualmente la calidad de representación de la variable z_j mediante $\hat{z}_j$.

En efecto, $\|z_j\| = 1$ y si la extremidad de $\hat{z}_j$ se ubica cerca del círculo unidad, también $\|\hat{z}_j\|$ será próximo de 1, pues obtenemos una buena calidad de representación.

En el ejemplo a continuación, este no se verifica para $\hat{z}_j$, pero sí se verifica que z_k, está bien representada por $\hat{z}_k$.

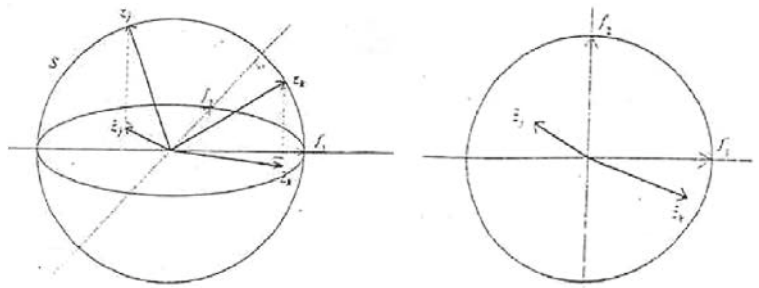


Figura 29. Representación de las variables.
La primera figura es una representación en el espacio de $1^{\perp}$;
la segunda figura es una aproximación en el primer plano principal.
Fuente: Adaptado de varios autores

Aplicación numérica

• Correlaciones de las variables con las componentes principales (Tabla 12)

$$\rho(x_1, C_1) = \frac{-26.84}{\sqrt{24.75}\sqrt{30.0}} = -0.985 \qquad \rho(x_1, C_2) = \frac{1.676^{**}}{\sqrt{24.75}\sqrt{3.75}} = 0.174$$

$$\rho(x_2, C_1) = \frac{-13.40}{\sqrt{9.0}\sqrt{30.0}} = -0.816 \qquad \rho(x_2, C_2) = \frac{-3.352}{\sqrt{9.0}\sqrt{3.75}} = -0.577$$

Donde:

$$cov(x_1, C_1) = \frac{1}{8}\underbrace{(-214.56)}_{*} = \frac{1}{n}\sum_{i=1}^{n} x_i c_i = -26.84$$

$$cov(x_1, C_2) = \frac{1}{8}\underbrace{(13.416)}_{**} = 1.677 \qquad cov\ (x_2, C_1) = -13.41$$

Tabla 12. Correlaciones de las variables con las componentes principales

	y_c		C_k		$cov\ (x_k, C_k)$			
	x_1	x_2	C_1	C_2	x_1C_1	x_1C_2	x_2C_1	x_2C_2
1	8	4	-8.94	0	-71.52	0	-35.76	0
2	3	4	-4.47	-2.236	-13.41	-6.708	-17.88	-8.944
3	5	0	-4.47	2.236	-22.35	11.18	0	0
4	-1	2	0	-2.236	0	2.236	0	-4.472
5	1	-2	0	2.236	0	2.236	0	-4.472
6	-5	0	4.47	-2.236	-22.35	11.18	0	0
7	-3	-4	4.47	2.236	-13.41	-6.708	-17.88	-8.944
8	-8	-4	8.94	0	-71.52	0	-35.76	0
	∑		0	0	*: -214.56	**:13.416	***: -107.28	-26.832
	$cov\ (x_{ki}, C_k)$				-26.84	1.677	-13.40	-3.354

✓ $\|\hat{z}_1\| = \sqrt{[\rho(x_1C_1)]^2 + [\rho(x_1C_2)]^2} = \sqrt{(-0.985)^2 + (0.174)^2} \Rightarrow$
$\|\hat{z}_1\| = \sqrt{0.970 + 0.03} \cong 1$

✓ $\|\hat{z}_1\| = \sqrt{(-0.816)^2 + (-0.577)^2} = \sqrt{0.666 + 0.3329} \cong 0.9993$

Las dos variables están bien representadas sobre el primer plano (Figura 30).

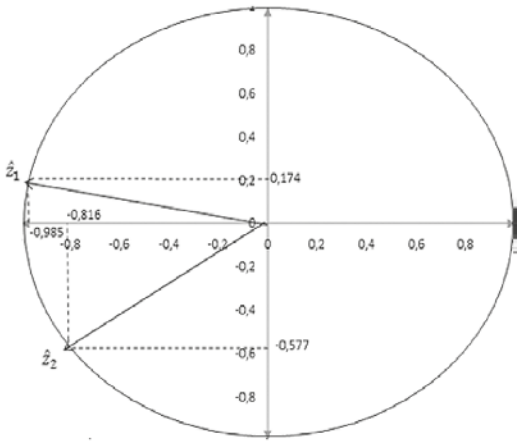


Figura 30. Representación aproximada de las variables del ejemplo numérico
Fuente: Adaptado de varios autores

5.11. Variables en el espacio de representación de los individuos

Los individuos están representados con puntos y las posiciones de los individuos están muy cerca cuando estos toman valores similares para cada variable. Las variables centradas y reducidas están representadas por vectores.

Para las variables bien representadas (es decir, las que tienen las extremidades de los vectores ubicadas cerca del círculo de correlación), el coseno del ángulo entre dos vectores es igual al coeficiente de correlación lineal de las dos variables que representan.

¿Será posible obtener una representación simultánea de los individuos y de las variables?

Hemos visto que, en el espacio de representación de los individuos, las variables están representadas por los ejes de las proyecciones: las variables iniciales ($y_1,...,y_p$) están representadas por los ejes generados por los vectores de la base canónica ($e_1, e_2,...,e_p$); las componentes principales ($C_1,...,C_p$) están representadas por los ejes generados por los autovectores ($u_1,...,u_p$).

Pues es posible, en el espacio de representación de los individuos, representar los ejes generados por los vectores ($e_1,...,e_p$).

Para $k = 1,...,p$, tenemos

$$u_k = \sum_{j=1}^{p} u_{jk} e_j \qquad (26)$$

Pero también

$$e_j = \sum_{j=1}^{p} u_{jk} u_k \qquad (27)$$

El vector e_j proyectado sobre el espacio principal de orden *S*, generado por los $(u_1, u_2,...,u_p)$ es:

$$\hat{e}_j = \sum_{j=1}^{p} u_{jk} u_k \qquad (28)$$

Si este vector está bien representado, el eje generado por este vector, puede considerarse como una buena representación de la variable y_j en el espacio de representación de los individuos.

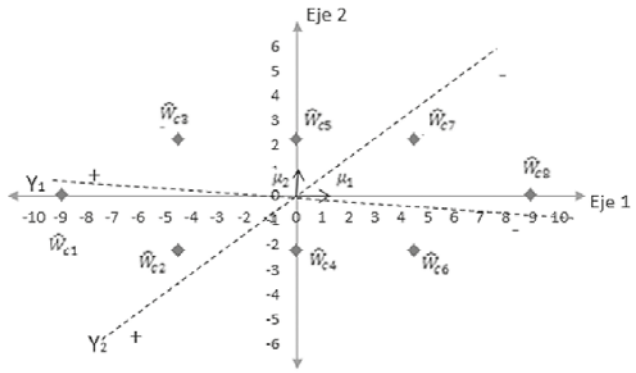


Figura 31. Representación simultánea en el primer plano factorial de individuos y variables
Fuente: Elaboración propia

5.12. Individuos y variables suplementarias

Podemos proyectar sobre los primeros planos factoriales (de $\mathbb{R}^p$ y de $\mathbb{R}^n$) individuos y variables que no han participado en el análisis (y por lo tanto, a la continuación de los ejes), pero que puede ayudar a la interpretación de las representaciones (Fine, 1996).

• Individuos suplementarios sobre el primer plano factorial

Sea $w_o = \begin{pmatrix} x_{o1} \\ x_{o2} \\ \vdots \\ x_{op} \end{pmatrix}$ ese individuo

Sean

$$w_{c0} = w_o - g, \qquad C_{o1} = \sum_{j=1}^{p} u_{1j} y_{0j} \quad y \quad C_{o2} = \sum_{j=1}^{p} u_{2j} y_{0j} \qquad (29)$$

Las coordenadas del individuo suplementario sobre la base (u_1,u_2) del primer plano factorial son entonces C_{o1} y C_{o2}.

• Variable suplementaria sobre el primer plano factorial

Sea $x_o = \begin{pmatrix} x_{1o} \\ x_{20} \\ \vdots \\ x_{no} \end{pmatrix}$ esa variable suplementaria,

$\rho(x_0, C_1)$ y $\rho(x_0, C_2)$ la correlación de x_0 con la primera y la segunda componente principal.

Las coordenadas de las variables suplementarias sobre la base (f_1,f_2) del primer plano factorial son entonces $\rho(x_0, C_1)$ y $\rho(x_0, C_2)$.

• Individuo suplementario

Sea $w_0 = \begin{pmatrix} 4 \\ 12 \end{pmatrix}$ un individuo suplementario por representar.

Tenemos $w_{c0} = w_0 - g = \begin{pmatrix} 4 \\ 12 \end{pmatrix} - \begin{pmatrix} 5 \\ 10 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$

$$\underset{u_{11}}{} \qquad \underset{u_{12}}{}$$
$$C_{o1} = (-0{,}89)(-1) + (-0{,}45)(2) = -0{,}01$$

$$\underset{u_{21}}{} \qquad \underset{u_{22}}{}$$
$$C_{o2} = (0{,}45)(-1) + (-0{,}89)(2) = -2{,}23$$

Entonces $\widehat{W}_{c0} = -0{,}01 * u_1 - 2.23 * u_2$

La calidad de representacion de w_{c0} por $\widehat{W}_{c0}$ se mide por:

$$\frac{\|\widehat{W}_{c0}\|}{\|W_c 0\|} = \underbrace{\frac{\sqrt{(-0{,}01)^2+(-2{,}23)^2}}{\sqrt{(-1)^2+(2)^2}}}_{\text{Coordenadas centradas en el total de variables}} = \frac{2{,}23}{2{,}236} \cong 1$$

• ***Variable suplementaria***

Sea $x_0 = \begin{pmatrix} 9 \\ 5 \\ 4 \\ 6 \\ 4 \\ 7 \\ 5 \\ 2 \end{pmatrix}$ una variable suplementaria por representar.

Tenemos $\bar{x}_0 = 5.25, \quad y_0 = \begin{pmatrix} 3.75 \\ -0.25 \\ -1.25 \\ 0.75 \\ -1.25 \\ 1.75 \\ -0.25 \\ -3.25 \end{pmatrix}$ y $\quad V(y_0) = 3.9375 \cong 3.94$

$$\rho(x_0, C_1) = \frac{-6.15}{\sqrt{3.94}\sqrt{30.0}} = -0.56 \qquad \rho(x_0, C_2) = \frac{-1.4}{\sqrt{3.94}\sqrt{3.75}} = -0.364$$

Entonces: $\hat{z}_0 = \rho(x_0, C_1)f_1 + \rho(x_0, C_2)f_2 = -0.56f_1 - 0.364f_2$

$$cov\,(x_0, C_1) = \frac{1}{n}\sum_{i=1}^{n} x_{0i}C_{1i}$$
$$= \frac{1}{8}[(3.75x - 8.94) + (-0.25x - 4.47) + (-1.25x - 4.47) + (0.75x0)$$
$$+ (-1.25x0) + (1.75x4.47) + (-0.25x4.47) + (-3.25x8.94)] = -6.15$$

$$cov\,(x_0, C_1) = \frac{1}{8}[(3.75x0) + (-0.25x - 2.24) + (-1.25x2.24) + (0.75x - 2.24)$$
$$+ (-1.25x2.24) + (1.75x - 2.24) + (-0.25x2.24) + (-3.25x0)] = -1.4$$

La calidad de representación de z_0 por $\hat{z}_0$ se mide por:

$$\|\hat{z}_0\| = \sqrt{\rho^2(x_0, C_1) + \rho^2(x_0, C_2)}$$

$$\|\hat{z}_0\| = \sqrt{(-0.56)^2 + (-0.364)^2} = 0.6679 = 66.79\%$$

5.13. Taller de Afianzamiento

ÁNALISIS EN COMPONENTES PRINCIPALES: ACP ($\mathbb{X}$, I_2, I_8).

Sea la matriz de datos (ocho individuos y dos variables).

$$Y' = \begin{matrix} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 \end{matrix}$$
$$\begin{bmatrix} 18 & 13 & 15 & 9 & 11 & 5 & 7 & 2 \\ 9 & 9 & 5 & 7 & 3 & 5 & 1 & 1 \end{bmatrix}$$

Realice geométricamente sobre papel cuadriculado, el ACP ($\mathbb{X}$, I_2, I_8) donde X es la matriz de datos centrados.

1) Diagrama de dispersión de Y (Figura 32).

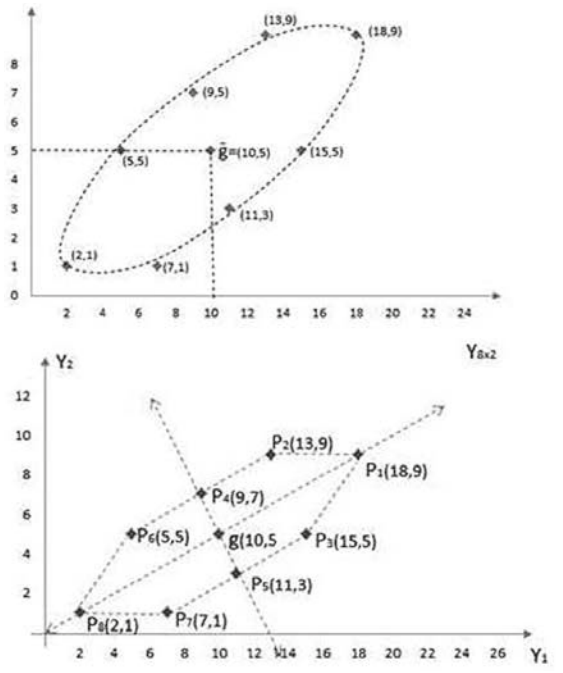


Figura 32. Diagrama de dispersión del ejemplo numérico
Fuente: Elaboración propia

2) Calcula el punto de gravedad y la matriz de datos centrados $\mathbb{X}$

$\tilde{g} = [10 \quad 5]$: centro de gravedad

$$\mathbb{X} = \mathbb{Y} - \mathbb{I}_n \tilde{g}^1 = \begin{bmatrix} 18-10 & 9-5 \\ 13-10 & 9-5 \\ 15-10 & 5-5 \\ 9-10 & 7-5 \\ 11-10 & 3-5 \\ 5-10 & 5-5 \\ 7-10 & 1-5 \\ 2-10 & 1-5 \end{bmatrix} \Rightarrow \begin{bmatrix} 8 & 4 \\ 3 & 4 \\ 5 & 0 \\ -1 & 2 \\ 1 & -2 \\ -5 & 0 \\ -3 & -4 \\ -8 & -4 \end{bmatrix} \begin{matrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ p_7 \\ p_8 \end{matrix} = \mathbb{X}$$

3) Grafique la nube de puntos centrados (Figura 33).

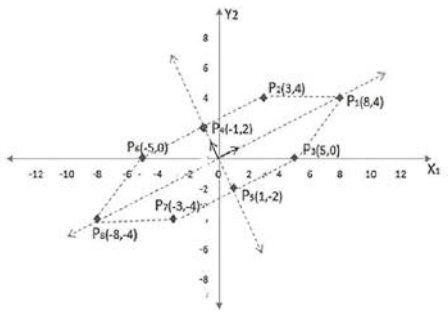
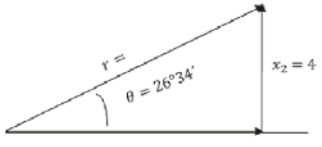


Figura 33. Nube de puntos centrados del ejemplo numérico
Fuente: Elaboración propia

4) Obtenga gráficamente los nuevos ejes, sobre la gráfica de individuos del punto anterior.

Para p_1: (8,4)



$$r^2 = x_1^2 + x_2^2 \Rightarrow r = \sqrt{8^2 + 4^2} = 8.9442$$
$$cos\ \theta = \frac{x_1}{r} = \frac{8}{8.9442} = 0.8944$$
$$sen\ \theta = \frac{x_2}{r} = \frac{4}{8.9442} = 0.4472$$

$$\theta = arc\cos(0.8944) = 26°34'$$
$$eje\ 1: \tilde{x}_1 = x_1 \cos\theta + x_2\ sen\ \theta$$
$$eje\ 2: \tilde{x}_2 = x_1(-\text{sen}\,\theta) + x_2\ cos\ \theta$$

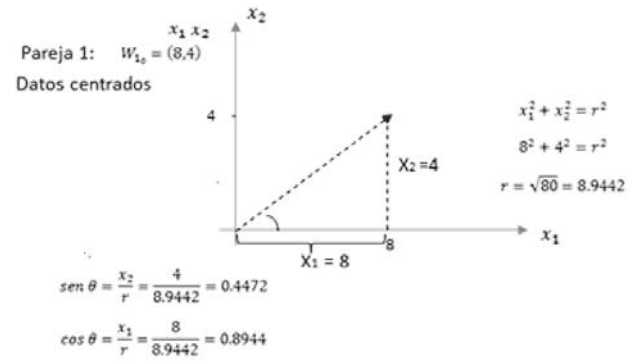
Si p_1: (8,4) $\Rightarrow \tilde{x}_{11} = 8(0.8944) + 4(0.4472)$

$$\tilde{x}_{11} = 8.944$$
$$\tilde{x}_{21} = 8(-0.4472) + 4(0.8944) = 0$$

$e_1 = (0.8944, 0.4472)$

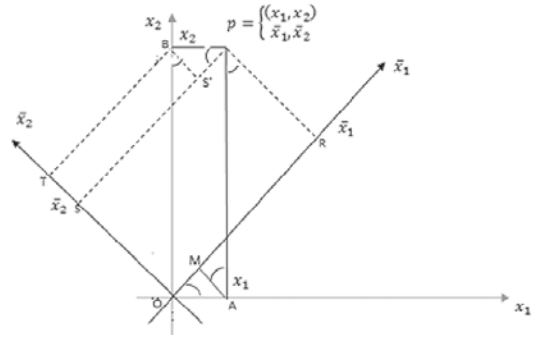
"Rotación" para $\rho_2(-1,2)$

$$r = \sqrt{(-1)^2 + (2)^2} = \sqrt{5}$$
$$\cos\theta = -\frac{1}{\sqrt{5}} = -0.4472$$
$$\text{sen}\,\theta = -\frac{2}{\sqrt{5}} = -0.8944$$
$$e_2 = (-0.4472, 0.8944)$$



Geometría elemental

Cuando se rotaron los ejes se dijo que era fácil demostrar que: $\bar{x}_1 = x_1\ cos\theta + x_2\ sen\theta$ y $\bar{x}_2 = -x_1\ sen\theta + x_2\ cos\theta$. Es decir, expresar los nuevos ejes en términos de los viejos y lo máximo de las coordenadas (Jiménez, 2012).



Consideraciones para la rotación de ejes.

1) Lo primero que puede verse al rotar los ejes es que P (A,B) es los ejes viejos, pasa a ser P (R,S) en los nuevos ejes.
2) El ángulo θ aparece en todos los lugares rotados < (por tener los lados respectivamente perpendiculares).

$$OA = BP = x_1 \quad ; \quad OB = AP = x_2 \qquad en\ ejes\ x_1x_2$$
$$OR = SP = \bar{x}_1 \quad ; \quad OS = RP = \bar{x}_2 \qquad en\ el\ sistema\ total$$

3) Todos los triángulos formados son rectángulos, lo que lleva a trabajar con relaciones trigonométricas simples.

$$OR = OM + MR \quad pero \quad OM = OA\cos\theta = x_1\cos\theta$$

$MR = MN + NR$ $pero$ $MN = AN\,\text{sen}\,\theta$; $NR = NP\ \text{sen}\,\theta \Rightarrow$ Por ello, $\bar{x}_1 = x_1\cos\theta + x_2 sen\,\theta$

En idéntica forma:

$$OS = OT - ST = OT - S'B$$

$$OT = OB\cos\theta = x_2\cos\theta$$

$$SB' = BP\ sen\,\theta = x_1 sen\,\theta = ST$$

$$OS = x_2\cos\theta - x_1\ sen\,\theta = -x_1\ sen\,\theta + x_2\cos\theta$$

El presentar así el último término o aún así:

$x_1\,(-sen\ \theta) + x_2\ cos\ \theta \Rightarrow$ Simplemente se entra a manejar matricialmente, lo cual tiene su ventaja.

Las nuevas coordenadas se pueden expresar así:

Componentes principales → Matriz de vectores propios

$$\begin{matrix} C_1 \leftarrow \\ C_2 \leftarrow \end{matrix}\begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & sen\,\theta \\ sen\,\theta & \cos\theta \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = A_x = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} X$$

Con lo cual ya podía asociarse a A la matriz de direcciones o matriz de disección coseno:

Para el ejemplo:

$$\begin{bmatrix} 0.8944 & 0.4472 \\ -0.4472 & 0.8944 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}\begin{bmatrix} eje\ 1 \\ eje\ 2 \end{bmatrix}\} \text{Rotación de ejes ortogonales (perpendicular)}$$

Donde:

$$\mathbb{Y} = \mathbb{X} - \overline{\mathbb{X}} = \begin{bmatrix} 8 & 4 \\ 3 & 4 \\ 5 & 0 \\ -1 & 2 \\ 1 & -2 \\ -5 & 0 \\ -3 & -4 \\ -8 & -4 \end{bmatrix}$$

$$\downarrow C_1 \qquad \downarrow C_2$$

La matriz $\underset{\downarrow}{C}$ $= \mathbb{X}u = \begin{bmatrix} 8 & 4 \\ 3 & 4 \\ 5 & 0 \\ -1 & 2 \\ 1 & -2 \\ -5 & 0 \\ -3 & -4 \\ -8 & -4 \end{bmatrix}\begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix} = \begin{bmatrix} 8.94 & 0 \\ 4.47 & -2.236 \\ 4.47 & 2.236 \\ 0 & -2.236 \\ 0 & 2.236 \\ 4.47 & -2.236 \\ 4.47 & 2.236 \\ 8.94 & 0 \end{bmatrix}$

$nuevos\ ejes$

Observaciones:

a) La matriz de vectores propios es ortogonal $AA' = I$.

b)

$$t_r(A\,A') = \underset{\downarrow}{\sum_i \sum_j a_{ij}^2} = \sum_{\Delta} \lambda_{\Delta}$$

Matriz ortogonal

c) A es una matriz definida positiva (no negativa)

d) Las distancias cuadradas C^2 y la densidad normal multivariada puede expresarse en términos de productos matriciales como formas cuadráticas.

Función cuadrática: $X'AX = \lambda_1 X'e_1e'_1X + \lambda_2 X'e_2e'_2X$

Si hacemos:

$$\left.\begin{matrix} X'e_1 = e_1^i x = y_1 \\ X'e_2 = e_2^i x = y_2 \end{matrix}\right\} \Rightarrow \underbrace{X'AX = \lambda_1 Y_1^2 + \lambda_2 Y_2^2 \geq 0}_{\text{Def. positiva}}$$

Entonces: $d_i^2 = X'AX > 0$ una distancia cuadrática es una forma cuadrática positiva que puede definirse (interpretarse) como una distancia al cuadrado (Figura 34).

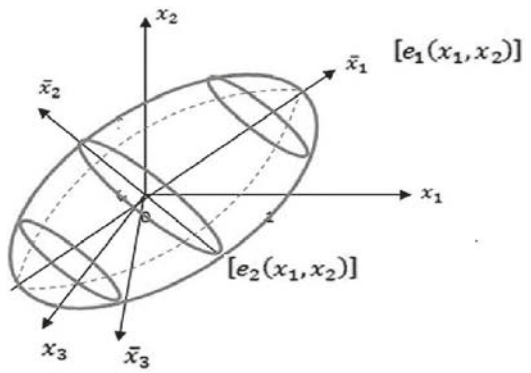


Figura 34. Elipse
Fuente: Adaptado de varios autores

$$d_i^2 = \sum_{i=1}^{p} \left(\frac{x_i - \mu_1}{\sqrt{\sigma_{ii}}}\right) = [x-\mu]\Sigma^{-1}[x-\mu] = \sum_{i=1}^{p} z_i^2 = x^2$$

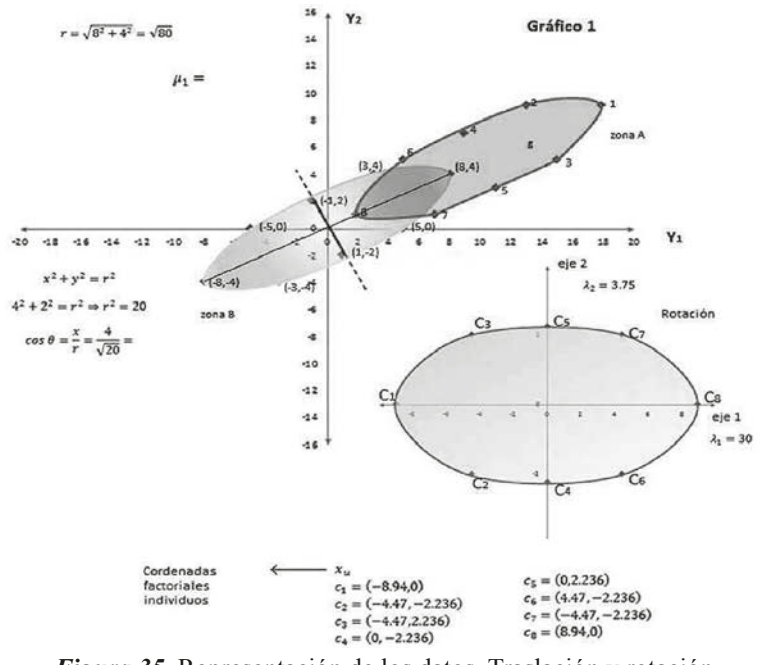


Figura 35. Representación de los datos. Traslación y rotación
Fuente: Elaboración propia

Geométricamente este resultado corresponde a una rotación del sistema de ejes (Figura 35). El primer eje tiene la dirección más alargada de la nube o sea la de mayor dispersión, que corresponde a la dirección de mayor inercia. El plano conformado por los dos primeros nuevos ejes, denominados factoriales, es la mejor fotografía de la nube de puntos.

5) Escriba la matriz con las nuevas coordenadas leyéndolas en la gráfica.

$$\mathbb{F} = \mathbb{X}u = \begin{bmatrix} 8 & 4 \\ 3 & 4 \\ 5 & 0 \\ -1 & 2 \\ 1 & -2 \\ -5 & 0 \\ -3 & -4 \\ -8 & -4 \end{bmatrix} \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix} = \begin{bmatrix} 8.94 & 0 \\ 4.47 & -2.236 \\ 4.47 & 2.236 \\ 0 & -2.236 \\ 0 & 2.236 \\ 4.47 & -2.236 \\ 4.47 & 2.236 \\ 8.94 & 0 \end{bmatrix}$$

$$\sum x_j x_k = 20$$

6) Dibuje el plano factorial de los individuos (Figura 36)

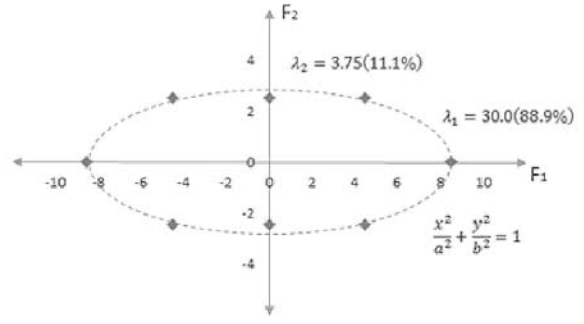


Figura 36. Plano factorial de los individuos del ejemplo numérico
Fuente: Elaboración propia

7) Obtenga visualmente un vector propio en la gráfica del punto 4 y normalice para obtener u_1 y obtenga visulamente u_2.

Para $\vec{e}_1$ se toma el punto $P(2,1)$

$$\text{Si } \mathbb{X} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \Rightarrow \sqrt{\mathbb{X}'\mathbb{X}} = \sqrt{(2 \quad 1)\begin{pmatrix} 2 \\ 1 \end{pmatrix}} = \sqrt{2^2 + 1^2} = \sqrt{4+1} = \sqrt{5}$$

$$u_1 = \frac{\mathbb{X}}{\sqrt{\mathbb{X}'\mathbb{X}}} = \frac{\begin{pmatrix}2\\1\end{pmatrix}}{\sqrt{5}} = \begin{pmatrix}2/\sqrt{5}\\1/\sqrt{5}\end{pmatrix} = \begin{pmatrix}0.8944\\0.4472\end{pmatrix}$$

Para $\vec{e}_2$ se tiene $\mathbb{X} = \begin{pmatrix}1\\-2\end{pmatrix}$; $\sqrt{\mathbb{X}'\mathbb{X}} = \sqrt{5}$

$$u_2 = \frac{\mathbb{X}}{\sqrt{\mathbb{X}'\mathbb{X}}} = \frac{\begin{pmatrix}1\\-2\end{pmatrix}}{\sqrt{5}} = \begin{pmatrix}1/\sqrt{5}\\-2/\sqrt{5}\end{pmatrix} = \begin{pmatrix}0.4472\\-0.8944\end{pmatrix}$$

8) A partir de los datos (coordenadas) del punto 5, calcule los dos valores propios.

$$I_1 = \lambda_1 = \sum_{i=1}^{I} p_i\, d^2(i,o) = \frac{1}{n}\sum_{i=1}^{n}(C_{i1} - C_i)^2 = 30$$

Donde: $\sum_{i=1}^{n}(C_{i1} - 0)^2 = (8.94)^2 + \cdots + (-8.94)^2 = 30$; $I_2 = \lambda_2 = 3.75$

El ACP completo corresponde a un cambio de referencia.

- La representación lineal es $(0, e_1, e_2, \ldots, e_p)$.
- El número macro de referencia es $(g, \mu_1, \mu_2, \ldots, \mu_p)$

Para la función característica, hallamos valores y vectores propios.

$$det(A - \lambda I) = 0$$

$$\left|\begin{pmatrix}1 & 0.7035\\0.7035 & 1\end{pmatrix} - \begin{pmatrix}\lambda & 0\\0 & \lambda\end{pmatrix}\right| = [(1-\lambda)^2 - (0.7035)^2] = 0$$

$$\Rightarrow 1 - \lambda = \pm(0.7035)\begin{cases}\lambda_1 = 1 + 0.7035 = 1.7035\\ \lambda_2 = 1 - 0.7035 = 0.2965\end{cases}$$

$$\sum \lambda_1 = 1.7035 + 0.2965 = 2 = Inercia$$

$$\cos(\mathbb{X}_1, \mathbb{X}_2) = 0.7035$$

$$\rho = \frac{\sum x_j\, x_k\, /n}{\sigma_j \sigma_k} = \frac{20/8}{\sqrt{24.75}\sqrt{9}} =$$

La correlación es el ángulo entre los dos vectores variables centrados.

$$\rho(h_j, h_k) = \frac{cov\,(r_j, r_k)}{\sigma_j \sigma_k} = \cos(\mathbb{X}_j, \mathbb{X}_k)$$

Centro de gravedad

$$g = \frac{1}{I}\sum_{i=1}^{I} Y_i = \frac{1}{I}\mathbb{Y}'\mathbb{I}_I$$

Centro de la nube de individuos: $\mathbb{Y}_c = \mathbb{Y} - \mathbb{I}_I g'$

Los datos transformados $z = \frac{x-\mu}{\sigma}$, nos brinda:

La matriz de correlación

$$\mathbb{Z} = \begin{bmatrix}1.608 & 1.333\\0.603 & 1.333\\1.005 & 0\\-0.201 & 0.666\\0.201 & -0.666\\-1.005 & 0\\-0.603 & -1.333\\1.608 & -1.333\end{bmatrix} \qquad \bar{x}_1 = 10.00 \quad S_1^2 = 24.75$$

$$\bar{x}_2 = 5.00 \quad S_1^2 = 9.0$$

$$s_1 = 4.975$$

$$s_2 = 3$$

$$cor\,(\mathbb{Z}) = \begin{pmatrix}1 & 0.7035\\0.7035 & 1\end{pmatrix}$$

Las correlaciones son el ángulo entre los dos vectores variables centrados.

$$\rho(r_j, r_k) = \frac{cov\,(r_j, r_k)}{\sigma_j \sigma_k} = \cos(\mathbb{X}_j, \mathbb{X}_k)$$

Por función característica, hallaremos valores y vectores propios.

$$det(A - \lambda I) = 0$$

$$(A - \lambda I) = 0 \Rightarrow \left|\begin{pmatrix}1 & 0.7035\\0.7035 & 1\end{pmatrix} - \begin{pmatrix}\lambda & 0\\0 & \lambda\end{pmatrix}\right| = 0$$

$$\left|\begin{pmatrix}1-\lambda & 0.7035\\0.7035 & 1-\lambda\end{pmatrix}\right| = 0 \Rightarrow [(1-\lambda)^2 - (0.7035)^2] = 0$$

$$(1-\lambda)^2 = 0.7035^2$$

$1-\lambda = \pm\sqrt{0.7035^2} \Rightarrow \begin{matrix} 1-\lambda = & 0.8387\ (1) \\ 1-\lambda = & -0.8387\ (2) \end{matrix}$

$1 - 0.8387 = \lambda$

Se va de $\mathbb{R}^n$ a $\mathbb{R}$ se consigue una nueva base.

$$\mathbb{W}\frac{1}{n}I_n = \underbrace{\frac{1}{n}XX'}$$

dim= $\min(X, X') = P$

Se hallan los valores propios.

$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_\alpha \geq \cdots \geq \mu_p \geq 0 \ldots 0$

Tienen asociados los vectores propios $v_1\ v_2 \ldots v_p$

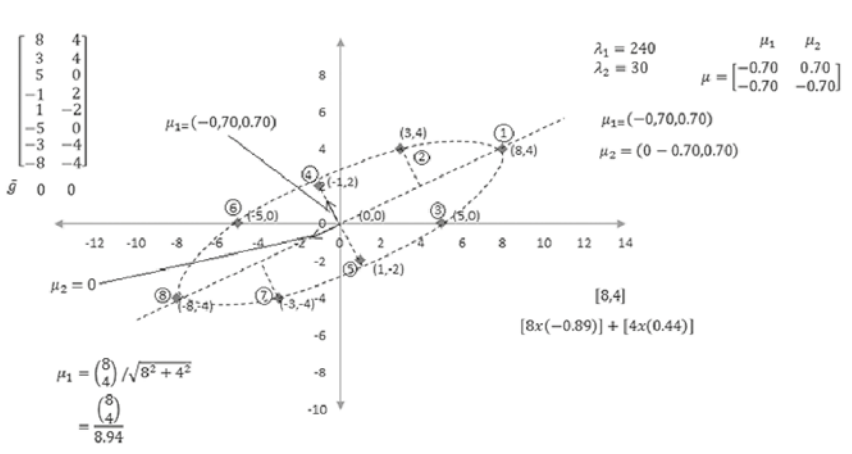


Figura 37. Traslación de los individuos (centrado) del ejemplo numérico
Fuente: Elaboración propia

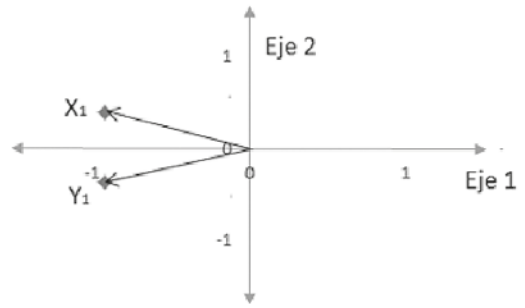
9) Obtenga los dos vectores de coordenadas de las variables (Figura 37).

Coordenadas

$$\mathbb{X}\ \underset{Y_c}{\mu_\alpha \downarrow} = \begin{bmatrix} \quad \end{bmatrix} \begin{bmatrix} -0.89 & -0.44 \\ 0.44 & -0.89 \end{bmatrix}$$

Pca$co:

10) Dibuje el primer plano factorial de las variables.



Ejemplo 1. Datos originales, $\mathbb{X} = y_{ij}$

$$\mathbb{X}^T = \begin{bmatrix} 18 & 13 & 15 & 9 & 11 & 5 & 7 & 2 \\ 9 & 9 & 5 & 7 & 3 & 5 & 1 & 1 \end{bmatrix}_{2x8} \qquad n = 8,\ p = 2$$

$$\mathbb{M} = \mathbb{I}_2; \mathbb{D} = \frac{1}{8}\mathbb{I}_8$$

Ejemplo 2. Datos centrados, $\mathbb{X} = y_{ij} - \overline{y}_j$

$$\mathbb{X}^T = \begin{bmatrix} 8 & 3 & 5 & -1 & 1 & -5 & -3 & -8 \\ 4 & 4 & 0 & 2 & -2 & 0 & -4 & -4 \end{bmatrix} \qquad \overline{Y}_1 = 10,\ \overline{Y}_2 = 5$$

$$\mathbb{M} = \mathbb{I}_2; \mathbb{D} = \frac{1}{8}\mathbb{I}_8$$

Ejemplo 3. Datos estandarizados, $\mathbb{X} = \frac{y_{ij} - \overline{y}_j}{s_j}$

$$\mathbb{X}^T = \begin{bmatrix} 1.608 & 0.603 & 1.005 & -0.201 & 0.201 & -1.005 & -0.603 & -1.608 \\ 1.333 & 1.333 & 0 & 0.666 & -0.666 & 0 & -1.3334 & -1.333 \end{bmatrix}$$

$$\mathbb{M} = \mathbb{I}_2; \mathbb{D} = \frac{1}{8}\mathbb{I}_8$$

Tarea. Hallar la matriz de varianza-covarianza para cada uno de los casos de los ejemplos 1, 2, 3.

III. TÉCNICAS MULTIVARIADAS BÁSICAS

6. ANÁLISIS EN COMPONENTES PRINCIPALES (ACP)

6.1. Dominio de aplicación

El análisis en componentes principales es útil en la lectura de tablas de "individuos" por variables cuantitativas. Se tiene por costumbre escribir en las filas a los "individuos" que representan las unidades estadísticas en un análisis. El objetivo principal del ACP, es comparar individuos según valores de las variables continuas que se parecen. Dos *individuos* se parecen porque obtienen más o menos los mismos indicadores en diferentes aspectos. Un segundo objetivo, ver relaciones entre variables que están describiendo a los individuos; un tercer objetivo, útil cuando se requieren otros análisis posteriores, es reducir la dimensionalidad del problema. Una alta correlación entre variables dará como resultado que unas pocas variables sintéticas resuman lo importante de la información de las variables originales (Lebart *et al.,* 1995).

6.2. Orígenes del Análisis en Componentes Principales

En 1901 Karl Pearson publicó un trabajo sobre el ajuste de un sistema de puntos en un multiespacio a una línea o un plano. Este enfoque fue retomado en 1933 por Hotelling, quien fue el primero en formular el análisis por componentes tal como se ha difundido hasta nuestros días, centrado en el análisis que sintetiza la mayor variabilidad del sistema de puntos.

El trabajo original de Pearson (1901) se centraba en aquellos componentes, o combinaciones lineales de variables originales, para los cuales la varianza no explicada fuera mínima. Estas combinaciones generaban un plano, función de las variables, en el cual el ajuste del sistema de puntos es "el mejor", por ser mínima la suma de las distancias de cada punto al plano de ajuste (Pla, 1986). Desde sus orígenes, el análisis por componentes principales ha sido aplicado en situaciones muy variadas: en psicología, medicina, meteorología, geografía, ecología, agronomía.

6.3. Fundamentos del método

El 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 análisis en componentes principales ponderado.

6.3.1. La nube de variables

La representación geométrica de las variables (nube de variables) es menos familiar pero igualmente importante. Conviene pensarla no como puntos sino como flechas que terminan en los puntos correspondientes a las coordenadas de las variables. Lo que interesa observar en las variables es su relación y la analogía geométrica en su representación es el ángulo entre las variables.

La representación gráfica tradicional del ACP es un gráfico bidimensional (llamado primer plano factorial) que captura la mayor proporción de la variabilidad presente en la muestra *n*. Las variables transformadas aparecen aquí como vectores cuya proyección sobre cada eje ortogonal representa la influencia de la variable respectiva sobre el correspondiente componente principal. El coseno del ángulo entre dos de las variables originales (en realidad entre los vectores que lo representan) medido en el espacio coordenado, es una medida directa de la correlación entre dichas variables. Así, si el ángulo es: *próximo a cero*, la correlación es estrecha y positiva; *cercano a 180°*, la correlación es también estrecha pero negativa; finalmente, si el ángulo es *cercano a 90°*, las variables están escasamente relacionadas.

El primer plano factorial corresponde al plano que corta a la hiper-esfera de tal manera que se logra la mejor proyección de las variables.

Una variable que quede exactamente sobre el plano factorial tendrá longitud 1 en su proyección, y su flecha terminará en un círculo de radio uno, que se denomina ***círculo de correlaciones***. Adicionalmente, cada individuo (fila) de la muestra puede ser representado en el nuevo espacio coordenado. Si el gráfico se superpone al anterior se denomina *biplot* (Gabriel, 1971).

6.4. El Análisis en Componentes Principales (ACP) como *ACP(X, M, D)*.

Las principales fórmulas del ***ACP(X, M, D)*** se resumen en la Tabla 4, de donde se pueden derivar las de un método particular una vez se han establecido las tres matrices***(X, M, D)***.

6.4.1. Análisis en Componentes Principales (ACP) de datos originales

$\boldsymbol{X} = [y_{ij}]$ $\quad\quad$ $\boldsymbol{M} = \boldsymbol{I}_p$ $\quad\quad$ $\boldsymbol{D} = \frac{1}{n}\boldsymbol{I}_n$

(Pearson, 1901; Fine, 1996)

6.4.2. Análisis en Componentes Principales (ACP) centrado por columnas

$\boldsymbol{X} = [y_{ij} - \overline{y}_j]$ $\quad\quad$ $\boldsymbol{M} = \boldsymbol{I}_p$ $\quad\quad$ $\boldsymbol{D} = \frac{1}{n}\boldsymbol{I}_n$

Con: $\overline{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$ (media ponderada por columnas)

(Escofier & Pagès, 1988-1998; Crivisqui, 1993)

6.4.3. Análisis en Componentes Principales (ACP) centrado por filas

$\boldsymbol{X} = [y_{ij} - \overline{y}_i]$ $\quad\quad$ $\boldsymbol{M} = \boldsymbol{I}_p$ $\quad\quad$ $\boldsymbol{D} = \frac{1}{n}\boldsymbol{I}_n$

Con: $\overline{y}_i = \frac{1}{p}\sum_{j=1}^{p} y_{ij}$ (media ponderada por filas)

(Vertel & Pardo, 2010)

6.4.4. Análisis en Componentes Principales (ACP) normado-centrado por columnas

$\boldsymbol{X} = \left[\frac{y_{ij} - \overline{y}_i}{\sigma_j}\right]$ $\quad\quad$ $\boldsymbol{M} = \boldsymbol{I}_p$ $\quad\quad$ $\boldsymbol{D} = \frac{1}{n}\boldsymbol{I}_n$

Con: $\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \overline{y}_i)}$ (desviación estándar)

(Noy-Meir, 1973; Noy-Meir *et al.,* 1975)

Las nuevas variables generadas a través de la técnica del ACP ponderado se denominan *componentes principales* y poseen algunas características estadísticas deseables (independencia, y en todos los casos no correlación). Esto significa, que si las variables originales no están correlacionadas, el análisis en componentes principales no ofrece ventaja alguna.

La literatura acerca de la construcción de los componentes principales, su uso y sus propiedades es muy amplia; en este libro se utilizará la *escuela francesa*. Casi en todos los libros de texto de análisis multivariado de datos se dedica un capítulo al análisis en componentes principales bajo la escuela anglosajona, pueden consultar, por ejemplo: Chatfields & Collins (1980), Morrison (1976), Kendall (1980), Harris (1967), Anderson (1984), Mardia y colaboradores (1982), que presentan un enfoque que acentúa los aspectos teóricos.

6.5. Elementos suplementarios

Es posible proyectar elementos suplementarios o ilustrativos (individuos, variables continuas y variables nominales) sobre los planos construidos en el ACP. Los elementos suplementarios permiten explorar asociaciones con los elementos activos o enriquecer los análisis. Los elementos que participan en el análisis se denominan activos, en el caso del ACP son variables continuas activas e individuos activos (Cabarcas & Pardo, 2001).

6.5.1. Individuos suplementarios

Un individuo que tiene los valores para todas las variables activas pero que no participó en el ACP se puede proyectar sobre los ejes obtenidos de la misma forma que los activos. Mediante este procedimiento se puede posicionar un

nuevo individuo con respecto a todos los activos para responder a objetivos preestablecidos, por ejemplo explorar su posible discriminación entre grupos.

6.5.2. Variables nominales ilustrativas

Como una variable nominal representa un partición (división en clases) de los individuos, lo que se proyecta son los centros de gravedad de cada subconjunto asociado a una modalidad.

6.5.3. Variables continuas

En el ACP normado la proyección de una variable continua ilustrativa equivale a su correlación con el eje, lo que da la clave para su interpretación.

6.6. Ejemplo de ACP centrado-ponderado

El ACP que se utiliza aquí es el normado-centrado por columnas, es decir, que las variables originales se estandarizan. En resumen, el ACP de la tabla [X], notado $\boldsymbol{ACP}(\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{D})$, es el $ACP\left(\left[\frac{y_{ij}-\overline{y}_i}{\sigma_j}\right], \boldsymbol{I}_p, \frac{1}{n}\boldsymbol{I}_n\right)$.

En el análisis en componentes principales se recomienda seguir los siguientes pasos:

6.6.1. Aplicación: Perfil socioeconómico de los departamentos caribeños colombianos

El objetivo de esta investigación es realizar la caracterización socioeconómica de la costa Caribe colombiana (Tabla 13). En este breve informe, se presentan algunos indicadores económicos y sociales de la región Caribe colombiana (DNP, 2005).

Con este ejercicio se pretende contribuir a facilitar la toma de decisiones. Por tal razón, política y administrativamente, la región Caribe está conformada por los departamentos de Atlántico, Bolívar, Cesar, Sucre, Córdoba, Magdalena, La Guajira y San Andrés.

Tabla 13. Datos socioeconómicos de los departamentos del Caribe

Departamentos/Variables	NBI	Analfabetismo	Desempleo	PIB *per cápita*
Atlántico	16,1	4,5	12,8	1,67
Bolívar	30,0	9,5	9,8	8,59
Cesar	35,7	13,7	6,2	1,36
Córdoba	35,8	15,8	13,6	1,20
La Guajira	37,5	14,4	5,9	3,02
Magdalena	30,6	13,4	7,1	0,75
Sucre	42,4	15,3	6,2	0,80

Nota: Necesidades Básicas Insatisfechas (NBI), analfabetismo, desempleo y Producto Interno Bruto (PIB) *per cápita* son evaluados en porcentaje (%). El archipiélago de San Andrés y Providencia no se tuvo en cuenta por presentar indicadores socioeconómicos atípicos.
Fuente: Elaboración propia

Estadísticas básicas

	NBI	Analfabetismo	Desempleo	PIB *per cápita*
Mínimo	16,10	4,50	5,9	0.750
1er. Cuartil	30,30	11,45	6,2	1.000
Mediana	35,70	13,70	7,1	1.360
3er. Cuartil	36,65	14,85	11,3	2.345
Máximo	42,40	15,80	13,6	8.590

Fuente: Elaboración propia

6.6.2. Análisis de tablas y gráficos descriptivos

En la aplicación de indicadores socioeconómicos de los departamentos caribeños, las variables continuas son evaluadas en porcentaje (%).

En la práctica, las variables suelen tener diferente escala aún en los casos en que las unidades de medida sean las mismas. En la Tabla 13 aparecen las variables continuas activas, las cuales se utilizan en el ACP, junto con sus estadísticas básicas.

Una variable estandarizada pierde las unidades de medida con lo cual se elimina el efecto de la escala de medida. En la Tabla 14 se observa que NBI (%) contribuye más a la inercia (66,89 %). Al hacer la estandarización de los datos, se presenta que cada variable contribuye en la misma cantidad a la inercia.

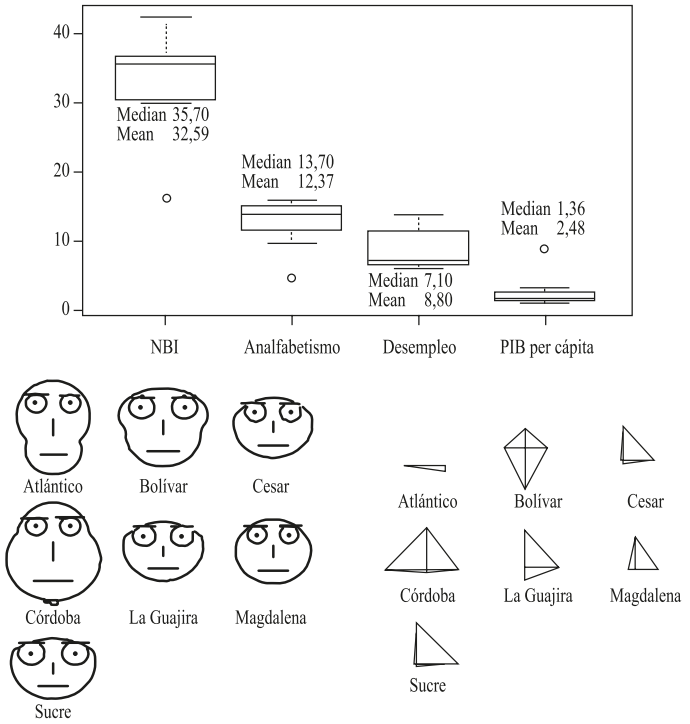


Figura 38. *Boxplot* e indicadores numéricos de las variables socioeconómicas; rostros de *Chernoff* y gráfico de *estrellas* para departamentos en estudio
Fuente: Elaboración propia

Como se sabe, la desviación estándar de una variable estandarizada es igual a uno y los valores de las covarianzas entre variables estandarizadas son iguales a sus coeficientes de correlación. Con la estandarización todas las variables contribuyen igual a la inercia (con 1), de modo que la *inercia total* es igual *al número de variables*.

Tabla 14. Varianza y contribución a la inercia de variables originales y estandarizadas

A. Datos originales					
Variable	**NBI**	**Analfabetismo**	**Desempleo**	**PIB perc**	**Inercia**
Varianza	70,48	16,24	10,81	7,83	105,36
% Inercia	66,89	15,41	10,26	7,43	100,00
B. Datos estandarizados					
Variable	**NBI**	**Analfabetismo**	**Desempleo**	**PIB perc**	**Inercia**
Varianza	1	1	1	1	4
% Inercia	25,0	25,0	25,0	25,0	100,00

Fuente: Elaboración propia

En el ACP normado-centrado las coordenadas de los individuos son los valores estandarizados de cada una de las variables. El valor de una variable estandarizada corresponde para un individuo a su distancia al promedio expresada en el número de desviaciones estándar (Tabla 15). En el caso de las variables evaluadas en los departamentos caribeños se observa que suelen tener mayor porcentaje unos aspectos que otros; con el ACP normado-centrado se hace jugar un papel similar a todas las variables.

Tabla 15. Tabla de datos de indicadores socioeconómicos estandarizados

```
> acp$tab
                NBI      analfab      desempleo        PIBperc
Atlantico -2.1209652   -2.1093932      1.3140762     -0.3142921
Bolivar   -0.3326644   -0.7694883      0.3285190      2.3566391
Cesar      0.4006676    0.3560319     -0.8541495     -0.4339436
Cordoba    0.4135331    0.9187920      1.5768914     -0.4956992
Guajira    0.6322461    0.5436186     -0.9527052      0.2067711
Magdalena -0.2554715    0.2756376     -0.5584824     -0.6693870
Sucre      1.2626543    0.7848015     -0.8541495     -0.6500883
```

Fuente: Elaboración propia

6.6.3. Descomposición de la inercia de la tabla

El primer paso en el Análisis en Componentes Principales normado-centrado (ACP) tiene como objeto reducir las dimensiones de la matriz de datos inicial, en el presente caso, el de una tabla de datos de INDICADORES SOCIOECONÓMICOS de los departamentos caribeños (Tabla 13).

La Tabla 16 es la matriz de correlaciones. Las Necesidades Básicas Insatisfechas (NBI) tienen correlación positiva con índice de analfabetismo (0,93), y correlación negativa con el índice de desempleo (-0,58).

Tabla 16. Matriz de correlaciones

	NBI	analfab	desempleo	PIBperc
NBI	1.0000000			
Analfab	0.9309641	1.0000000		
Desempleo	-0.5892419	-0.4603070	1.0000000	
PIBperc	-0.1450457	-0.3346681	0.0974725	1.0000000

Fuente: Elaboración propia

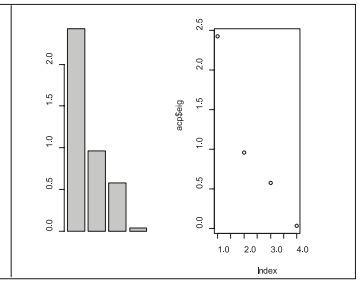
De esta manera, se obtienen los distintos ejes factoriales o direcciones principales de alargamiento de la nube de puntos que explican las variaciones que

se producen en dicha matriz (Tabla 17), los cuales posteriormente permitirán la representación factorial de la información contenida en la tabla. Cada eje factorial viene acompañado de su propio valor, y del porcentaje de inercia, que representan la varianza explicada contenida en cada eje, así como su importancia relativa porcentual.

Tabla 17. Valores propios, inercia acumulada, porcentaje de inercia, y porcentaje de inercia acumulada de los ejes factoriales para el ACP centrado-normado

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


 Eigenvalues * 1000
  Eigenvalue CumInertia Percent CumPercent
1     2424.8      2424.8    60.6       60.6
2      961.0      3385.8    24.0       84.6
3      576.1      3961.9    14.4       99.0
4       38.1      4000.0     1.0      100.0
```



Fuente: Elaboración propia

La decisión para saber cuántos ejes es conveniente analizar está soportada en la distribución de valores propios (acp$eig): [1] 2.425 (60,6 %), 0,961 (24 %), 0,576 (14,4 %), 0,038 (0,95 %). En este caso, se decide utilizar los dos primeros ejes para la síntesis del perfil socioeconómico de los departamentos caribeños (84,6 % de la inercia, es decir, el porcentaje de la variabilidad total acumulado en los dos primeros componentes principales alcanza a 84,6 %). Para el presente estudio a manera de ejemplo solo se analizarán los dos primeros ejes.

6.6.4. Extracción de los ejes factoriales (Tablas 18, 19 y 20)

Antes de interpretar los resultados obtenidos del ACP centrado-normado, se debe definir cada uno de los ejes factoriales. Como el ACP centrado-normado es un ACP ponderado, para encontrar las coordenadas factoriales (filas, columnas) se utilizan las fórmulas de la Tabla 9.

Las coordenadas factoriales y ayudas a la interpretación (coordenadas factoriales, contribuciones, cosenos cuadrados, distancias al cuadrado) para filas (*departamentos*) y columnas (*variables de estudio*) se presentan en las Tablas 18, 19 y 20. Brevemente, se explica cómo se realizan algunas ayudas a la interpretación para el caso particular de ACP.

Tabla 18. Distancias al cuadrado, contribución de cada individuo a la inercia

IDEN	Atlántico	Bolívar	Cesar	Córdoba	Guajira	Magdalena	Sucre
Peso	0,14	0,14	0,14	0,14	0,14	0,14	0,14
Dist2	10,77	6,36	1,21	3,75	1,65	0,90	3,36
Inercia	1,54	0,91	0,17	0,53	0,24	0,13	0,48
% Inercia	38,47	22,73	4,30	13,38	5,87	3,21	12,00

Fuente: Elaboración propia

Tabla 19. Coordenadas y ayudas a la interpretación de los departamentos

	Coordenadas			Contribuciones		Cosenos²		
	G1	G2	G3	G1	G2	G1	G2	Dist2
Atlántico	-3,09	0,96	-0,53	56,3	13,7	88,7	8,6	10,77
Bolívar	-1,37	-2,07	0,43	11,1	64,0	29,5	67,7	6,36
Cesar	0,95	0,08	-0,54	5,4	0,1	75,5	0,6	1,21
Córdoba	0,19	0,94	1,68	0,2	13,1	1,0	23,5	3,75
Guajira	1,10	-0,57	-0,31	7,2	4,9	73,9	19,9	1,65
Magdalena	0,43	0,51	-0,58	1,1	3,9	20,1	28,9	0,90
Sucre	1,79	0,15	-0,15	18,8	0,4	95,0	0,7	3,36

Tabla 20. Coordenadas y ayudas a la interpretación de las variables

	Coordenadas			Contribuciones		Cosenos z		
	G1	G2	G3	G1	G2	G1	G2	plano
NBI	0,95	-0,18	0,23	37,0	3,3	89,6	3,2	92,8
Analfabetismo	0,94	0,06	0,32	36,2	0,4	87,8	0,4	88,1
Desempleo	-0,72	0,31	0,62	21,3	9,9	51,7	9,5	51,7
PIBperc	-0,37	-0,91	0,19	5,5	86,4	13,4	83,1	96,4

Fuente: Elaboración propia

Calidad de la representación

Ejemplo: Para la primera fila (Atlántico)

Fila 1 (Atlántico)	F_1^2 **(1)** $= (-3.0916)^2$ $= 9.5576$	$\|i\|^2$ $= dist^2$ $= 10.7736$	$cos_s^2(i) = \frac{F_s^2(i)}{\|i\|^2}$	$cos^2_{s=1}$ **(i = 1) = 88.7%**

Ejercicio: Para la primera columna (NBI)

Columna 1 (NBI)				

Contribución absoluta

Ejemplo: Para la primera fila (Atlántico) en el eje 1

Fila 1 (Atlántico)	F_1^2 **(1)** $= (-3.0916)^2$ $= 9.5576$	$p_i = 0.1428$	$\lambda_s = 2.4248$	$con^2_{s=1}(i = 1) = \frac{p_i F_s^2(i)}{\lambda_s}$ $= 56.3\%$

Ejercicio: Para la primera columna (NBI) en el eje 1

Columna 1 (NBI)				

6.6.5. Interpretación de los ejes factoriales

La nube de variables tiene en el caso normado una representación más sencilla. La nube de individuos cambia con respecto al no normado pero su interpretación sigue siendo la misma. Las ayudas presentadas en las Tablas 18, 19 y 20 permiten controlar y complementar las lecturas de los planos factoriales. La contribución a la inercia del eje ayuda a encontrar un significado del eje.

Departamentos-Individuos (filas), en el primer eje factorial departamentos Atlántico, Sucre, Bolívar, La Guajira y Cesar son los que más contribuyen (acumulan el 98,8 % = 56,3 % + 18,8 % + 11,1% + 7,2 % + 5,4 % de la inercia del eje), entonces el eje 1 se interpreta principalmente como la contraposición de departamentos del Atlántico y Bolívar con los departamentos Cesar, La Guajira y Sucre.

En el segundo eje, se contraponen los departamentos de Bolívar con Atlántico y Cesar (contribuyen con el 64,0 % + 13,4 % +13,1 % = 90,2 % de la inercia del segundo eje).

Categorías-variables (columnas), en Figura 39 se observa que el primer eje tiene una alta correlación positiva con los índices de analfabetismo (0,94) y NBI (Necesidades Básicas Insatisfechas) (0,94). El primer eje tiene una correlación negativa no tan fuerte con el índice de desempleo (-0,72).

Hay una correlación fuerte y positiva entre las variables Analfabetismo y Necesidades Básicas Insatisfechas. Estas a su vez, tienen una correlación negativa con la variable Desempleo.

El segundo eje separa hacia arriba departamentos con menos Producto Interno Bruto *per cápita*, y hacia abajo, con más Producto Interno Bruto *per cápita*.

6.6.6. Interpretación del plano factorial (1-2)

Una vez que se describen los ejes (I y II) que van a permitir caracterizar el estudio, el siguiente paso en la investigación es analizar los planos factoriales que se forman con la unión de los ejes (en forma de pares) que el investigador de acuerdo al análisis decidió tomar en cuenta.

Se representa en la mayoría de las aplicaciones, cada uno de los individuos-filas que forman la muestra como un punto en el plano factorial (Figura 39). Es también, posible identificarlos con un número, que se prefiere muchas veces omitir para facilitar la interpretación general. Para esta aplicación, como tiene pocos individuos se colocaron sus codificaciones (nombres).

Cada zona del plano factorial definido por los componentes principales sintetiza una problemática diferente, de manera que una vez conocida la ubicación de una finca en el plano es posible sacar conclusiones acerca de su situación respecto a la producción.

<u>*Lectura simultánea:*</u> (Figura 39)
Primer grupo: Sucre, La Guajira y Cesar presentan altos índices de Necesidades Básicas Insatisfechas y Analfabetismo, como porcentajes bajos de desempleo. Esto significa, que el subempleo es una fuente de trabajo (tomada en las estadísticas nacionales como fuente de empleo formal) en las zonas urbanas, y en la zona rural por ser una zona agropecuaria, el capital humano se dedica a trabajar por días (jornalero).

Segundo grupo: Atlántico presenta altos índices de desempleo y bajos niveles de NBI y Analfabetismo.

Tercer grupo: Bolívar presenta un fuerte aporte económico del Producto Interno Bruto *per cápita*.

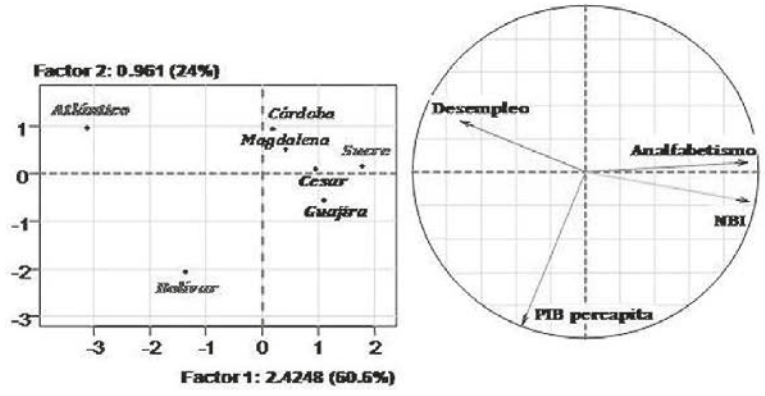


Figura 39. Plano factorial 1-2 del *ACP:* filas-individuos y columnas-variables
Fuente: Elaboración propia

6.6.7. Integración de los resultados en su contexto

Se debe resaltar que el ACP centrado-normado por sí solo no explica el fenómeno que se está estudiando. El investigador, en última instancia, es el que da sentido (ubica en el contexto) a los resultados obtenidos por medio de la técnica aplicada. El análisis y conclusiones a las que llegue, se fundamentan principalmente en el grado (nivel de conocimiento o manejo) que tiene sobre sus materiales.

6.7. Guía para el análisis de un ACP centrado-normado

1. Cuál es la dimensión del espacio de representación (rango de la matriz de correlaciones)
2. ¿Realiza análisis normado o no normado? ¿Por qué?
3. ¿Cuántos ejes selecciona para el análisis? ¿Por qué?
4. ¿Cuál es la variable que más contribuye al primer eje? ¿Cuál la que menos? (cómo obtuvo esa información).
5. ¿Cuáles son las variables más correlacionadas? (cómo obtuvo la información).
6. ¿Cuál es la variable mejor representada en el primer plano factorial? ¿Cuál la peor? (cómo obtuvo la información).

7. ¿Qué representa el primer eje? ¿Qué nombre le asignaría?
   ¿Qué representa el segundo eje? ¿Qué nombre le asignaría?
8. ¿Cuál es el individuo mejor representado en el primer plano factorial? ¿Por qué? Ubique sobre el gráfico de individuos al peor representado sobre el primer plano factorial. ¿Cómo lo hizo?
9. ¿Qué características tienen los departamentos según sus ubicaciones en el plano? (a la derecha, a la izquierda, arriba, abajo).
10. Supongamos que usted desea visitar un departamento del Caribe colombiano con buenas características socioeconómicas. De dos departamentos, ¿cuáles visitaría? ¿Por qué? ¿Cuáles son las características de los dos departamentos?
11. Seleccione dos departamentos que definitivamente no visitaría ¿Por qué? ¿Qué características tienen?
12. Haga un pequeño resumen del análisis.

ENTREGA COMO TRABAJO DE INVESTIGACIÓN (Reporte, Artículo, etc)

Produzca un documento respondiendo a las 12 preguntas, copiando los elementos necesarios de los archivos de salida.

Reporte escrito (Orden = como un artículo de carácter técnico), así:

a. Nombre (título); b. Autor(es); c. Palabras o frases clave; d. *Abstract*: (1-10 líneas máximo); e. Introducción; f. Desarrollo del tema; g. Conclusiones; Bibliografía en estricto orden alfabético, así: Autor(es), fecha, nombre del texto y/o artículo(s), editorial.

El trabajo debe ser lo más conciso, preciso y hacer un buen manejo del papel; pueden utilizar el respaldo. Reciclar (ECOLOGÍA).

7. ANÁLISIS DE CORRESPONDENCIAS SIMPLES (ACS)

El objeto es el análisis de *tablas de frecuencias*, resultado del volumen creciente de datos (presencia-ausencia, conteos, porcentajes, tablas de contingencia, valoración, similitud, afinidad, confusión, etc.) de fenómenos pecuarios para extraer conocimiento y que sirvan de apoyo a la toma de decisiones.

Para analizar *tablas de frecuencias,* el método descriptivo multivariado más útil es el Análisis de Correspondencias Simples (ACS); (Fisher, 1940; Williams, 1952; Tenenhaus & Young, 1985; Escofier & Pagès, 1988-1998; Lebart *et al.*, 1995; Cabarcas & Pardo, 2001; Dray & Chessel, 2003; Vertel & Pardo, 2010); la técnica es frecuentemente utilizada en ecología para la distribución de especies de flora y fauna (Birks & Austin, 1994). En Avilez *et al.* (2010), utilizan el ACS para realizar una caracterización productiva de explotaciones lecheras.

Los requisitos que deben cumplir las *tablas de datos* que se analizan bajo este método, son:

a) Los datos que contienen las tablas deben ser todos positivos.
b) Las magnitudes en la tabla deben ser del mismo orden.
c) Tanto las filas como las columnas de la tabla deben ser susceptibles de ser sumadas.

En el ACS se busca la mejor representación simultánea de dos conjuntos, constituidos por filas y columnas de una *tabla de frecuencias (T)* a través de una reducción de dimensión que permita aislar el ruido para examinar las relaciones existentes entre las variables (Fine, 1996; Fernández, 2002; Vertel & Pardo, 2010). También, para visualizar las asociaciones de las categorías fila y columna. El ACS se puede ver como la aplicación simultánea de dos Análisis en Componentes Principales (ACP). En el ACS se pueden utilizar variables suplementarias para analizar objetivos preestablecidos, al igual que en un ACP sobre los ejes factoriales se pueden proyectar filas y columnas que no hayan participado en el análisis.

A partir de la T se obtiene la tabla de frecuencias relativas notada por **F**; las notaciones del ACS se derivan de esta última tabla (Pardo, 2009).

La M-distancia al cuadrado entre dos filas i y l, y la D-distancia al cuadrado entre las columnas k y q de X son:

$$d^2(i,l) = \sum_{k=1}^{K} \frac{1}{f_{.k}} \left(\frac{f_{ik}}{f_{i.}} - \frac{f_{lk}}{f_{l.}} \right)^2 ; d^2(k,q) = \sum_{i=1}^{I} \frac{1}{f_{.k}} \left(\frac{f_{ik}}{f_{j.}} - \frac{f_{iq}}{f_{.q}} \right)^2 \quad (30)$$

Esta distancia tiene dos propiedades muy importantes para la interpretación de las salidas del ACS:

1) Equivalencia distribucional

El ACS no se modifica si se unen dos puntos que tienen el mismo perfil, el peso del punto colapsado es la suma de los pesos de los puntos que se unen. Esto permite unir filas o columnas con perfiles parecidos, para simplificar las tablas originales. Esta propiedad hace que el ACS sea robusto ante la "arbitrariedad" en la conformación de las categorías de una variable en un estudio.

2) Representación simultánea

En el ACS las relaciones de transición son:

$$F_s(i) = \frac{1}{\lambda_s} \sum_{j=1}^{J} \frac{f_{ij}}{f_{i.}} G_s(j) \quad (31)$$

$$G_s(i) = \frac{1}{\lambda_s} \sum_{i=1}^{I} \frac{f_{ij}}{f_{\cdot j}} F_s(i) \quad (32)$$

Estas expresiones muestran relaciones que hacen posible técnicamente la representación simultánea y permiten su interpretación. Un punto fila se ubica en el promedio ponderado por los valores de su perfil, de las coordenadas de todos los puntos columna, dilatado por el inverso de la norma del vector propio.

***Ejercicio:* Tabla de datos para conteos (frecuencias absolutas)**

Causas que afectan las pérdidas en un hato ganadero del sistema de producción doble propósito (Ejercicio 2). *A- Accidentes; B- Desconocidas; C- Enfermedades infecciosas, gastrointestinales y parasitarias; D- Deficiencias nutricionales; E- Problemas genéticos*

Tabla de contingencia (Tabla de datos)

	A	**B**	**C**	**D**	**E**	***Marginal filas***
Finca 1	15	54	231	149	0	449
Finca 2	30	29	126	51	1	237
Finca 3	12	51	533	125	0	721
Marginal columnas	57	134	890	325	1	1407

El análisis de este estudio está basado en el trabajo de Martínez y Brandi (1997) "Factores que afectan las pérdidas en un rebaño doble propósito" y está orientado por la siguiente pregunta:

1. La distribución de *causas que afectan las pérdidas* está asociada a la *finca*.
2. ¿Qué fincas globalmente se parecen?

7.1. Análisis de la tabla T: Tabla de frecuencias relativas

La tabla a analizar se nota [**T**], es de distribución 3x5 y de término general t_{ij}

$t_{12} = 54$, significa que la *finca 1* presenta 54 animales de un hato ganadero del sistema de producción doble propósito que mueren por causas desconocidas (D).

La tabla de frecuencias relativas asociada a la tabla [**T**] se nota [**F**] y su término general es $f_{ij} = \frac{t_{ij}}{n}; n = \sum_j t_{ij}$.

Tabla 21. Frecuencias relativas

Notación

$$\boldsymbol{F} = \begin{pmatrix} f_{11} & f_{12} & f_{13} & f_{14} & f_{15} \\ f_{21} & f_{22} & f_{23} & f_{24} & f_{25} \\ f_{31} & f_{32} & f_{33} & f_{34} & f_{35} \end{pmatrix}$$

Frecuencias relativas

Valores	*A*	*B*	*C*	*D*	*E*	*Marginal filas*
Finca 1	1,07	3,84	16,42	10,59	0,0	31,91
Finca 2	2,13	2,06	8,96	3,62	0,07	16,84
Finca 3	0,85	3,62	37,88	8,88	0,00	51,24
Marginal columnas	4,05	9,52	63,26	23,10	0,07	100,00

Fuente: Elaboración propia

El total de la tabla suma 100 %, al interior de la tabla se tiene la distribución de frecuencias conjunta entre las dos variables (*fincas* y *causas de pérdidas*).

Por ejemplo, el 37,88 % del total de los animales pertenecía a la *finca 3* y murió por causa de *enfermedades infecciosas, gastrointestinales o parasitarias (C)*; el 10,59 % *por causa de deficiencias nutricionales* (D) y pertenecía a la *finca 1*.

7.2. Marginales filas y columnas

$$f_{i.}: marginales\ fila \quad f_{i.} = \sum_{j=1}^{J} f_i j \qquad f_{.j}: marginales\ columna \quad f_{.j} = \sum_{i=1}^{I} f i_j$$

La última columna de la Tabla 21 es la distribución marginal filas de la variable *fincas* (Tabla 22): de los animales del hato ganadero se muestra que el 31,91 % corresponde a animales muertos de la finca 1, el 16,84 % de la finca 2 y el 51,24 % de la finca 3.

Ejemplo: $f_{2.} = \sum_{j=1}^{5} f_{2j} \rightarrow$ La marginal de la fila 2 (Finca 2)

$$f_{2.} = f_{21} + f_{22} + f_{23} + f_{24} + f_{25}$$
$$f_{2.} = 0{,}0213 + 0{,}0206 + 0{,}0896 + 0{,}0362 + 0.0007 = 0.1684$$

Interpretación: De los animales muertos del hato ganadero, el 16,84 % pertenecía a la finca 2.

Tabla 22. Notación y marginales filas de la aplicación en estudio

Notación

$$\boldsymbol{D_I} = \begin{pmatrix} f_{1.} & 0 & 0 \\ 0 & f_{2.} & 0 \\ 0 & 0 & f_{3.} \end{pmatrix}$$

Valores

$$\boldsymbol{D_I} = \begin{pmatrix} 31{,}91 & 0 & 0 \\ 0 & 16{,}84 & 0 \\ 0 & 0 & 51{,}24 \end{pmatrix}$$

Fuente: Elaboración propia

La última fila de la Tabla 21 es la distribución marginal columnas *causas de pérdidas* (Tabla 23): 4,05 % de los animales murió por *causas de accidentes*

(A); 9,52 % por *causas desconocidas (B)*; 63,26 % por causas *infecciosas, gastrointestinales y parasitarias (C);* 23,10 % por *deficiencias nutricionales (D) y* 0,07 % a *causa de problemas genéticos.*

Tabla 23. Notación y marginales columnas de la aplicación en estudio

Notación

$$\boldsymbol{D_J} = \begin{pmatrix} f_{.1} & 0 & 0 & 0 & 0 \\ 0 & f_{.2} & 0 & 0 & 0 \\ 0 & 0 & f_{.3} & 0 & 0 \\ 0 & 0 & 0 & f_{.4} & 0 \\ 0 & 0 & 0 & 0 & f_{.5} \end{pmatrix}$$

Valores

$$\boldsymbol{D_J} = \begin{pmatrix} 4{,}05 & 0 & 0 & 0 & 0 \\ 0 & 9{,}52 & 0 & 0 & 0 \\ 0 & 0 & 63{,}26 & 0 & 0 \\ 0 & 0 & 0 & 23{,}10 & 0 \\ 0 & 0 & 0 & 0 & 0.07 \end{pmatrix}$$

Fuente: Elaboración propia

Ejemplo: $f_{.3} = \sum_{i=1}^{5} f_{i3} \rightarrow$ La marginal de la columna 3 (*C-Enfermedades infecciosas*)

$$f_{.3} = f_{13} + f_{23} + f_{33}$$
$$f_{.3} = 0{,}1642 + 0{,}0896 + 0{,}3788 = 0{,}6327$$

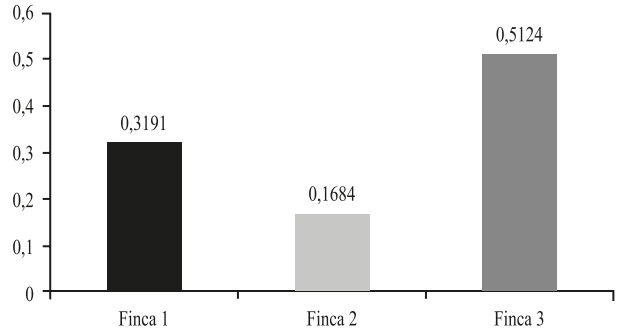
Interpretación: De los animales del hato ganadero escogidos en el estudio, el 63,26 % murió por *enfermedades infecciosas, gastrointestinales y parasitarias (C)*.

Se definen matrices diagonales: $D_I = diag(f_{i.})$ y $D_J = diag(f_{.j})$ que van a corresponder a los pesos de las filas (Figura 40) y a los pesos de las columnas respectivamente (Figura 41), en el análisis de correspondencias simples como un análisis en componentes principales ponderado.

En nuestro ejemplo:
$D_I = diag(0.3191,0.1684,0.5124)$
$D_J = diag(0.0405,0.0952,0.6326,0.2310,0.0007)$

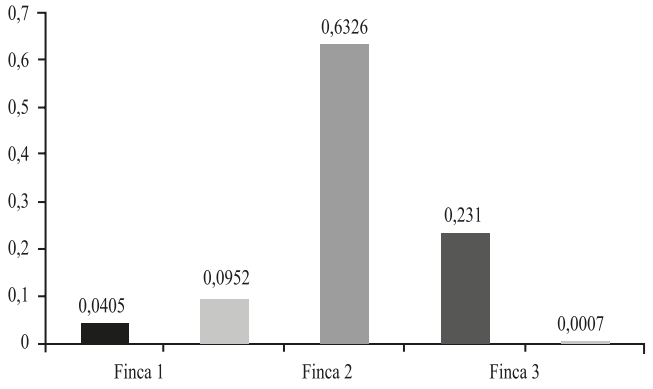
En resumen, sobre gráficos de representación simultánea de filas y columnas, la posición relativa de los puntos de un mismo conjunto (filas o columnas) se interpreta en términos de distancia, mientras que la posición de un punto de un conjunto y todos los puntos del otro conjunto se interpreta en términos de baricentro.

La proximidad de una fila y de una columna no tiene sentido en sí mismo.



Individuos	Finca 1	Finca 2	Finca 3
fi.	0,3191	0,1684	0,5124

Figura 40. Histograma de los pesos de las filas
Fuente: Elaboración propia



Columnas	A	B	C	D	E
f.j	0,0405	0,0952	0,6326	0,231	0,0007

A- Accidentes; B- Desconocidas; C- Enfermedades infecciosas, gastrointestinales y parasitarias; D- Deficiencias nutricionales; E- Problemas genéticos
Figura 41. Histograma de los pesos de las columnas
Fuente: Elaboración propia

7.3. Tablas de perfiles fila y columna

La lectura interesante de la información contenida en una tabla de contingencia es la comparación entre filas y entre columnas. En la tabla de frecuencias relativas las filas y las columnas están influenciadas por el peso relativo de

sus marginales. La comparación se facilita obteniendo las distribuciones condicionales o perfiles de cada una de las filas y de cada una de las columnas.

Para obtener la distribución condicional de la fila i, se dividen todas las celdas de esa fila por el valor total de la fila. De manera análoga, se obtienen las condicionales de las columnas. Se llega entonces a dos tablas: una de perfiles fila y otra de perfiles columna.

A partir de la Tabla 21 se obtiene la Tabla 22, de perfiles fila: por ejemplo para la fila 2 (finca 2): 2,13/16,84 = 0,1266, 2,06/16,84 = 0,1224, 8,96/16,84 = 0,5318, 3,62/16,84 = 0,2152 y 0,07/16,84 = 0,0042 expresados en porcentaje (100 %): 12,66, 12,24, 53,18, 21,52 y 0,42 %.

Un perfil fila *i* se nota: $\left\{\frac{f_{ij}}{f_{i.}}, j = 1, \dots, J\right\}$ matricialmente: $\boldsymbol{D_I^{-1}F}$

Para cada *causa de pérdidas* (columna) se tiene una distribución de frecuencia entre las tres fincas (filas), que se denomina distribución condicional o perfil fila (Tabla 24).

Tabla 24. Notación y valores de los perfiles fila para la aplicación en estudio

Notación

$$\boldsymbol{D^{-1}F} = \begin{pmatrix} \frac{f_{11}}{f_{1.}} & \frac{f_{12}}{f_{1.}} & \frac{f_{13}}{f_{1.}} & \frac{f_{14}}{f_{1.}} & \frac{f_{15}}{f_{1.}} \\ \frac{f_{21}}{f_{2.}} & \frac{f_{22}}{f_{2.}} & \frac{f_{23}}{f_{2.}} & \frac{f_{24}}{f_{2.}} & \frac{f_{25}}{f_{2.}} \\ \frac{f_{31}}{f_{3.}} & \frac{f_{32}}{f_{3.}} & \frac{f_{33}}{f_{3.}} & \frac{f_{34}}{f_{3.}} & \frac{f_{35}}{f_{3.}} \end{pmatrix}$$

Valores

Valores	*A*	*B*	*C*	*D*	*E*	*Suma*
Finca 1	3,34	12,03	51,45	33,19	0,00	100
Finca 2	12,66	12,24	53,18	21,52	0,42	100
Finca 3	1,66	7,07	73,93	17,34	0,00	100

Fuente: Elaboración propia

De la Tabla 24 y la Figura 42 se pueden comparar fácilmente los perfiles fila: en la *finca 1* se dan más pérdidas por causas de *enfermedades infecciosas, gastrointestinales y parasitarias* (51,45 %), por *deficiencias nutricionales* (33,19 %), por causas *desconocidas* (12,03 %), y luego por *accidentes* (3,34 %), finalmente por *problemas genéticos* no tuvo pérdidas (0 %).

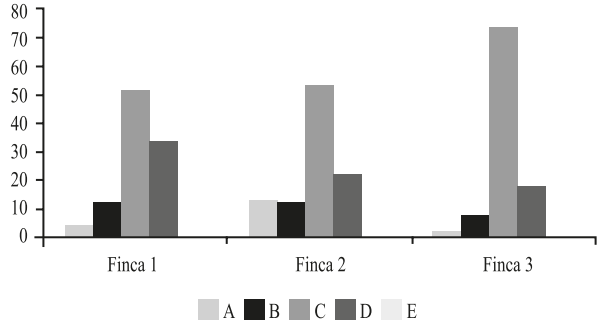


Figura 42. Perfiles fila. *Fincas vs. causas de pérdidas*
Fuente: Elaboración propia

Las pérdidas en la *finca 3* ocurren por causas de *enfermedades infecciosas, gastrointestinales y parasitarias* (73,93 %). Los perfiles fila de las *fincas 1* y *3* son los más parecidos en su forma. En ambos las causas de pérdidas se ordenan según frecuencia así: *A- Accidentes; B- Desconocidas; C- Enfermedades infecciosas, gastrointestinales y parasitarias; D- Deficiencias nutricionales; E- Problemas genéticos.*

La Tabla 25 contiene los perfiles columna expresados en porcentaje, calculados a partir de la Tabla 21, dividiendo la celda en cada columna por la marginal, por ejemplo para la columna 3 (*C- Enfermedades infecciosas, gastrointestinales y parasitarias*):

0,1642/0,6326 = 0,2595 = 25,95 %
0,0896/0,6326 = 0,1416 = 36,51 %
0,3788/0,6326 = 0,5988 = 59,88 %

Un perfil columna *i* se nota: $\left\{\frac{f_i j}{f_{\cdot j}}, i = 1, \ldots, I\right\}$ matricialmente: $\boldsymbol{F'D_J^{-1}}$

Cada finca tiene su distribución de *causas de pérdidas* en las frecuencias (condicionales o perfiles de columna Tabla 25 y Figura 43).

A partir de la Tabla 25, se pueden comparar los cinco perfiles columna: las causas de pérdidas *B* y *D* son muy parecidas en su distribución en todas las fincas. Por problemas genéticos (*E*) se presentan solamente en la finca 2.

Tabla 25. Notación y valores de los perfiles columna para la aplicación en estudio

Notación

$$\boldsymbol{F'D_J^{-1}} = \begin{pmatrix} \frac{f_{11}}{f_{\cdot 1}} & \frac{f_{12}}{f_{\cdot 2}} & \frac{f_{13}}{f_{\cdot 3}} & \frac{f_{14}}{f_{\cdot 4}} & \frac{f_{15}}{f_{\cdot 5}} \\ \frac{f_{21}}{f_{\cdot 1}} & \frac{f_{22}}{f_{\cdot 2}} & \frac{f_{23}}{f_{\cdot 3}} & \frac{f_{24}}{f_{\cdot 4}} & \frac{f_{25}}{f_{\cdot 5}} \\ \frac{f_{31}}{f_{\cdot 1}} & \frac{f_{32}}{f_{\cdot 2}} & \frac{f_{33}}{f_{\cdot 3}} & \frac{f_{34}}{f_{\cdot 4}} & \frac{f_{35}}{f_{\cdot 5}} \end{pmatrix}$$

Fuente: Elaboración propia

Tabla 26. Perfiles columna

Valores

Valores	*A*	*B*	*C*	*D*	*E*
Finca 1	26,32	40,31	25,95	45,84	0,00
Finca 2	52,65	21,64	14,16	15,69	100,00
Finca 3	21,06	38,07	59,88	38,46	0,00
Suma	100,00	100,00	100,00	100,00	100,00

Fuente: Elaboración propia

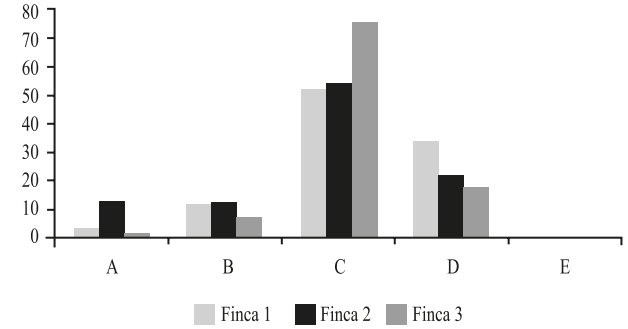
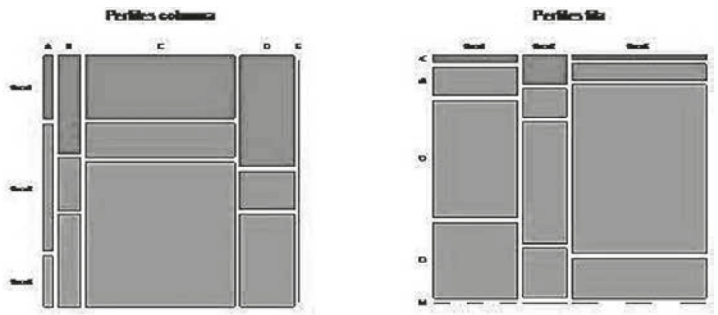


Figura 43. Perfiles columna. *Causas de pérdidas* vs. *fincas*
Fuente: Elaboración propia

De los perfiles fila y columna en conjunto se puede concluir principalmente que hay una correspondencia fuerte entre la finca 3 y las pérdidas por causas de *enfermedades infecciosas, gastrointestinales y parasitarias (C)*. También, se puede observar una correspondencia entre las causas por *problemas genéticos* (*E*) y la finca 2 (Figura 44).



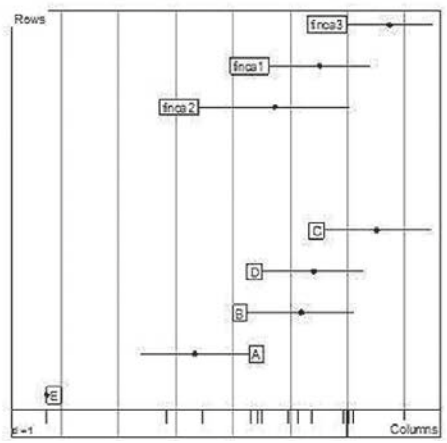


Figura 44. Gráficas de los perfiles fila y columna en conjunto
Fuente: Elaboración propia

En el ACS se busca una representación más adecuada para analizar simultáneamente perfiles fila y columna obtenidas a partir de una Tabla de Contingencia (TC). Cuando se tiene TC de gran tamaño es muy difícil obtener una síntesis apropiada de forma como se hizo en el ejemplo.

Para el ACS se parte de la representación de los perfiles fila en un espacio multidimensional, donde las columnas son los ejes y simétricamente de otra nube de perfiles columna, donde las líneas son los ejes. Para ello se requiere del uso de una distancia apropiada: la distancia ji-cuadrado entre distribuciones (Cabarcas & Pardo, 2001).

Mientras en ACP, las filas y las columnas objetos de naturaleza bien diferente (individuos y variables), las filas y las columnas de un ACS de una tabla de frecuencias son de la misma naturaleza, a saber, clases de individuos. Según este simple punto de vista, no es nada escandaloso ver aparecer todas estas clases sobre un mismo gráfico.

7.4. El análisis de correspondencias simples, ACM (T) como un ACP (X, M, D)

El ACS (Figura 45) de la tabla **T** (Tabla 3) se obtiene mediante el ***ACP (X, M, D)*** de la tabla de datos ***X*** cuyo término general está dado por $p_{ij} = \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}}$ usando $\boldsymbol{D} = \boldsymbol{D_I} = diag(fi.)$ y $\boldsymbol{M} = \boldsymbol{DJ} = diag(f.j)$ (Greenacre, 2007; Vertel & Pardo, 2010).

La matriz de frecuencias estandarizadas, tiene término general:

$$p_{ij} = \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}} \qquad (33)$$

Para: $p_{12} = \frac{f_{12} - f_{1.}f_{.2}}{f_{1.}f_{.2}} = \frac{0.0384 - (0.3191x0.0952)}{(0.3191x0.0952)} = 0{,}2628$

La tabla de frecuencias estandarizadas *x* es:

```
> acs$tab
                A          B          C          D         E
Finca1 -0.1753605  0.2628062 -0.1866645  0.43664896 -1.000000
Finca2  2.1245836  0.2848101 -0.1595221 -0.06839338  4.936709
Finca3 -0.5891671 -0.2572816  0.1686811 -0.24943988 -1.000000
```

Todas las fórmulas del ACS se pueden derivar de las fórmulas correspondientes al ***ACP (X, M, D)*** (ver Tabla 9).

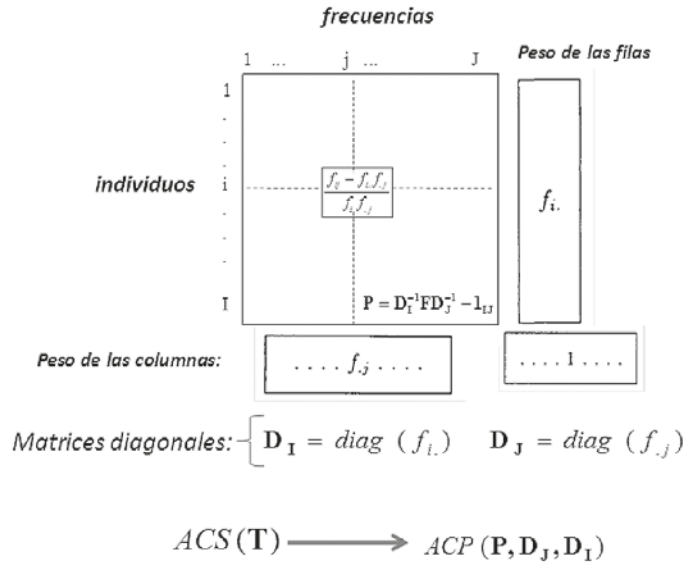


Figura 45. Análisis de correspondencias simples de una tabla de frecuencias T
Fuente: Elaboración propia

Para poder hacer la nube de variables (frecuencias) se buscan los ejes de proyección que conservan lo más posible las normas de las variables originales. La solución se consigue con los valores y vectores propios de las matrices $\boldsymbol{P^t D_I P D_J}$, esta matriz tiene valores propios que son iguales a los valores propios de $\boldsymbol{P D_J P^t D_I}$ y los restantes *I - S* valores propios son iguales a cero; cada valor propio μ_s tiene asociado el vector propio v_s.

En cada ACP ponderado del ACS, los planos factoriales y sus ayudas a la interpretación son análogas a los de los individuos en el ACP clásico (Ver Tabla 9).

Las dos nubes de puntos están en un subespacio de dimensión 2 porque:

min (I,J) – 1 = min(3,5) –1 = 3 – 1 = 2

Las *I* filas de la Tabla *T* conforman la nube de puntos N_I en $\mathbb{R}^J$ y las *J* columnas conforman la nube de puntos N_J en $\mathbb{R}^I$. Las métricas son: $M = D_J = diag(f.j)$ y $D = D_I = diag(fi.)$.

Las coordenadas son: las filas de la Tabla **X** y las columnas de la Tabla **X**.

La inercia es la traza *tr(X' DXM)* o *tr(XMX' D)*

En nuestro ejemplo:

Inercia (ACS) es:

$$\underbrace{tr(\boldsymbol{X}' \boldsymbol{D_I X D_J})}_{\text{JxI IxI IxJ JxJ}} \quad o \quad \underbrace{tr(\boldsymbol{X D_J X' D_I})}_{\text{IxI}}$$

JxJ

La inercia total del ACS es:

$$Inercia\ (ACS) = \sum_{s=1}^{S=\min(I,J)-1=2} \mu_s = \mu_1 + \mu_2 = 0.08911435$$

El significado de gráficas a la nube de puntos N_I sobre el nuevo sistema de ejes, es decir, sobre la base de los vectores propios normados: $\{\mu_1 \ldots \mu_J\}$, es hacer una rotación del sistema de ejes. Como los propios están ordenados por la cantidad de inercia que recogen, las proyecciones sobre los primeros ejes son las "mejores", en el sentido de conservar la mayor inercia posible. El primer plano factorial conformado por los ejes 1 y 2, generados por μ_1 y μ_2, respectivamente, retiene una inercia igual a $\mu_1 + \mu_2$.

Una primera decisión que debe tomar el analista, usuario del ACP ponderado, es determinar el número de ejes *S* que va a retener para la interpretación (Tabla 27).

Tabla 27. Valores propios de la aplicación Causas de pérdidas en ganado bovino

Eje	μ_s	$\sum_{s=1}^{t=1,\ldots S} \mu_s$	$\mu_s\ \phi^2$
1	59,5	59,5	66,8 %
2	29,6	89,1	100,0 %
	Suma = 89,1	-	-

Fuente: Elaboración propia

La lectura de la tabla y de la gráfica permite tomar la decisión de cuántos ejes analizar. En este ejemplo los dos primeros valores propios retienen 100 % de la inercia, no se pierde información al leer el primer plano factorial, con esto la lectura se hace más fácil.

Las coordenadas factoriales filas para los dos primeros ejes se buscan con: $FS = XM\mu_s$ al reemplazar para el ejemplo $F_S = XD_j^{*}{}^{\mu_s}\Big/\sqrt{f_{\cdot j}}$

Los resultados son:

```
        Eje 1   Eje 2
finca1 -0.1510  0.2277
finca2 -0.3963 -0.2608
finca3  0.2243 -0.0561
```

Con las coordenadas factoriales (filas, columnas) se realiza el primer plano o mapa factorial del ACS de frecuencias (Figura 46).

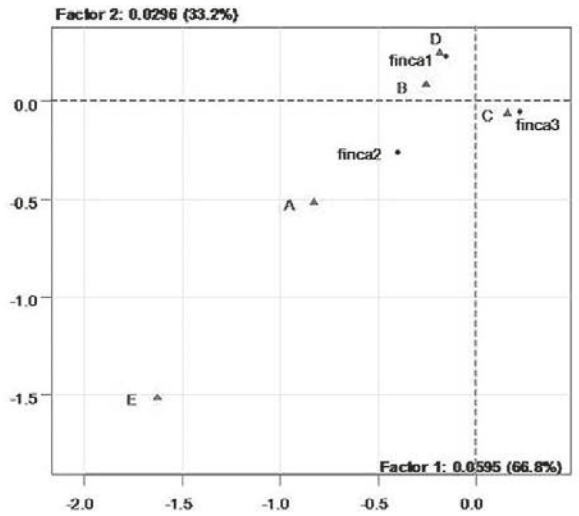


Figura 46. Plano factorial 1-2 Fincas vs. causas de pérdidas
Fuente: Elaboración propia

Para las ayudas de interpretación

Contribución absoluta del punto *i* en el eje *s*

La contribución de un individuo es: $Con_s(i) = \frac{p_i F_s^2(i)}{\lambda_s}$; λ_s es el valor propio en el eje

$F_x^2(i)$ es el valor de la coordenada al cuadrado en el eje s, p_i es el peso de las filas.

Por ejemplo para el individuo 2 en el eje 1.

$$Con_1(2) = \frac{p_2 x F_1^2(2)}{\lambda_1} = \frac{(0{,}1684)(-0{,}3963)^2}{0{,}0595} = 0{,}4445$$

Contribuciones:

```
Row Contributions
       Eje 1 Eje 2
finca1  12.2  55.9
finca2  44.5  38.7
finca3  43.3   5.4
```

Multiplicado el resultado por 100 %

Las fincas que más contribuyen a la formación del primer eje son la *finca 2* y la *finca 3* (con el 87,8 % de la inercia del primer eje). El primer eje se interpreta como la contraposición de la *finca 2* y la *finca 3*. El segundo eje se interpreta como la contraposición entre la *finca 1* y la *finca 2* (contribuyen con el 94,6 % de la inercia del segundo eje).

Calidad de la representación

Sobre el eje *s* el coseno al cuadrado de un punto-fila *i* es:

$$Cos_s^2(i) = \frac{F_{(s)}^2(i)}{\| i \|^2} \qquad (34)$$

$$\| i \|^2 = d^2(i,o) = \sum_{j=1}^{J} f_{\cdot j}\, x_{ij}^2 \qquad (35)$$

Donde: f_j es la marginal columna j y x_{ij} es la frecuencia estandarizada en la fila i y columna j.

d^2 *(i,o)*: es el cuadrado de la distancia al centro de gravedad, el cual coincide con el origen de la representación.

Para el individuo 1 en el eje 1, d^2 *(i,o)* es

$$d^2(i,o) = (-0{,}1753^2\ x\ 0{,}0405) + (0.2628^2 x\ 0.0952) + (-0.1867^2\ x 0.6326\ ) + (0.4367^2\ x\ 0.2310) + (-1.0000^2\ x\ 0.0007)$$

$$d^2(i,o) = 0.0012 + 0.0065 + 0.0221 + 0.044 + 0.0007$$

$$d^2(i,o) = 0{,}0745$$

La calidad de representación del individuo 1 en el eje 1 es:

$$Cos_1^2(1) = \frac{F_1^2(i)}{\| 1 \|^2} = \frac{(-0{,}1510)^2}{0{,}0745} = 30.5\%$$

Las calidades de representación son:

```
Representation Quality of the Rows
       Axis1 Axis2 con.tra
finca1 -30.5  69.5    26.7
finca2 -69.8 -30.2    42.5
finca3  94.1  -5.9    30.7
```

Los signos (negativos o positivos) muestran en qué lado del eje se encuentran.

El valor del coseno al cuadrado coincide con la relación de contribuciones del individuo *i* a la inercia: contribución a la inercia proyectada sobre el eje *s*/ contribución a la inercia total, y se llama también **Contribución relativa**.

El coseno al cuadrado también se define para un punto como el cuadrado de la relación entre la norma de la proyección sobre la norma en el espacio completo.

d^2 *(i,o)*: Distancia de un punto al origen, en el espacio completo, es un buen complemento en la lectura de los ejes factoriales, está dada por la norma el vector-individuo.

La *M*-distancia al cuadrado entre dos filas *i* y *l* es:

$$d^2(i,\ell) = \sum_{j=1}^{J} \frac{1}{f.j}\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{\ell j}}{f_{\ell .}}\right)^2 \qquad (36)$$

Para la *finca 1* (1) y la *finca 2* (2) es:

$$d^2(1,2) = \frac{1}{0.0405}\left(\frac{0.0107}{0.3191} - \frac{0.0213}{0.1684}\right)^2 + \frac{1}{0.0952}\left(\frac{0.0384}{0.3191} - \frac{0.0206}{0.1684}\right)^2 + \frac{1}{0.6326}\left(\frac{0.1642}{0.3191} - \frac{0.0896}{0.1684}\right)^2 + \frac{1}{0.2310}\left(\frac{0.1059}{0.3191} - \frac{0.0362}{0.1684}\right)^2 + \frac{1}{0.0007}\left(\frac{0.0000}{0.3191} - \frac{0.0007}{0.1684}\right)^2$$

$$d^2(1,2) = 0.5589$$

Para las columnas (Causas de pérdidas) se definen los mismos indicadores.

En resumen:

Filas (Fincas)	**Coordenadas – Factoriales**		**Contribución**		**Calidad de represen-tación**		**Remain**
	Eje 1	*Eje 2*	*Eje 1*	*Eje 2*	*Eje 1*	*Eje 2*	Resto
Finca 1	-0,1510	0,2277	12,2	55,9	30,5	69,5	0
Finca 2	-0,3963	-0,2608	44,5	38,7	69,8	30,2	0
Finca 3	0,2243	-0,0561	43,3	5,4	94,1	5,9	0

Columnas (Causas de pérdidas)	**Coordenadas – Factoriales**		**Contribución**		**Calidad de representación**		**Remain**
	Eje 1	*Eje 2*	*Eje 1*	*Eje 2*	*Eje 1*	*Eje 2*	Resto
A	-0,8244	-0,5181	46,3	36,7	71,7	28,3	0
B	-0,2511	0,0812	10,1	2,1	90,5	9,5	0
C	0,1600	-0,0663	27,2	9,4	85,4	14,6	0
D	-0,1850	0,2435	13,3	46,2	36,6	63,4	0
E	-1,6247	-1,5156	3,1	5,5	53,5	46,5	0

Los dos primeros valores propios retienen el 100 % de la inercia total (Eje 1 = 66,8 %, Eje 2 = 33,2 %). En el lado positivo del primer eje (Figura 46) se ubica la *finca 2* que presenta pérdidas por causa de *accidentes* (A) en mayor proporción, opuesta a la *finca 3* que presenta pérdidas por causa de *enfermedades infecciosas, gastrointestinales y parasitarias* (C). En el segundo eje se contraponen la *finca 1* y la *finca 2* (contribuyen con un 94,6 % en la formación del eje), la *finca 1* presenta pérdidas por causa de *deficiencias nutricionales* (A) y la finca 2 por *accidentes* (A).

7.5. Guía para el análisis

1. ¿Qué porcentaje de la inercia total es recogida por el primer eje factorial, por el segundo eje factorial y por el primer plano factorial? ¿Qué se puede concluir de esta constatación?
2. Análisis de la proyección de la nube de puntos-categorías sobre el primer plano factorial.
   a. ¿Cuáles son las dos categorías más contributivas al primer eje factorial? ¿Cuáles son sus coordenadas y cuáles sus pesos relativos? ¿Qué tan bien representadas están esas categorías en el primer plano factorial? ¿Cuál es la categoría que está más mal representada en el primer plano factorial? ¿Puede decirse que está muy mal representada? ¿Cómo podría denominarse al primer eje factorial?
   b. ¿Cuáles son las dos categorías más contributivas al segundo eje factorial? ¿Cuáles son sus coordenadas y sus pesos relativos? ¿Qué tan bien representadas están esas categorías en el primer plano factorial? ¿Cómo podría denominarse al segundo eje factorial?
3. Análisis de la proyección de la nube de puntos-individuos sobre el primer plano factorial.
   a. ¿Cuáles son las dos filas-individuos más contributivos al primer eje factorial? ¿Cuáles son sus coordenadas y cuáles sus pesos relativos? ¿Qué tan bien representadas están estas filas-individuos en el primer plano factorial? ¿Cuáles son las dos filas-individuos más mal representadas en el primer plano factorial? ¿Puede concluirse de lo anterior que algunas filas-individuos están muy mal representadas? ¿De qué manera estos resultados son útiles para ayudar a la caracterización del primer eje factorial?
   b. ¿Cuáles son las dos filas-individuos más contributivos al segundo eje factorial? ¿Cuáles son sus coordenadas y cuáles sus pesos relativos? ¿Qué tan bien representadas están estas filas-individuos en el primer plano factorial? ¿De qué manera estos resultados son útiles para ayudar a la caracterización del segundo eje factorial?
4. Análisis de la proyección conjunta de las dos nubes de puntos
   a. ¿Puede evidenciarse algún patrón de comportamiento con respecto a los perfiles de las fincas 1 y 3? Teniendo en cuenta las proyecciones de las categorías, ¿cómo se puede caracterizar dicho patrón?
   b. Liste las fincas más cercanas a cada una de las categorías. ¿Se puede evidenciar algún patrón especial en estos grupos?
5. Escriba en un párrafo las conclusiones más relevantes del análisis.

ENTREGA COMO TRABAJO DE INVESTIGACIÓN (Reporte, Artículo, etc.)

Produzca un documento respondiendo a las 12 preguntas, copiando los elementos necesarios de los archivos de salida.

Reporte escrito (Orden = como un artículo de carácter técnico) así:
a. Nombre (título); b. Autor(es); c. Palabras o frases clave; d. *Abstract*: (1-10 líneas máximo); e. Introducción; f. Desarrollo del tema; g. Conclusiones; Bibliografía en estricto orden alfabético, así: Autor(es), fecha, nombre del texto y/o artículo(s), editorial.

El trabajo debe ser lo más conciso, preciso y con un buen manejo del papel, pueden utilizar el respaldo. Reciclar (ECOLOGÍA).

8. ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES (ACM)

Este análisis se extiende al caso de más de dos variables cualitativas utilizando codificaciones especiales de los datos, que le otorgan propiedades específicas e interesantes como para merecer un tratamiento como método independiente (Cabarcas & Pardo, 2001).

8.1. Dominio de aplicación

El análisis de correspondencias múltiples se utiliza en el análisis de tablas de individuos descritos por variables categóricas. Compara individuos a través de las modalidades de las variables. Encuentra asociaciones entre variables a través de sus modalidades. Es el método apropiado para abordar el análisis multivariado de las encuestas y para explotar bases de datos con información cualitativa. El ejemplo reducido para la presentación del método es una tabla de datos de seis fincas descritas por sus respuestas correspondientes a la dinámica de ordeño, las cuales son categóricas y modalidades. Los individuos son similares a las modalidades. La asociación entre variables se presenta porque son similares los individuos que asumen las mismas modalidades de diferentes variables.

8.2. Fundamentos del método

Se parte de una tabla de *individuos* x *variables categóricas*. La tabla puede ser numérica pero los números están indicando la modalidad de la variable que asume el individuo de la fila. Sin embargo los números de la tabla no tienen significado aritmético, es decir no tiene ningún sentido sumarlos u obtener alguna estadística descriptiva. Una tabla así se suele denominar de código condensado y aquí se denota con la letra **Q,** de tamaño (*n,p*), donde *n* representa al número de individuos y *p* el número de variables.

Para mostrar los elementos del método utilizaremos la Tabla 28 (primer recuadro), donde puede leerse por ejemplo que la *Vaca 13*, viene de la **FINCA** *GA-GF*, tiene **CRUCE** *criollo*, ha tenido *dos partos* (**NP-** Número de partos), que el 2° parto lo tuvo en la **ÉPOCA** de *invierno* (INV), y que el **SEXO** de la cría del 2° parto fue *macho* (M).

Tabla 28. Resultados de encuesta de dinámica de ordeño

	FINCA	CRUCE	NP	ÉPOCA	SEXO
Vaca1	GA-GF	indicus	n1	INV	M
Vaca2	GA-GF	indicus	n2	VER	M
Vaca3	GA-GF	indicus	n2	VER	M
Vaca4	GA-GF	indicus	n3	INV	M
Vaca5	GA-GF	indicus	n4	VER	M
Vaca6	GA-GF	indicus	n3	INV	M
Vaca7	GA-GF	indi-crio	n4	VER	M
Vaca8	G07	indi-euro	n3	VER	M
Vaca9	G07	indi-euro	n4	VER	H
Vaca10	GA-GF	indi-euro	n1	INV	H
Vaca11	GA-GF	indi-euro	n3	INV	M
Vaca12	GA-GF	indi-euro	n4	INV	H
Vaca13	GA-GF	criollo	n2	INV	M
Vaca14	GA-GF	indi-crio	n2	INV	H
Vaca15	GA-GF	indi-crio	n4	INV	H
Vaca16	GA-GF	indi-crio	n2	INV	M
Vaca17	GA-GF	indi-euro	n3	INV	H
Vaca18	GA-GF	indicus	n2	INV	M
Vaca19	GA-GF	indi-euro	n3	VER	M
Vaca20	G07	indi-euro	n3	VER	M

Fuente: Elaboración propia

A partir de la tabla **Q** se pueden construir dos tablas con significado numérico: la Tabla Disyuntiva Completa (TDC) y la tabla de Burt.

8.3. Tabla Disyuntiva Completa y tabla de Burt

Una variable categórica asigna a cada individuo de una población una modalidad y divide a la población en tantos subconjuntos como modalidades tenga. Por ejemplo la finca donde está la vaca puede ser: GA-GF, G07.

Las vacas de este estudio se dividen entonces en dos grupos según su sitio de origen (FINCA). La codificación disyuntiva completa se hace recurriendo a una variable indicadora por cada modalidad, es decir una variable que toma el valor de 1 si el individuo asume la modalidad y cero si no. Por ejemplo para el sitio de ordeño se tiene:

FINCA GA-GF: 0 – No, 1 – Sí; FINCA G07: 0 – No, 1 – Sí

La variable indica a su vez la pertenencia o no a cada uno de los grupos. El nombre disyuntiva completa de esta codificación se debe a que se exige a cada individuo pertenecer a una y solo una de las modalidades, entonces aparece siempre 'uno' en un solo lugar bajo las modalidades pertenecientes a una sola variable.

La Tabla 29 es la Tabla Disyuntiva Completa (TDC) derivada de la tabla **Q** que representa a las 20 vacas de una zona descritas por aspectos relacionados a la dinámica de ordeño (Tabla 28, primer recuadro).

Tabla 29. Tabla disyuntiva completa del ejemplo

	FINCA		**CRUCE**				**NP**				**ÉPOCA**		**SEXO**	
	G07	GA-GF	Cr	I-Cr	I-Eu	I	n1	n2	n3	n4	INV	VER	H	M
Vaca1	0	1	0	0	0	1	1	0	0	0	1	0	0	1
Vaca2	0	1	0	0	0	1	0	1	0	0	0	1	0	1
Vaca3	0	1	0	0	0	1	0	1	0	0	0	1	0	1
Vaca4	0	1	0	0	0	1	0	0	1	0	1	0	0	1
Vaca5	0	1	0	0	0	1	0	0	0	1	0	1	0	1
Vaca6	0	1	0	0	0	1	0	0	1	0	1	0	0	1
Vaca7	1	0	0	1	0	0	0	0	0	1	0	1	0	1
Vaca8	1	0	0	0	1	0	0	0	1	0	0	1	0	1
Vaca9	0	1	0	0	1	0	0	0	0	1	0	1	1	0
Vaca10	0	1	0	0	1	0	1	0	0	0	1	0	1	0
Vaca11	0	1	0	0	1	0	0	0	1	0	1	0	0	1
Vaca12	0	1	0	0	1	0	0	0	0	1	1	0	1	0
Vaca13	0	1	1	0	0	0	0	1	0	0	1	0	0	1
Vaca14	0	1	0	1	0	0	0	1	0	0	1	0	1	0
Vaca15	0	1	0	1	0	0	0	0	0	1	1	0	1	0
Vaca16	0	1	0	1	0	0	0	1	0	0	1	0	0	1
Vaca17	0	1	0	0	1	0	0	0	1	0	1	0	1	0
Vaca18	0	1	0	0	0	1	0	1	0	0	1	0	0	1
Vaca19	0	1	0	0	1	0	0	0	1	0	0	1	0	1
Vaca20	1	0	0	0	1	0	0	0	1	0	0	1	0	1
Suma	3	17	1	4	8	7	2	6	7	5	12	8	6	14

Fuente: Elaboración propia

La TDC denotada por **Z** es de tamaño (*n,m*), donde *m* es el número total de modalidades. La suma de cada una de sus filas es igual a *p*, el número de variables y la suma de cada columna es el número de individuos que asume la modalidad respectiva.

En el ejemplo, la TDC es de tamaño 20 x 14. Son $p = 5$ variables, cada una con 2 o 4 modalidades. La suma de las filas es igual a 5, el número de variables. El número de individuos que asume cada una de las modalidades aparece en la parte inferior de la Tabla 29.

Sometiendo la tabla disyuntiva completa a un análisis de correspondencias simples se logran los objetivos que se persiguen en una descripción multivariada de una tabla de individuos x variables categóricas.

El análisis de correspondencias de la tabla de Burt, que es una tabla que yuxtapone todas las tablas de contingencia de las variables cruzadas de dos en dos, produce planos equivalentes para las modalidades. Para el ejemplo, la tabla de Burt (B) (Tabla 30) tiene tamaño 14 x 14.

La tabla B es simétrica y por lo tanto, es suficiente mostrar la parte triangular inferior. Está conformada por 25 subtablas. Las cinco subtablas diagonales son a su vez diagonales y contienen las frecuencias marginales de cada una de las variables. Las 10 subtablas de la parte inferior son las tablas de contingencia entre parejas de variables y las 10 subtablas de la parte superior (no se muestran), son las transpuestas de las anteriores.

En la Tabla 30 se puede leer, por ejemplo, 3 vacas del cruce criollo (Cr) están en la finca G07, 8 vacas son del cruce *Indicus-Europeo* (I-Eu) de las cuales 3 vienen de la finca G07 y el resto (5) de la finca GA-GF.

Tabla 30. Tabla de Burt (B)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1-FINCA G07	3	0												
2-FINCA GA-GF	0	17												
3-CRUCE Cr	0	1	1	0	0	0								
4-CRUCE I-Cr	0	4	0	4	0	0								
5-CRUCE I-Eu	3	5	0	0	8	0								
WSQ6-CRUCE I	0	7	0	0	0	7								
7-NP - n1	0	2	0	0	1	1	2	0	0	0				
8-NP – n2	0	6	1	2	0	3	0	6	0	0				
9-NP – n3	2	5	0	0	5	2	0	0	7	0				
10-NP – n4	1	4	0	2	2	1	0	0	0	5				
11-EPOCA INV	0	12	1	3	4	4	2	4	4	2	12	0		
12-EPOCA VER	3	5	0	1	4	3	0	2	3	3	0	8		
13-SEXO H	1	5	0	2	4	0	1	1	1	3	5	1	6	0
14-SEXO M	2	12	1	2	4	7	1	5	6	2	7	7	0	14

Fuente: Elaboración propia

8.4. Análisis de correspondencias de la TDC

El análisis de correspondencias múltiples es un análisis de correspondencias simples de la tabla Z, es decir que esta tabla se toma como una tabla de contingencia particular y se somete a las transformaciones necesarias para el análisis: construcción de una tabla de perfiles fila y otra de perfiles columna.

Las marginales de la tabla de frecuencias relativas **F**, son los pesos asociados a los puntos perfiles y las ponderaciones que juegan en la distancia ji-cuadrado.

El total de la tabla Z es *np* y por lo tanto, un elemento de la tabla F derivada es: $f_{ij} = \frac{z_{ij}}{np}$

Las sumas de las filas de **F** son todas iguales a *1/n*, que será el peso de cada individuo en la nube de perfiles fila.

La suma de una columna *j* de la tabla F es: $f_{.j} = \frac{z_{.j}}{np} = \frac{n_j}{np}$,, se introduce $n_j = z_{.j}$, pues esta notación explicita mejor el significado de la suma de las columnas de **Z**: número de individuos que asumen la modalidad *j*.

Nube de perfiles fila

Los perfiles fila que representan a los individuos del análisis son histogramas cuyas barras solo pueden tomar dos valores: cero, cuando el individuo no asume la modalidad o *1/m*, el inverso del número de variables, cuando la asume. Todos los individuos tienen asociado el mismo peso: *1/n.*

La distancia ji-cuadrado entre dos perfiles individuos se transforma en:

$$d^2(i,l) = \frac{1}{p}\sum_{j=1}^{m}\frac{n}{n_j}(z_{ij} - z_{lj})^2 \qquad (37)$$

Como los elementos de z_{ij} son 1 o 0, dos individuos están próximos si asumen más o menos las mismas modalidades. Cuando un individuo asume una modalidad de frecuencia baja aparece más alejado de los demás.

Nubes de modalidades

El perfil *j* de una modalidad se obtiene dividiendo cada elemento de la tabla **Z** de la columna *j* por el total de la columna n_j. Entonces el perfil de una modalidad toma valores 0 o $1/n_j$, o sea que en el histograma aparece cero o una barra de altura $1/n_j$.

Los histogramas de dos modalidades se pueden diferenciar tanto en la posición de las barras como en su altura. Son más altas las barras del perfil de las modalidades que son asumidas por pocos individuos. El peso de cada modalidad en el análisis es proporcional a su frecuencia n_j.

La distancia entre dos modalidades *j* y *k* es igual al porcentaje de individuos que poseen *j* pero no *k* más el porcentaje de individuos que poseen *k* pero no *j*. Es decir que esta distancia crece con el número de individuos que poseen una y solo una de las modalidades *j* o *k* y decrece con la frecuencia de cada una de estas modalidades (Escofier & Pagés, 1998, p.58; Cabarcas & Pardo, 2001).

En Lebart *et al*. (1995) aparecen demostrados los siguientes hechos y sus consecuencias prácticas:

- El número de ejes que soporta la nube de modalidades o de individuos es *m-p*, número de modalidades menos número de variables. En el ejemplo 14 modalidades menos 5 variables = 9 ejes (ver Tabla 31).
- La distancia de una modalidad al centro de gravedad es más grande si su frecuencia es más baja.
- La parte de la inercia debida a una modalidad de una variable es más grande si la modalidad tiene frecuencia más baja. Se deben evitar las modalidades de muy baja frecuencia.
- La parte de la inercia debida a una variable (suma de la inercia de sus modalidades) es función creciente del número de modalidades de la variable. Se debe equilibrar el número de las modalidades de las variables.
- La inercia total de la nube es $I = \frac{m}{p} - 1$, es decir que solo depende del número de modalidades y del número de variables y no de los datos de la tabla. La inercia total no tiene significado estadístico. En el ejemplo I = 14/5 – 1 = 1,8 (ver Tabla 31).
- La subnube de modalidades asociada a una variable tiene el mismo centro de gravedad general.

Las relaciones cuasi-bibaricéntricas

En el análisis de correspondencias múltiples las relaciones cuasi-bibaricéntricas adquieren una forma más sencilla:

Coordenada de un individuo i sobre el eje α

$$\psi_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \frac{1}{p} \sum_{j \epsilon m(i)} \varphi_{\alpha j} \qquad (38)$$

Un individuo se ubica en el promedio aritmético de las coordenadas de las modalidades que asume, alejado por el inverso de la raíz del valor propio.

Coordenada de una modalidad j sobre el eje α

$$\varphi_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \frac{1}{n_j} \sum_{j \epsilon I(j)} \psi_{\alpha i} \qquad (39)$$

Una modalidad se ubica en el promedio de las coordenadas de los individuos que la asumen, dilatada por el inverso del valor propio. Estas relaciones además de permitir la representación simultánea de los individuos y de las modalidades son básicas para la lectura de los planos factoriales.

8.5. El Análisis de Correspondencias Múltiples, ACM (Q) como un ACP(X,M,D)

Para una tabla de datos **[Q]** formada por variables cualitativas (Tenenhaus & Young, 1985; Doledec & Chessel, 1994; Chessel, 1992, p.32), notamos **Z:** tabla disyuntiva completa y $\boldsymbol{D_m}$: la matriz diagonal de frecuencias de las modalidades **m**. El término general de la tabla $\boldsymbol{Z}$ es Z_{im} igual a 1 si la categoría **m** es observada en una fila **i**).

El ACM ($\boldsymbol{Q}$) es el $\boldsymbol{ACP(ZD_m^{-1} - 1_{nm}, \frac{1}{p} D_m, \frac{1}{n} I_n)}$. Donde: $\boldsymbol{1_{nm}}$ es la matriz de $\boldsymbol{n}$ filas, y $\boldsymbol{m}$ (modalidades) columnas donde los términos valen 1, $\boldsymbol{p}$ es el número de variables.

$$\boldsymbol{X} = \left[\frac{z_{im} n}{n_m} - 1\right] \qquad \boldsymbol{M} = \frac{1}{p} \boldsymbol{D_m} \qquad \boldsymbol{D} = \frac{1}{n} \boldsymbol{I_n}$$

Todas las fórmulas del ACM se pueden derivar de las fórmulas correspondientes al ***ACP(X, M, D)*** (ver Tabla 9).

En el Análisis de Correspondencias Múltiples (ACM) se recomienda seguir los siguientes pasos:

a. Descomposición de la inercia de la tabla

El primer paso en el Análisis de Correspondencias Múltiples (ACM) tiene

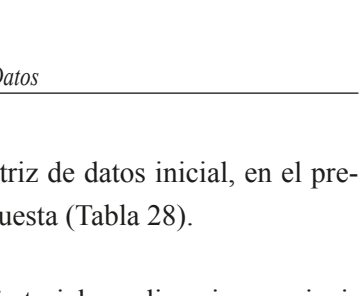
como objeto reducir las dimensiones de la matriz de datos inicial, en el presente caso, el de una tabla de datos de una Encuesta (Tabla 28).

De esta manera, se obtienen los distintos ejes factoriales o direcciones principales de alargamiento de la nube de puntos que explican las variaciones que se producen en dicha matriz (Tabla 31), los cuales posteriormente permitirán la representación factorial de la información contenida en la tabla.

Cada eje factorial viene acompañado de su valor propio, y del porcentaje de inercia, que representan la varianza explicada contenida en cada eje, así como su importancia relativa porcentual.

Tabla 31. Valores propios, inercia acumulada, porcentaje de inercia y porcentaje de inercia acumulada de los ejes factoriales

```
> dudi.tex(acm,job="ayudas.acm")
 Eigenvalues * 1000
  Eigenvalue CumInertia Percent CumPercent
1      438.7      438.7    24.4       24.4
2      363.5      802.1    20.2       44.6
3      284.2     1086.3    15.8       60.4
4      226.6     1312.9    12.6       72.9
5      171.1     1484.1     9.5       82.4
6      125.9     1610.0     7.0       89.4
7       98.7     1708.7     5.5       94.9
8       64.5     1773.2     3.6       98.5
9       26.8     1800.0     1.5      100.0
```



Fuente: Elaboración propia

De acuerdo a lo observado en la Tabla 31, se pude indicar que el primer factor con un valor propio igual a 438,7 (multiplicado por 1000) explica el 24,4 % de la varianza, el segundo factor explica el 20,2 % (valor propio = 363,5) y el tercero con un 15,8 % (valor propio = 284,2), entre los tres ejes factoriales se explica el 60,4 % de la variación.

Para el presente estudio a manera de ejemplo solo se analizarán los dos primeros ejes.

b. Extracción de los ejes factoriales (Tablas 32 y 33)

Antes de interpretar los resultados obtenidos del ACM, se debe definir cada uno de los ejes factoriales. La Figura 47 muestra las subnubes de las categorías por cada una de las variables, sobre el primer plano factorial. Se observa una muy buena separación de las categorías de FINCA y de ÉPOCA y una buena separación para las de CRUCE y NP (número de partos).

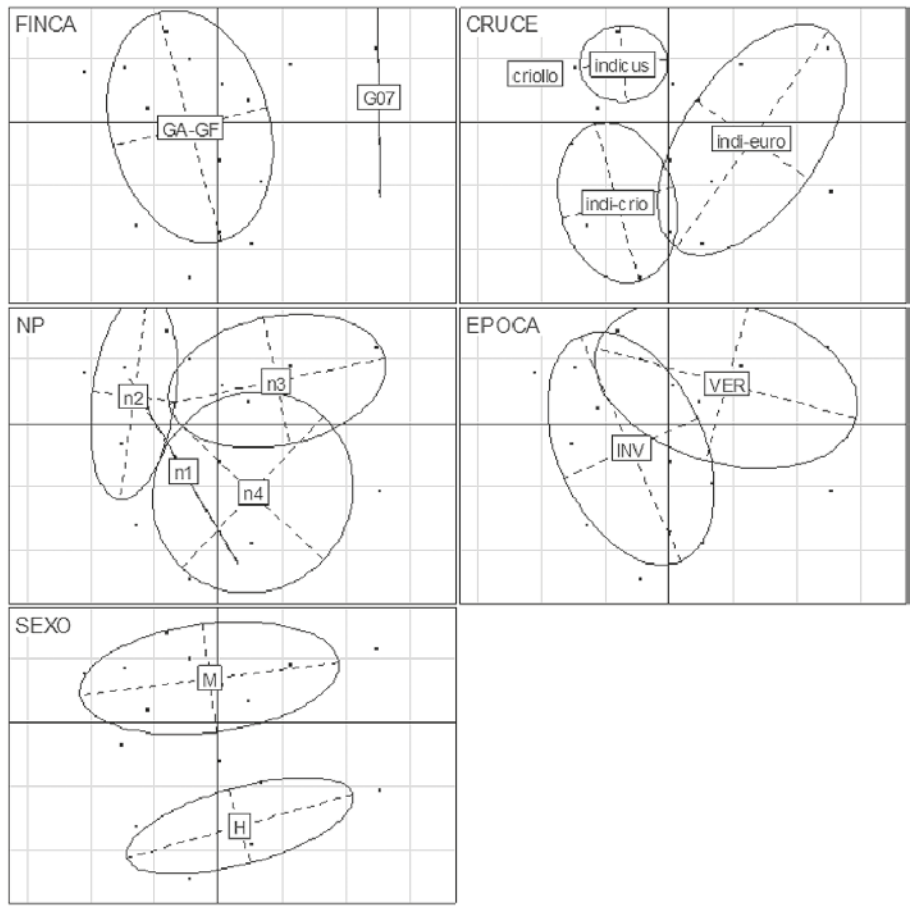


Figura 47. Subnubes en el primer plano
Fuente: Elaboración propia

Como el ACM es un ACP ponderado, para encontrar las coordenadas factoriales (filas, columnas) se utilizan las fórmulas de la Tabla 9. Las coordenadas factoriales y ayudas se interpretan para filas (*vacas*) y columnas (*modalidades de las variables de estudio*).

Tabla 32. Coordenadas y ayudas a la interpretación de individuos-vacas (filas)

Row Coordinates			Row Contributions			Repr. Quality of the Rows			
	Axis1	Axis2		Axis1	Axis2		Axis1	Axis2	con.tra
Vaca1	-0.5604	0.1042	Vaca1	3.6	0.2	Vaca1	-12.9	0.4	6.7
Vaca2	-0.3975	0.7066	Vaca2	1.8	6.9	Vaca2	-12.6	39.7	3.5
Vaca3	-0.3975	0.7066	Vaca3	1.8	6.9	Vaca3	-12.6	39.7	3.5
Vaca4	-0.2264	0.4908	Vaca4	0.6	3.3	Vaca4	-5.1	24.2	2.8
Vaca5	0.0296	0.2910	Vaca5	0.0	1.2	Vaca5	0.1	6.1	3.9
Vaca6	-0.2264	0.4908	Vaca6	0.6	3.3	Vaca6	-5.1	24.2	2.8
Vaca7	0.0084	-0.3094	Vaca7	0.0	1.3	Vaca7	0.0	-5.3	5.1
Vaca8	1.2507	0.5761	Vaca8	17.8	4.6	Vaca8	71.4	15.2	6.1
Vaca9	1.2754	-0.5436	Vaca9	18.5	4.1	Vaca9	58.1	-10.6	7.8
Vaca10	0.0101	-0.8617	Vaca10	0.0	10.2	Vaca10	0.0	-27.1	7.6
Vaca11	0.2339	0.1641	Vaca11	0.6	0.4	Vaca11	5.9	2.9	2.6
Vaca12	0.2586	-0.9555	Vaca12	0.8	12.6	Vaca12	4.3	-59.5	4.3
Vaca13	-1.0569	0.3851	Vaca13	12.7	2.0	Vaca13	-24.7	3.3	12.6
Vaca14	-0.6500	-0.8135	Vaca14	4.8	9.1	Vaca14	-22.2	-34.8	5.3
Vaca15	-0.2229	-1.2291	Vaca15	0.6	20.8	Vaca15	-2.4	-74.2	5.7
Vaca16	-0.7602	-0.1743	Vaca16	6.6	0.4	Vaca16	-38.0	-2.0	4.2
Vaca17	0.3441	-0.4750	Vaca17	1.4	3.1	Vaca17	9.1	-17.3	3.6
Vaca18	-0.7390	0.4260	Vaca18	6.2	2.5	Vaca18	-50.0	16.6	3.0
Vaca19	0.5754	0.4447	Vaca19	3.8	2.7	Vaca19	30.3	18.1	3.0
Vaca20	1.2507	0.5761	Vaca20	17.8	4.6	Vaca20	71.4	15.2	6.1

Fuente: Elaboración propia

Tabla 33. Coordenadas y ayudas a la interpretación de las categorías-variables (columnas)

Column Coordinates			Column Contributions		
	Comp1	Comp2		Comp1	Comp2
FINCA.G07	1.9009	0.3364	FINCA.G07	24.7	0.9
FINCA.GA.GF	-0.3354	-0.0594	FINCA.GA.GF	4.4	0.2
CRUCE.criollo	-1.5957	0.6387	CRUCE.criollo	5.8	1.1
CRUCE.indi.crio	-0.6133	-1.0476	CRUCE.indi.crio	3.4	12.1
CRUCE.indi.euro	0.9812	-0.2229	CRUCE.indi.euro	17.6	1.1
CRUCE.indicus	-0.5430	0.7621	CRUCE.indicus	4.7	11.2
NP.n1	-0.4154	-0.6282	NP.n1	0.8	2.2
NP.n2	-1.0068	0.3418	NP.n2	13.9	1.9
NP.n3	0.6907	0.5373	NP.n3	7.6	5.6
NP.n4	0.4073	-0.9112	NP.n4	1.9	11.4
EPOCA.INV	-0.4524	-0.3384	EPOCA.INV	5.6	3.8
EPOCA.VER	0.6786	0.5076	EPOCA.VER	8.4	5.7
SEXO.H	0.2555	-1.3486	SEXO.H	0.9	30.0
SEXO.M	-0.1095	0.5780	SEXO.M	0.4	12.9

Representation Quality of the Columns				Cumulated Representation Quality of the Columns			
	Comp1	Comp2	con.tra		Comp1	Comp2	remain
FINCA.G07	63.8	2.0	9.4	FINCA.G07	63.8	65.8	34.2
FINCA.GA.GF	-63.8	-2.0	1.7	FINCA.GA.GF	63.8	65.8	34.2
CRUCE.criollo	-13.4	2.1	10.6	CRUCE.criollo	13.4	15.6	84.5
CRUCE.indi.crio	-9.4	-27.4	8.9	CRUCE.indi.crio	9.4	36.8	63.2
CRUCE.indi.euro	64.2	-3.3	6.7	CRUCE.indi.euro	64.2	67.5	32.5
CRUCE.indicus	-15.9	31.3	7.2	CRUCE.indicus	15.9	47.1	52.9
NP.n1	-1.9	-4.4	10.0	NP.n1	1.9	6.3	93.7
NP.n2	-43.4	5.0	7.8	NP.n2	43.4	48.5	51.5
NP.n3	25.7	15.6	7.2	NP.n3	25.7	41.2	58.8
NP.n4	5.5	-27.7	8.3	NP.n4	5.5	33.2	66.8
EPOCA.INV	-30.7	-17.2	4.4	EPOCA.INV	30.7	47.9	52.1
EPOCA.VER	30.7	17.2	6.7	EPOCA.VER	30.7	47.9	52.1
SEXO.H	2.8	-78.0	7.8	SEXO.H	2.8	80.8	19.2
SEXO.M	-2.8	78.0	3.3	SEXO.M	2.8	80.8	19.2

Fuente: Elaboración propia

c. Interpretación de los ejes factoriales

Las ayudas presentadas en las Tablas 32 y 33 permiten controlar y complementar las lecturas de los planos factoriales. La contribución a la inercia del eje ayuda a encontrar un significado del eje.

Vacas-individuos (filas), en el primer eje factorial las vacas 8, 9, 13 y 20 son las que más contribuyen (acumulan el 66,8 % de la inercia del eje), entonces el eje 1 se interpreta principalmente como la contraposición de las vacas 8, 9 y 20 con la vaca 13. En el segundo eje, se contraponen la vaca 12 con la vaca 15.

Categorías-variables (columnas), en el primer eje factorial la FINCA G07, el CRUCE indi.euro y el número de partos (NP) n2 son los que más contribuyen (acumulan el 56,2 % de la inercia del eje), entonces el eje 1 se interpreta principalmente como la contraposición de las vacas que tienen el cruce *B. indicus* x *B. europeo* y vienen de la finca G07 con las vacas que tienen 2 partos. En el segundo eje, se contraponen las vacas que tuvieron partos en la ÉPOCA seca con las que tuvieron partos en la ÉPOCA de invierno.

d. Interpretación del plano factorial (1-2)

Una vez que se describen los ejes (I y II) que van a permitir caracterizar el

estudio, el siguiente paso en la investigación es analizar los planos factoriales que se forman con la unión de los ejes (en forma de pares) que el investigador de acuerdo al análisis decidió tomar en cuenta.

Lectura simultánea: (Figura 48)
Primer grupo: vacas 8, 9 y 20 del cruce *indicus x europeo*, pertenecen a la finca G07, están en su tercer parto y los partos lo tienen en la época de verano.

Segundo grupo: vacas 13, 14, 16 y 18, pertenecen a la finca GA-GF, están en su segundo parto y los partos lo tienen en la época de invierno.

Tercer grupo: vacas 10, 12, 14 y 15, pertenecen a la finca GA-GF, tienen los partos en la época de invierno y las crías son machos.

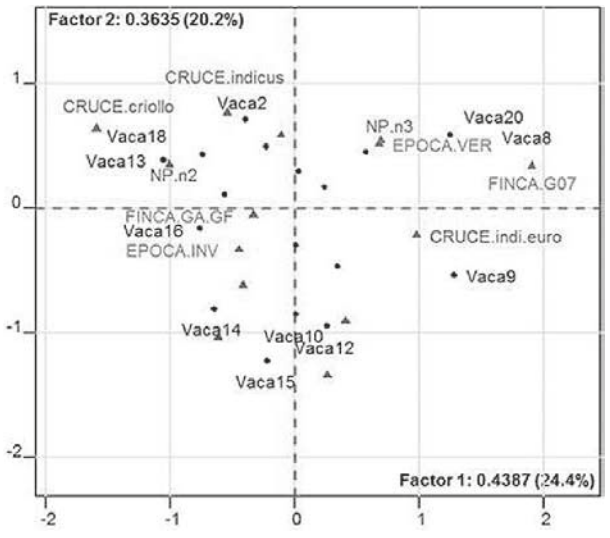


Figura 48. Plano factorial 1-2 del *ACM:* filas-individuos y columnas-variables
Fuente: Elaboración propia

e. Integración de los resultados en su contexto

Se debe resaltar que el ACM por sí solo no explica el fenómeno que se está estudiando. El investigador, en última instancia, es el que da sentido (ubica en el contexto) a los resultados obtenidos por medio de la técnica aplicada. El análisis y conclusiones a las que llegue, se fundamentan principalmente en el grado (nivel de conocimiento o manejo) que tiene sobre sus materiales.

9. INVESTIGACIONES

Se presentan en este capítulo investigaciones científicas del sector agropecuario publicadas en revistas científicas en las cuales se ha realizado parte del análisis estadístico con análisis multivariado de los datos.

Referencia	Fecha de recepción	Fecha de aprobación
Ricardo Pérez C. M., Alexander Pérez C., Melba Vertel M. Caracterización nutricional, físico-química y microbiológica de tres abonos orgánicos para uso en agroecosistema de pasturas en la subregión sabanas del departamento de Sucre, Colombia. Revista *Tumbaga* (2010), 5, 27-37	Día/mes/año 2/03/2010	Día/mes/año 12/03/2010

CARACTERIZACIÓN NUTRICIONAL, FÍSICOQUÍMICA Y MICROBIOLÓGICA DE TRES ABONOS ORGÁNICOS PARA USO EN AGROECOSISTEMAS DE PASTURAS EN LA SUBREGIÓN SABANAS DEL DEPARTAMENTO DE SUCRE, COLOMBIA

Ricardo Pérez C.[1*], M.Sc, Alexánder Pérez C.[2], Dr. Melba Vertel M.[3], M.Sc.
[1,2,3]. Universidad de Sucre, Campus Universitario Puerta Roja, Sincelejo, Colombia.
*Correspondencia: rimanper7@hotmail.com

RESUMEN

El objetivo del presente trabajo fue caracterizar nutricional, física, química y microbiológicamente tres abonos orgánicos, para su uso en agroecosistemas de pasturas en la subregión Sabanas del departamento de Sucre. Fueron preparados tres abonos orgánicos (composta de pollinaza, composta de bovinaza y lombricompost) con materias primas procedentes de la zona de estudio. A cada abono orgánico se le hizo caracterización nutricional, fisicoquímica y microbiológica. Para relacionar sus parámetros nutricionales, fisicoquímicos y microbiológicos, se emplearon técnicas multivariadas (análisis en componentes principales o análisis de correspondencias simples) y el programa estadístico R, 2009. Los resultados muestran que la composta de pollinaza pre-

senta una mayor contribución nutricional, mayor retención de humedad y una alta volatilización; la composta de pollinaza y el lombricompost muestran la más baja densidad física (0,40g/cm^3). Los tres abonos tienen concentraciones de Cd, Cr, Hg, Ni y Pb por debajo de los niveles máximos permisibles por la NTC 5167 de 2004 y concentraciones de Cu y Zn inferiores a las máximas permitidas por la Comunidad Europea-Real Decreto 824 de 2005. Con relación a la diversidad de comunidades microbiológicas, la composta de bovinaza presentó mayor diversidad poblacional de hongos y bacterias. El contenido nutricional, la facilidad de liberación, las características físicas, químicas (metales pesados) y microbiológicas de un abono orgánico son determinantes de su calidad, razón por lo cual estos deben ser evaluados antes de su aplicación en cualquier agroecosistema.

Palabras clave: Abonos orgánicos, suelo, microbiota, pasturas, sabanas.

METODOLOGÍA

El presente trabajo se realizó en el municipio de Sampués, departamento de Sucre, Colombia.

Se elaboraron tres abonos orgánicos (composta de pollinaza, composta de bovinaza y lombricompost) utilizando la siguiente infraestructura y materiales: casa con techo de plástico agrícola IT-01 y polisombra, medias canecas de plástico color azul con dimensiones de 0,90 m de largo y 0,60 m de ancho, madera para levantar las canecas a 1 m del suelo, palas, carretilla, tanque para almacenamiento de agua, regaderas y guantes impermeables. En la preparación de los abonos se empleó por cada 100 kg de sustrato: 75 kg de forraje (37,5 kg de gramíneas y 37,5 kg de leguminosas) y 25 kg de residuos animales. Para la composta de pollinaza se utilizó forraje de pasto elefante morado (*Pennisetum sp),* acacia forrajera (*Leucaena leucocephala,* C.L) y cama de pollos de engorde (compuesta por cisco y estiércol de pollos).

En la composta de bovinaza se empleó forraje de pasto elefante morado, acacia forrajera y estiércol fresco de vacunos. En el lombricompost fue utilizado forraje de pasto elefante morado*,* acacia forrajera y estiércol de vacunos (con

previo manejo durante 14 días en lugar fresco en cobertura), en este sustrato en la etapa de enfriamiento se hizo la siembra de *Eisenia foetida* y se utilizó como alimento de las lombrices durante el proceso de lombricompostaje; el proceso de compostaje y lombicompostaje tuvo una duración de cuatro meses.

Para relacionar los parámetros nutricionales, fisicoquímicos y microbiológicos de los abonos orgánicos se utilizaron técnicas descriptivas multivariadas (análisis de componentes principales o de correspondencias simples (Pardo & del Campo, 2007), mediante el programa estadístico R, 2009 (4,5) (*R Development*, 2009).

RESULTADOS

Caracterización nutricional

Al relacionar los parámetros nutricionales de los abonos orgánicos (composta de pollinaza, composta de bovinaza y lombricompost), mediante el análisis de componentes principales se observa en la componente 1, que la composta de bovinaza y el lombricompost presentan respectivamente los mayores valores de pH (7,37-7,36), K_2OT (1,2-1,1 %) y K_2OD (0,98-0,91 %), mientras que la composta de pollinaza, los más altos de NT (1,51 %), P_2O_5T (1,66 %), P_2O_5 A (1,46 %), CaO (1,98 %), Mn (0,024 %), COO (11,00 %), S (0,16 %) y CIC (36,5 %). En la componente 2, la composta de bovinaza presenta las mayores contribuciones de Bo (0,004 %), C/N (7,44), MgO (0,75 %), ceniza (13,46 %) y Fe (0,15 %) en contraposición con el lombricompost que es el de menor contribución de C/N (6,86 %), MgO (0,70 %), ceniza (12,57 %) y Fe (0,09 %) (Figura 1).

Al analizar la calidad nutricional de los tres abonos orgánicos, mediante análisis de correlación entre variables y el clúster aglomerativo se observa que la composta de bovinaza y el lombricompost son similares en cuanto a contribución de pH, K_2OT, K_2OD, NT, P_2O_5T, P_2O_5 A, CaO, Mn, COO, S y CIC, mientras que la composta de pollinaza contribuye con los valores más altos en estas variables nutricionales, a excepción de pH y K_2O total y disponible, y la composta de bovinaza muestra los valores más altos de Bo y C/N (Figuras 1 y 2).

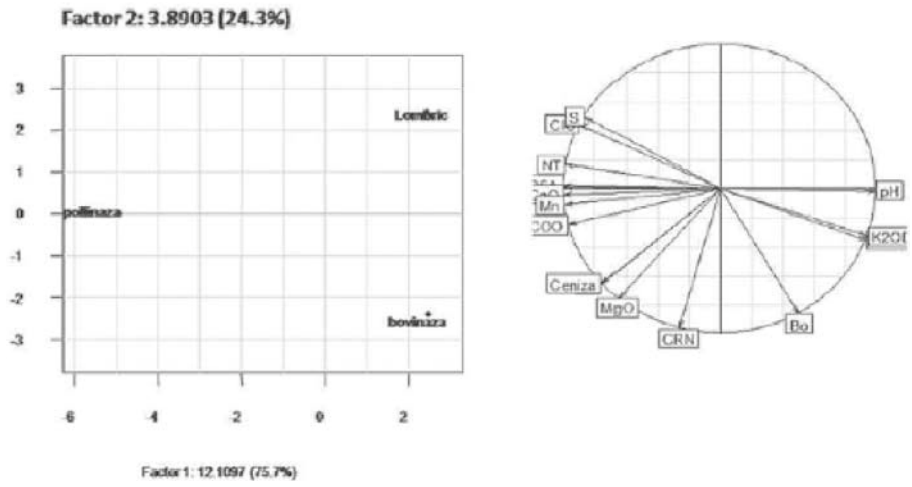


Figura 1. Análisis en componentes principales para calidad nutricional de los abonos composta de pollinaza, composta de bovinaza y lombricompost. T: total, D: disponible, A: asimilable, K_2OT: potasio total, K_2OD: potasio disponible, NT: nitrógeno total, P_2O_5T: fósforo total, P_2O_5A: fósforo asimilable, CaO: óxido de calcio, Mn: manganeso, COO: carbono orgánico oxidable, S: azufre, CIC: capacidad de intercambio catiónico, Bo: boro, Fe: hierro, C/N: relación carbono nitrógeno, primera componente: eje de las X, segunda componente: eje de las Y

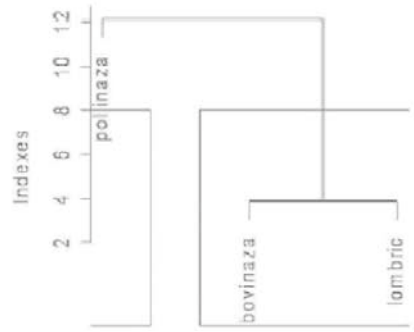


Figura 2. Análisis de clúster aglomerativo de calidad nutricional en tres abonos orgánicos

Caracterización física

Al relacionar los parámetros físicos de los abonos orgánicos, mediante el análisis de componentes principales la primera componente muestra que la composta de pollinaza presenta la más alta retención de humedad (78,05 %) y la mayor pérdida por volatilización (25,50 %), mientras que el lombricompost

tiene la más baja retención de humedad (64,73 %) y la menor pérdida por volatilización (21,82 %) (Figura 3). En la segunda componente, la composta de bovinaza presenta la mayor densidad (0,46g.cm^{-3}), mientras que la composta de pollinaza y el lombricompost presentan la más baja (0,40g.cm^{-3}) (Figura 3).

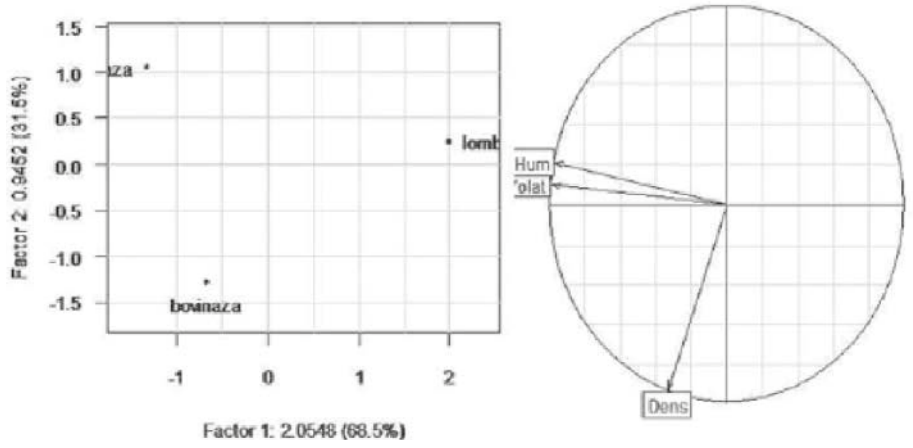


Figura 3. Análisis en componentes principales para calidad física de los abonos composta de pollinaza, composta de bovinaza y lombricompost.
Hum: humedad, volat: volatilización, dens: densidad

Caracterización de metales pesados

Al relacionar las concentraciones de metales pesados en los abonos orgánicos, mediante el análisis de componentes principales se observa en la componente 1, que la composta de bovinaza presenta la mayor concentración de Hg (0,06mg.kg^{-1}) y Cd (0,38 mg.kg^{-1}), mientras que la composta de pollinaza muestra los valores más altos de Ni (6,60 mg.kg^{-1}) y Pb (0,27 mg.kg^{-1}). En la componente 2, el lombricompost presenta la mayor concentración de Cr (8,80 mg.kg^{-1}) y la composta de pollinaza, la mayor concentración de Cu (53,20 mg.kg^{-}) y Zn (255,50 mg.kg^{-}). La cuantificación de Cr, Cd, Pb, Ni, Zn y Cu se hizo mediante espectrofotometría de absorción y la de Hg por espectrofotometría de absorción en frío (Figura 4).

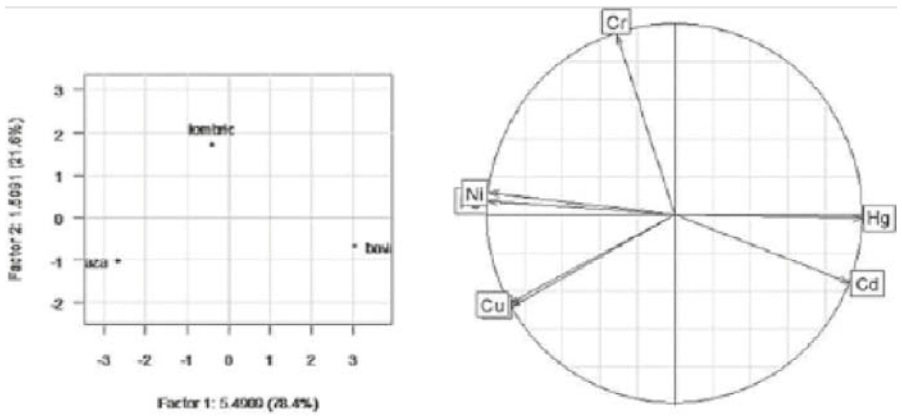


Figura 4. Análisis en componentes principales para la caracterización de metales pesados en la composta de pollinaza, la composta de bovinaza y el lombricompost

Caracterización microbiológica

Al relacionar la densidad poblacional de hongos (UFC.g^{-1}) en los abonos orgánicos, mediante el análisis de componentes principales se observa en la componente 1, que la composta bovinaza presenta las mayores densidades de *Penicillium sp* (400 x10^{3}UFC.g^{-1}), Moniliaceae (200 x10^{3}UFC.g^{-1}) y Levadura (400 x10^{3}UFC.g^{-1}), mientras que el lombricompost muestra la mayor densidad de *Aspergillus sp* (1.800 x10^{3}UFC.g^{-1}). En la componente 2, la composta de pollinaza presenta la mayor densidad de *Rhizopus sp* (100 x10^{3}UFC.gr^{-1}) (Figura 5).

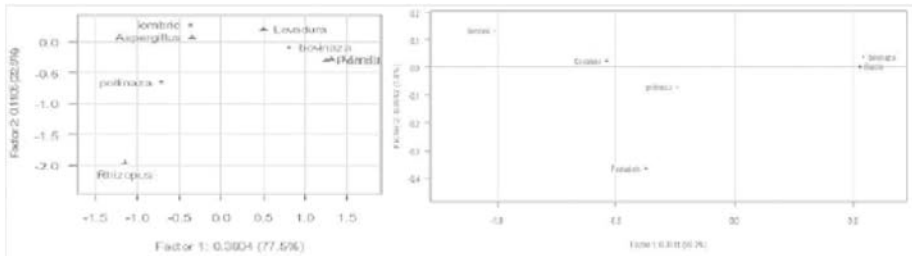


Figura 5. Análisis en correspondencias simples para densidades poblacionales de hongos y bacterias en abonos orgánicos. UFC: Unidades Formadoras de Colonias

Al relacionar la densidad de bacterias (UFC.g^{-1}) en los abonos orgánicos, mediante el análisis de componentes principales se observa en la componente 1,

que la composta de bovinaza presenta la mayor densidad perteneciente al grupo Bacilo Gram-negativo (5.500 x10^{3}UFC.g^{-1}), mientras que en la composta de pollinaza se observa la mayor densidad de Cocobacilo Gram-negativo (4.200 x10^{3}UFC.gr^{-1}). En la componente 2, la composta de pollinaza presenta la mayor densidad del género Pseudomonas (420 x10^{3}UFC.g^{-1}), mientras que el lombricompost muestra la mayor densidad de Cocobacilo Gram-negativo (2.000 x10^{3}UFC.g^{-1}) (Figura 5).

CONCLUSIONES

La calidad de un abono orgánico se determina con base en el agroecosistema en que será utilizado, por lo que es necesario identificar en la caracterización de los abonos: contenido nutricional, facilidad de suministro de nutrientes, características físicas, químicas (metales pesados) y microbiológicas, que más inciden en el agroecosistema objetivo. Teniendo en cuenta que el área ganadera de la subregión Sabanas del departamento de Sucre presenta suelos degradados, especialmente por compactación, y un periodo de sequía de 4 a 6 meses con altas temperaturas, se espera que el lombricompost ofrezca las mejores ventajas en el agroecosistema de pasturas, dado que es el abono de más rápida liberación de nutrientes, el de menor volatilización, menor densidad, tiene bajas concentraciones de metales pesados y la mayor densidad poblacional de hongos pertenecientes al género *Aspergillus*, el cual ha sido evidenciado como uno de los microorganismos participantes en el proceso de liberación de nutrientes, y además este abono presenta una alta densidad de bacterias del grupo Cocobacilo Gram-negativo, del cual se han reportado algunas especies como biodegradadoras de residuos tóxicos en el suelo.

AGRADECIMIENTOS

Los autores expresan sus agradecimientos a Sue Caribe, Universidad de Sucre, al Laboratorio Natural Control y al doctor Cristo Pérez Cordero.

BOTERO A., Luz; VERTEL M., Melba; FLÓREZ M., Lisbeth; MEDINA P., Javier
CALIDAD COMPOSICIONAL E HIGIÉNICO-SANITARIA DE LECHE CRUDA ENTREGADA EN ÉPOCA SECA POR PRODUCTORES DE GALERAS, SUCRE
Vitae, vol. 19, núm. 1, enero-abril, 2012, pp.S314-S316
Universidad de Antioquia
Medellín, Colombia

CALIDAD COMPOSICIONAL E HIGIÉNICO-SANITARIA DE LECHE CRUDA ENTREGADA EN ÉPOCA SECA POR PRODUCTORES DE GALERAS-SUCRE

COMPOSITIONAL, HYGIENIC AND HEALTH QUALITY OF RAW MILK DELIVERED BY THE DRY SEASON PRODUCERS, GALERAS-SUCRE

Luz Botero A.[1], Melba Vertel M.[2*], Lisbeth Florez M.[1] Javier Medina P.[1]
[1] Departamento de Zootecnia, Facultad Ciencias Agropecuarias, Universidad de Sucre.
[2] Grupo de Investigación Estadística y Modelamiento Matemático Aplicado a Calidad Educativa, Universidad de Sucre.
*Correspondencia: melba.vertel@unisucre.edu.co

RESUMEN

El objetivo del presente trabajo fue caracterizar calidad composicional, higiénica y sanitaria de leche cruda entregada en época seca por proveedores asociados a la Cooperativa Agropecuaria de Galeras (Sucre) como un aporte para la salud pública. El análisis de varianza y las pruebas de comparación de promedios para variables sanitarias y composicionales se basaron en el diseño experimental bloques completos aleatorizados (tratamientos: clasificación de hatos ganaderos; bloques: fincas muestreadas). Para tipificar hatos de acuerdo a buenas prácticas de manejo de ordeño y causas de contaminación en leche cruda utilizaron análisis descriptivos multivariados. La certificación de hatos ganaderos se evaluó con estadística univariada. El recuento de aerobios mesófilos mostró diferencias significativas, mayores valores en hatos pequeños-grandes (superando rangos permitidos). Los resultados generales

mostraron bajos niveles tecnológicos de las fincas asociadas (infraestructura y prácticas de manejo), aunque la leche producida presenta buenos porcentajes en su composición. La incidencia de mastitis fue alta (> 20 %).

Palabras clave: Leche cruda, calidad microbiológica, salud pública, análisis en componentes principales, sistema de producción doble propósito.

INTRODUCCIÓN

La leche es el líquido secretado por la glándula mamaria de los mamíferos, pudiendo variar su composición entre diferentes especies y dentro de la misma especie por efecto de factores relacionados con la raza, intervalo entre ordeños, cuartos de la ubre, estaciones climáticas, alimentación, enfermedades, temperatura ambiental, edad, entre otros.

Dentro de las enfermedades, la mastitis es una de las que más afecta la producción y calidad de la leche; se estima que 33,3 % está afectada por mastitis en uno o más cuartos. La producción de leche es una actividad económica del municipio de Galeras-Sucre. Las fincas asociadas son manejadas en el sistema de producción doble propósito, producción 1-3 litros/vaca/día. El presente trabajo caracterizó la calidad composicional e higiénico-sanitaria de leche cruda entregada en época seca por productores de este municipio.

MATERIALES Y MÉTODOS

El estudio se realizó a los hatos ganaderos asociados de la zona rural de Galeras-Sucre (Colombia). El análisis estadístico tiene *análisis univariado* utilizando diseño de bloques completos aleatorizados (*tratamientos* - clasificación FEDEGAN: T1-pequeño, T2-mediano y T3-grande). Se tomaron muestras según NTC para mediciones microbiológico-composicionales. Mastitis evaluada por test California y certificación por ficha técnica. Se realizó análisis de varianza y pruebas DMS (5 %). También, análisis multivariado para tipicar sistemas de producción lechera. Para el análisis estadístico se utilizó software R y los paquetes *agricolae* y *ade4*.

RESULTADOS Y DISCUSIÓN

Aerobios mesófilos: El ANAVA muestra diferencias significativas (Tabla 1). Pequeñas-grandes fincas presentaron contaminación y límites establecidos.

Tabla 1. Resultados de promedios, desviaciones y análisis de varianza (ANAVA)

Variables	**Clasificación de hatos ganaderos**			**ANAVA (*p-value*)**	
	Pequeño (T1)	*Mediano (T2)*	*Grande (T3)*	*Bloques*	*Tratamiento*
Aerobios mesófilos	13,93 ± 0,64 [a]	11,80±0,54 [b]	13,08±0,51 [ab]	0,11 [NS]	0,03*
Grasa %	4,06±0,13	4,42±0,22	4,16±0,30	0,62 [NS]	0,95[NS]
Proteína %	3,57±0,10	3,76±0,06	3,57±0,12	0,47 [NS]	0,98[NS]
Sólidos totales %	12,30±0,33	13,21±0,27	12,66±0,46	0,34 [NS]	0,96[NS]

*: Diferencias significativas al nivel del 5 %, NS: No significativo

Grasas, proteínas y sólidos totales (%): No hubo diferencias significativas para bloques y tratamientos (Tabla 1). Los valores encontrados estuvieron encima de lo establecido en grasas y proteínas, mientras sólidos totales, en rango de aceptación. En el ACP ponderado, hatos ganaderos producen menor volumen de producción aunque tienen buena composición química (Figura 1).

Caracterización del ordeño: La Figura 2 muestra relación entre fincas con prácticas y condiciones antes, durante y después del ordeño. En la parte superior hatos medianos-grandes relacionados a la forma (cantinas de aluminio) y desinfección de recipientes y utilizan agua potable para limpieza (mayor nivel tecnológico e infraestructura adecuada). Las fincas (hatos pequeños) se agrupan en la parte izquierda, reflejan bajo nivel tecnológico.

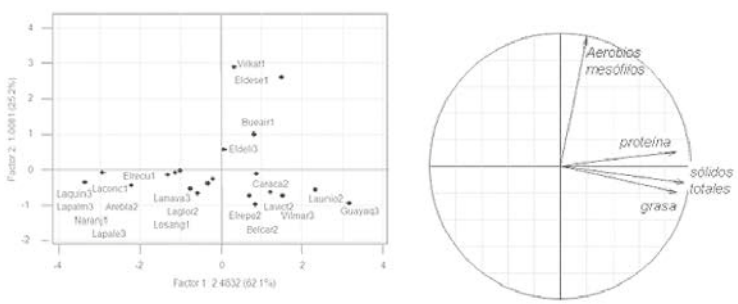


Figura 1. Plano factorial 1-2 del análisis en componentes principales de recuento de aerobios mesófilos y variables composicionales de leche cruda en fincas ganaderas

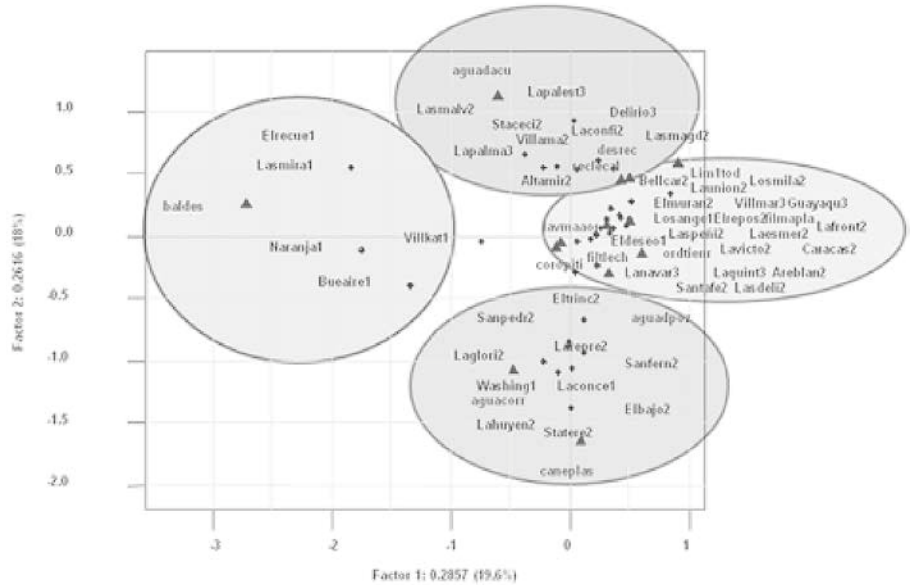


Figura 2. Primer plano factorial del análisis de correspondencias simples de prácticas de ordeño de ganado bovino manejado bajo el sistema de producción doble propósito en Galeras (Sucre)

Incidencia de mastitis: Alta (>20 %), puede estar relacionada a falta de conocimiento de los propietarios-trabajadores sobre buenas prácticas agropecuarias.

Certificación de hatos libres de aftosa y brucelosis: Teniendo en cuenta ficha técnica se encontró que fincas asociadas a la cooperativa realizan vacunación de aftosa-brucelosis, mas no cuentan con certificación de hatos libres de aftosa.

CONCLUSIONES

La falta de implementación de buenas prácticas de ordeño y el bajo nivel tecnológico de los hatos en estudio son la principal causa de contaminación de la leche producida. Aunque, la calidad composicional de la leche cruda, expresada en porcentaje de proteína, grasa y sólidos totales en los tres tratamientos, se encuentran por encima de los mínimos establecidos por la ley.

Infectio. 2014;18(3):93-99

DETECCIÓN DE *TOXOPLASMA GONDII* POR AMPLIFICACIÓN DEL GEN B1 EN CARNES DE CONSUMO HUMANO

Diana Marcela Campo-Portacio, Maira Alejandra Discuviche-Rebolledo, Pedro José Blanco-Tuirán, Yina Margarita Montero-Pérez, Kelly Estela Orozco-Méndez, Yulenis Margarita Assia-Mercado
Grupo de Investigaciones Biomédicas, Universidad de Sucre, Colciencias, Sincelejo, Sucre, Colombia

Palabras clave: *Toxoplasma gondii*, carne, ADN, pollo, cerdo, res.
Key words: *Toxoplasma gondii*, meat, DNA, chicken, pork, beef.

RESUMEN

Objetivo

Determinar la frecuencia de formas parasitarias de *Toxoplasma gondii (T. gondii)* en diferentes tipos de carne de consumo humano comercializadas en Sincelejo-Sucre, mediante la amplificación del gen B1 por la técnica de reacción en cadena de la polimerasa (PCR).

MATERIALES Y MÉTODOS

Se realizó un estudio descriptivo que determinó la infección por *Toxoplasma gondii* en 120 muestras de carnes de consumo humano, obtenidas en dos tipos de expendios del municipio de Sincelejo. De cada sector se tomaron 60 muestras distribuidas así: 20 muestras de carne de res, 20 muestras de carne de cerdo y 20 muestras de carne de pollo. Estas muestras fueron sometidas a una extracción de ADN mediante el método de altas concentraciones de sales y a una PCR anidada para amplificar una región específica del material genómico de *T. gondii* correspondiente al gen B1.

RESULTADOS

Se detectó ADN de *Toxoplasma gondii* en el 32 % de las carnes analizadas. Dentro de este porcentaje se encontraron en proporciones similares, formas parasitarias de *T. gondii* en carne de pollo (35 %), cerdo (32,5 %) y res (27,5 %), por lo cual no se observó diferencia estadística al realizar el análisis por tipo de carne. Así mismo se encontró una frecuencia de formas parasitarias de *T. gondii* de 36,6 % en las muestras recolectadas en el mercado público y 26,7 % en las de los supermercados de cadena.

CONCLUSIONES

Esta investigación demuestra la alta frecuencia de formas parasitarias de *T. gondii* en diferentes tipos de carne de consumo humano comercializadas en el municipio de Sincelejo, e indica un elevado riesgo de infección por el parásito en la población.

MATERIALES Y MÉTODOS

Tipo de estudio

Se realizó un estudio descriptivo en el que se detectó ADN de *T. gondii* en muestras de carne de res, pollo y cerdo, obtenidas en diferentes expendios ubicados en el municipio de Sincelejo.

Área de estudio

Este estudio fue realizado en la ciudad de Sincelejo, departamento de Sucre, ubicado al noroeste del país a 9° 18" de latitud norte, 75° 23" de longitud oeste del meridiano de Greenwich. Tiene una extensión total de 28.410,31 ha. La temperatura media anual está cercana a los 27,15 °C ± 0,4; con una mínima promedio anual de 19,7 °C y una máxima de 35,3 °C.

Muestra de estudio

Se recolectaron 120 muestras de músculo esquelético en dos lugares dedi-

cados a la venta de productos para la canasta familiar: el mercado público y supermercados de cadena de la ciudad de Sincelejo. De cada sector se tomaron 60 muestras distribuidas así: 20 muestras de carne de res, 20 muestras de carne de cerdo y 20 muestras de carne de pollo. Cada una con un peso aproximado de 5g. Fueron empacadas en bolsas resellables, y transportadas desde el lugar de expendio hasta el Laboratorio de Investigaciones Biomédicas de la Universidad de Sucre donde fueron conservadas a -20 °C por 24h de forma previa a la manipulación.

Extracción de ADN genómico

La extracción de ADN se llevó a cabo mediante el método de altas concentraciones de sales. Se maceraron 150 mg de la muestra en 500 µL de buffer de lisis (SDS [*sodium dodecyl sulfato*, BioAmérica, Miami, Estados Unidos] 0,1 %, EDTA [*disodium salt dihydrate*, BioAmérica, Miami, Estados Unidos] 1 mM, Tris-HCl [BioAmérica, Miami, Estados Unidos] 10 mM) por 10 min. A la muestra homogeneizada se le adicionaron 6,2 µL de proteinasa K (*proteinase K*, Promega, Madison, Estados Unidos); se incubó a 55 °C por 4 h y se inactivó la enzima a 94 °C por un minuto. Luego se adicionaron 150 µL de NaCl (*sodium chloride*, Amresco Bioexpress, Ohio, Estados Unidos) 6 M y se centrifugó a 4 °C y 12.000 rpm durante 10 min. El sobrenadante de cada muestra fue depositado en un nuevo vial y a cada uno de ellos se le agregó el doble de su volumen en etanol absoluto (Merck, Alemania) y se almacenó a -20 °C durante toda la noche, para precipitar el ADN total. Terminada la precipitación del ADN, se centrifugó a 4 °C y 12.000 rpm durante 10 min y se realizaron 3 lavados con 1.000 µL de etanol al 70 %. Finalmente el ADN se secó a 55 °C por 20 min y se resuspendió el ADN en 100 µL de tampón TE (Tris-HCl Bioamérica 10 mM, EDTA Bioamérica 0,1 mM).

Cuantificación y verificación de la integridad del ADN por reacción en cadena de la polimerasa

Luego del proceso de extracción, se realizó la cuantificación del ADN obtenido en un espectrofotómetro Nanodrop 2000. A cada una de las muestras

se le determinó su concentración y su pureza de acuerdo a la proporción OD 260 nm/OD 280 nm. Se obtuvo un promedio de concentraciones de 567 ng/µL y un promedio de proporción OD 260 nm/OD 280 nm de 1,75. Adicionalmente, se realizó una PCR convencional con el fin de verificar la integridad del ADN extraído, a través de la cual fue amplificado un fragmento de 359 pb de la región conservada del gen mitocondrial cyt b de vertebrados, utilizando los cebadores (cyt b1): 5'-CCA TCC AAC ATC TCA GCA TGA TGA AA-3' y (cyt b2): 5'-GCC CCT CAG AAT GAT ATT TGT CCT CA-3' sintetizados por Integrated DNA Technologies IDT.

Amplificación del gen B1

La detección del ADN del parásito se realizó mediante una PCR anidada con el fin de aumentar la especificidad y sensibilidad de la técnica; para lo cual se emplearon dos pares de cebadores, sintetizados por Integrated DNA Technologies IDT, que amplifican una región del gen B1 de *T. gondii.*

ANÁLISIS DE DATOS

Los resultados obtenidos fueron introducidos en una base de datos diseñada en Microsoft Excel 2007, para su posterior análisis mediante el programa R versión 2.7.1. Se hallaron frecuencias de infección por tipo de carne, por lugar de recolección, intervalos de confianza y se realizó análisis de correspondencia múltiple.

RESULTADOS

De las 120 muestras de carnes de cerdo, pollo y res analizadas mediante PCR anidada y después de visualizar los productos obtenidos en los geles de agarosa, fue posible detectar ADN del parásito *T. gondii* en 38 de las muestras (Figura 1), las cuales representan el 32 % de la muestra de estudio. Esto permite afirmar con un 95 % de confianza que en Sincelejo las carnes de consumo humano que se comercializan están infectadas por *T. gondii* en porcentajes que fluctúan del 23,6 al 40,8 %.

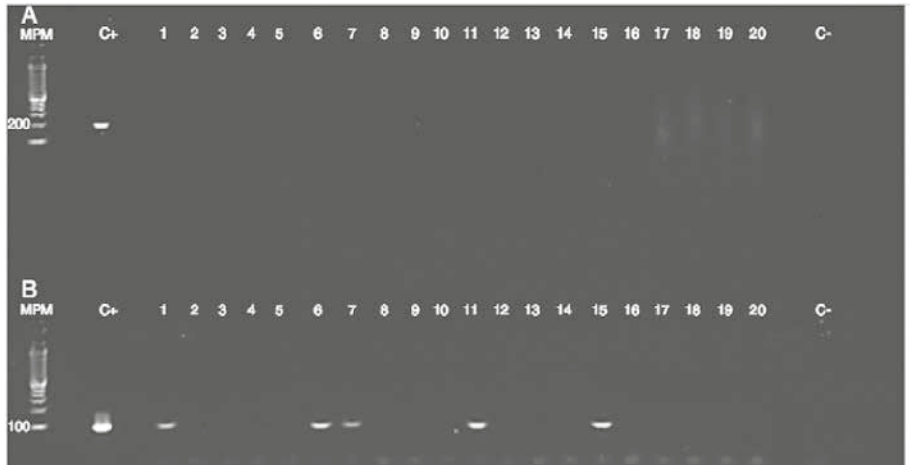


Figura 1. Electroforesis en gel de agarosa al 2 %

Teniendo en cuenta el lugar de recolecta, se obtuvo que en 36,6 % (22 de 60) de las muestras recolectadas en el mercado público se detectó ADN de *T. gondii* y en los supermercados de cadena solo se encontraron 16 muestras positivas para la infección con el parásito de un total de 60 analizadas, lo que representa un 26,7 %. Estos resultados no presentan diferencia estadísticamente significativa (Tabla 1).

Al realizar el análisis de independencia con un 95 % de confianza, se obtuvo que la presencia de formas parasitarias de *T. gondii* en carnes es independiente del lugar de recolecta, es decir que se pueden consumir proporciones similares de carne infectada con el parásito procedente de ambos tipos de expendios (X^2_{cal} = 0,9628; p = 0,3265 > 0,05).

Tabla 1. Tabla de contingencia r x n para la frecuencia de infección por *T. gondii*, de acuerdo al sitio de expendio y el tipo de carne

Variable	N° de muestras evaluadas	N° de muestras positivas	Prevalencia (%)	IC 95 %	X^2 (p)
Lugar					0,9628 (0,3265)
Supermercados de cadena	60	16	26,7	16,5-39,9	
Mercado público	60	22	36,6	24,9-50,2	

Tipo de carne					1,016 (0,6017)
Cerdo	40	13	32,5	19,0-49,2	
Pollo	40	14	35	21,2-51,7	
Res	40	11	27,5	13,2-41,5	

IC: intervalo de confianza.

El análisis de proporciones realizado para la detección de formas parasitarias por lugar de recolecta mostró que, con un 95 % de confianza, las carnes que se comercializan en el mercado público están infectadas con *T. gondii* en proporciones que van de 24,8 a 50,1 %, en tanto las carnes que se comercializan en los supermercados de cadena están infectadas con el parásito en porcentajes que van de 16,4 a 39,8 %.

Con relación al tipo de carne, se encontró que la carne de pollo fue la que presentó mayor porcentaje de infección por *T. gondii*, con un total de 14 muestras positivas, las cuales representan el 35 %. En segundo lugar se encuentra la carne de cerdo, con 13 (32,5 %) muestras positivas y por último, la carne de res con 11 (27,5 %) muestras positivas. Estos resultados no presentan diferencia estadística (Tabla 1).

Al realizar el análisis de independencia se obtuvo que, con un 95 % de confianza, la detección de formas parasitarias de *T. gondii* es independiente del tipo de carne, es decir que en Sincelejo se pueden consumir proporciones similares de carnes de pollo, res y cerdo, infectadas con el parásito causante de la toxoplasmosis (X^2 = 1,0; p = 0,6).

El análisis de proporciones, con un 95 % de confianza, para los tipos de carne, mostró que en Sincelejo se puede consumir carne infectada de pollo en 21,1 a 51,7 %, carne de cerdo en 19 a 49,2 % y carne de res en 13,2 a 41,5 %.

El análisis de correspondencia múltiple mostró que la presencia de *T. gondii* en los tres tipos de carnes estudiados (cerdo, pollo y res) y los sitios de recolecta son independientes, lo que permite observar en forma agrupada, en el

plano 1 de la Figura 2, las muestras positivas de estos tres tipos de carne y de cualquiera de estos lugares.

Además, sugiere que es posible encontrar el mismo porcentaje de carnes positivas para la infección por *T. gondii* en ambos lugares y en todos los tipos de carne aquí analizados.

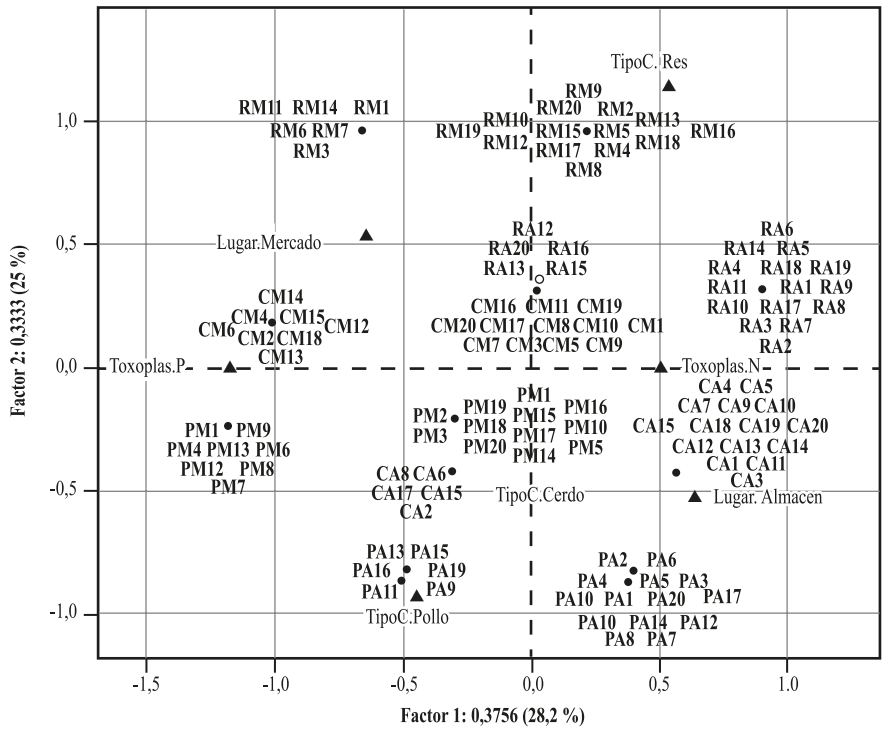


Figura 2. Plano de análisis de correspondencias múltiples

AGRADECIMIENTOS

Este estudio fue realizado en el marco del proyecto código 112949326224 financiado por COLCIENCIAS (contrato número 728-2009). Agradecemos al Grupo de Investigaciones Biomédicas de la Universidad de Sucre el apoyo brindado para la ejecución de este proyecto y a la docente Melba Liliana Vertel Morinson por su asesoría en el análisis estadístico de los datos.

APÉNDICE

APÉNDICE I. Software R

R (*R Development Core Team* 2014) es un entorno especialmente diseñado para el tratamiento de datos, cálculo y desarrollo gráfico. El lenguaje de programación R forma parte del proyecto GNU (RefBib) y puede verse como una implementación alternativa del lenguaje S, desarrollado en AT&T Bell Laboratories. Se presenta como software libre, donde este término se refiere a la libertad de los usuarios para ejecutar, copiar, distribuir, estudiar, cambiar y mejorar el software. Se trata de un lenguaje creado específicamente para la visualización y exploración de datos, permite trabajar con facilidad con vectores, matrices y ofrece diversas herramientas para el análisis de datos, modelización y programación estadística.

En la web http://www.r-project.org/index.html se encuentra disponible toda la información acerca de R. La instalación de R se realiza a través de la CRAN (*Comprehensive R Archive Network*). En resumen, R proporciona un entorno de trabajo especialmente preparado para el análisis estadístico de datos. Con excepción de algunas funciones, la mayoría están localizadas en el paquete *base*. Algunas de las funciones para análisis de datos están en *base* pero la gran mayoría de los métodos estadísticos disponibles en R están distribuidos como paquetes *packages*.

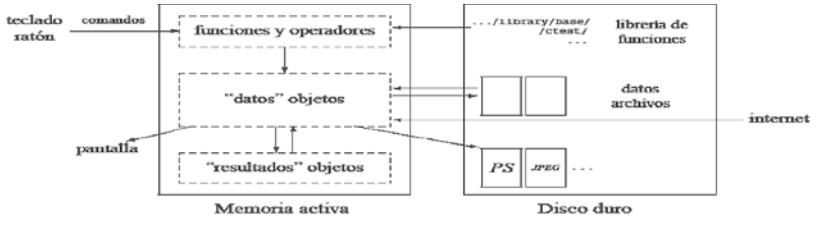


Figura 49. Una visión esquemática del funcionamiento de R
Fuente: Adaptado de varios autores

Entre los paquetes de mayor utilización en este libro están: *ade4* y *FactoClass*.

ade4: Es un software desarrollado en el Laboratorio de Biometría y Biología Evolutiva (UMR 5558) de la Universidad de Lyon 1. Se caracteriza por: implementación de funciones gráficas y estadísticas, suministro de datos digitales, elaboración de una documentación técnica y temática, e inclusión de referencias.

En Thioulouse *et al*. (2005), hace un breve resumen de las principales clases definidas en el paquete *ade4* para los métodos factoriales de análisis de tablas de datos *(*por ejemplo: el análisis en componentes principales). Este paquete es una reescritura completa del software ADE4 (Thioulouse *et al.*, 1997), http://pbil.univlyon1.fr/ADE-4/ para R medioambiental. El paquete está disponible en el CRAN *ade4*, también se puede utilizar directamente en línea, gracias al sistema Rweb (http://pbil.univ-lyon1.fr/Rweb/).

Del paquete estadístico *ade4* para realizar el análisis factorial que corresponde a una tripleta particular (ver sección 4.6) de una tabla de datos, se utilizaron en particular las funciones:

dudi.pca: Análisis en Componentes Principales (ACP)
dudi.coa: Análisis de Correspondencias Simples (ACS)
dudi.acm: (Análisis de Correspondencias Múltiples (ACM)

La descomposición de una tripleta particular ***ACP(X, M, D)*** es una función del *ade4*, que retorna un objeto de tipo *dudi* con los valores y vectores propios, coordenadas factoriales de filas y columnas. Las funciones *dudi* de ade4 reciben los datos en un objeto *data.frame* y utilizan todas las columnas como activas. Las demás ayudas a la interpretación (coordenadas factoriales, contribuciones, cosenos cuadrados, distancias al cuadrado) se obtienen con la función *inertia.dudi* de ade4.

FactoClass: Es un paquete de R que combina métodos factoriales con análisis de conglomerados, en la exploración multivariada de tablas de datos. Utiliza funciones de *ade4* (Thioulouse *et al.*, 2004) para realizar el análisis factorial de los datos.

El paquete ade4 tiene varias funciones para obtener los planos factoriales; sin embargo, en *FactoClass* se incluye la función *planfac* o *plot.dudi* que recibe un objeto *dudi* y produce un plano factorial similar a los del paquete FactoMineR (Husson *et al.*, 2011) o a los de ade4. Programaron una función complementaria llamada *dudi.tex* que presenta en forma tabular los resultados de valores y vectores propios, coordenadas factoriales de filas y columnas, y demás ayudas a la interpretación que se requieren en forma amigable.

Ahora estamos listos para utilizar R, *ade4* y *FactoClass* para ejecutar los métodos básicos de análisis multivariado de datos. Sin embargo es más cómodo utilizar un editor de texto para almacenar los comandos de R en forma organizada.

Para el tema que nos ocupa, el objeto principal de *ade4* es *dudi*, el cual se obtiene con la función *as.dudi*, que es llamada por las funciones que realizan cualquier método factorial, por ejemplo *dudi.pca*, *dudi.coa*, *dudi.mca*.

La función interna *as.dudi (df, col.w, row.w)* realiza un ***ACP(X, M, D)*** donde:

df: data.frame con *n* filas y *p* columnas ***(X)***
col.w: un vector numérico con los pesos de las filas (*col.w[i]* = D[i,i])
row.w: un vector numérico con los pesos de las columnas (*row.w[j]* = M[j, j])

La salida de *as.dudi* es un objeto *dudi* que es un *list* de los objetos:
tab: ***(X)***
cw: pesos de las filas (D[i,i])
lw: pesos de las columnas (M[j,j])
eig: valores propios, un vector con *min(n, p)* componentes
nf: entero, número de ejes guardados
c1: ejes principales (vectores propios en $\mathbb{R}^p$), *data.frame* con *p* filas y *nf* columnas
l1: componentes principales, *data.frame* con *n* filas y *nf* columnas
co: coordenadas de las columnas, *data.frame* con *p* filas y *nf* columnas
li: coordenadas de las filas, data.frame con *n* filas y *nf* columnas
call: llamado original de la función *as.dudi*

Para el ejemplo del Capítulo 6, que realiza un Análisis en Componentes Principales (ACP), la salida en R es la siguiente:

```
acp<-dudi.pca(datos,scann=FALSE,nf=3); acp
Duality diagramm
class: pca dudi
$call: dudi.pca(df = datos, scannf = FALSE, nf = 3)

$nf: 3 axis-components saved
$rank: 4
eigen values: 2.425 0.961 0.5761 0.03807
   vector   length   mode        content
1  $cw      4        numeric     column weights
2  $lw      7        numeric     row weights
3  $eig     4        numeric     eigen values

   data.frame  nrow     ncol     content
1  $tab        7        4        modified array
2  $li         7        3        row coordinates
3  $l1         7        3        row normed scores
4  $co         4        3        column coordinates
5  $c1         4        3        column normed scores
other elements: cent norm
```


APÉNDICE II. Script en R para los resultados de los capítulos del libro

A lo largo del libro se presentan resultados gráficos, numéricos y tabulares de las prácticas con diferentes conjuntos de datos con los que pretendíamos ilustrar los contenidos.

Los conjuntos de datos han sido creados específicamente para el fin de este libro. Todos ellos se encuentran disponibles en este apéndice. De esta manera se podrán reproducir las prácticas incluidas a lo largo de los capítulos.

REPRESENTACIÓN GRÁFICA DE LOS DATOS

```
#Script generado en Software R para Capitulo 3

#Datos originales
NBI<-c(16.1,30.0,35.7,35.8,37.5,30.6,42.4);NBI
analfab<-c(4.5,9.5,13.7,15.8,14.4,13.4,15.3);analfab
desempleo<-c(12.8,9.8,6.2,13.6,5.9,7.1,6.2);desempleo
PIBperc<-c(1.67,8.59,1.36,1.20,3.02,0.75,0.80);PIBperc
#Base de datos
datos<-data.frame(NBI,analfab,desempleo,PIBperc);datos

#Resumen numérico
summary(datos)

#Grafico histograma
hist(NBI); hist(analfab); hist(desempleo); hist(PIBperc)

#Gráfico de los 4 histogramas en una sola salida
par(mfrow = c(2, 2))
hist(NBI); hist(analfab); hist(desempleo); hist(PIBperc)

#Plano cartesiano básico
plot(analfab,NBI)

#Plano cartesiano con detalles
plot(analfab,NBI,type="p",col="blue",lwd=4);grid(col="black")

#Gráficos de columnas
barplot(t(datos),beside=TRUE, col=rainbow(5))
boxplot(datos)
```

```
#Gráfico de tallos y hojas
stem(NBI); stem(analfab); stem(desempleo); stem(PIBperc)
pairs(datos,col = rainbow(5), pch = 15)
plot(datos, col = rainbow(5), pch = 15)

#Librerias que se necesitan para gráficos especiales
library(rscproxy)
library(TeachingDemos)

#Gráfico de Chernoff
# cuyo código es:
faces2(datos)

# Gráfico de estrellas
stars(datos)

# Histogramas simulados
x<-rnorm(100,mean=10,sd=1) #introduzca sus datos en x
#el histograma
histo<-hist(x,prob=T,main="",ylim=range(hist(x)$density))
#la densidad
z<-pretty(histo$breaks,n=50)
y<-dnorm(z,mean=mean(x),sd=sd(x))
lines(z,y,lty=3,lwd=2,col="blue")
#para más detalles de la función hist
#digite en la consola ?hist
# programa para verificar el TLC con las siguientes
# distribuciones: binomial(20,0.5), uniforme,
# chi- cuadrada y una exponencial
win.graph()
par(mfrow=c(2,2))
aux<-matrix(0,100,1000)
muestras<- matrix(rbinom(aux,20,0.5),100,1000)
binomial<-apply(muestras,2,mean)
hist(binomial,col=5:8)
muestras<-matrix(runif(aux),100,1000)
uniforme<-apply(muestras,2,mean)
hist(uniforme,col=3:7)
muestras<-matrix(rchisq(aux,3),100,1000)
chicuadrado<-apply(muestras,2,mean)
hist(chicuadrado,col=6:1)
muestras<-matrix(rexp(aux),100,1000)
exponencial<-apply(muestras,2,mean)
hist(exponencial,col=3:5)
par(oma=c(1,1,1,1),new=T,font=2,cex=1)
```



```
# Ejemplo 1.
y<-c(1.38,1.40,1.42,1.54,1.30,1.55,1.50,1.60,1.41,1.34);y
i<-c(0.05,0.15,0.25,0.35,0.45,0.55,0.65,0.75,0.85,0.95);i
q<- c(qnorm(0.05),qnorm(0.15),qnorm(0.25),qnorm(0.35),qnorm(0.45),
qnorm(0.55),qnorm(0.65),qnorm(0.75),qnorm(0.85),qnorm(0.95));q
ej1<-data.frame(y,i,q);ej1

#Gráfico q-q
y<-c(1.38,1.40,1.42,1.54,1.30,1.55,1.50,1.60,1.41,1.34);y
qqnorm(y)
qqline(y) # pasa la línea
#Prueba Ji-cuadrado
y<-c(1.38,1.40,1.42,1.54,1.30,1.55,1.50,1.60,1.41,1.34);y

library(nortest)
pearson.test(y)

# Prueba K-S
y<-c(1.38,1.40,1.42,1.54,1.30,1.55,1.50,1.60,1.41,1.34);y
ks.test(y,"pnorm",1.4,0.1)

#Prueba Shapiro-Wilks
y<-c(1.38,1.40,1.42,1.54,1.30,1.55,1.50,1.60,1.41,1.34);y
shapiro.test(y)

#Prueba de Asimetria y kurtosis
library(moments)
y<-c(1.38,1.40,1.42,1.54,1.30,1.55,1.50,1.60,1.41,1.34);y
agostino.test(y)
anscombe.test(y)

#Prueba omnibus
y<-c(1.38,1.40,1.42,1.54,1.30,1.55,1.50,1.60,1.41,1.34);y
library(moments)
b1<-skewness(y)^2;b1
b2<-kurtosis(y);b2
n<-length(y);n
X2<-((n*b1)/6)+(n*(b2-3)^2/24);X2
pvalor<-pchisq(X2,2,lower.tail=FALSE);pvalor
```

UTILIZACIÓN DEL ÁLGEBRA Y LA GEOMETRÍA EN ACP (𝕏,𝕄,𝔻)

```
# Hallar para todos los casos valores y vectores propios

#Activación de paquetes
Library (xtable)
Library (Factoclass)

#Datos originales:
x1←c(18,13,15,9,11,5,7,2); x1
x2←c(9,9,5,7,3,5,1,1); x2
datos < -data.frame(x1,x2);datos
A←(7/8). Var(datos); A

#ACP datos originales
acp ← dudi.pca(datos,center=F,scale=F,scannf=F,nf=2);acp
R/    λ1 = 155       λ2 = 3,75      acp.c$tab

#ACP Datos centrados
acp.c <- dudi.pca(datos,center=T,scale=F,scannf=F,nf=2);acp.c
R/    λ1 = 30        λ2 = 3,75      acp.c$tab

#ACP Datos estandarizados
acp.z <- dudi.pca(datos,center=T,scale=T,nf=2);acp.z
R/    λ1 = 1.7035    λ2 = 0,2965

acp.z1 <- dudi.pca(datos, scannf= F =, nf= 2); acp; acp.z1
R/    λ1 = 1.7035    λ2 = 0,2965
```


ANÁLISIS EN COMPONENTES PRINCIPALES

```
# Activación de paquetes
library(ade4)
library(xtable)
library(FactoClass)

# La base de datos "caribe.txt" debe ser creada en un archivo de
# Texto y guardado en una carpeta, que se debe habilitar en
# Console, Archivo, cambiar el dir…
           NBI  analfab  desempleo   PIBperc
Atlantico  16.1      4.5       12.8      1.67
Bolivar    30.0      9.5        9.8      8.59
Cesar      35.7     13.7        6.2      1.36
Cordoba    35.8     15.8       13.6      1.20
Guajira    37.5     14.4        5.9      3.02
Magdalena  30.6     13.4        7.1      0.75
Sucre      42.4     15.3        6.2      0.80

# Lectura de datos
datos<-read.table("caribe.txt",header=TRUE); datos
#Matriz de correlación
cor(datos)
#El Análisis en componentes principales estandarizado
acp<-dudi.pca(datos,scann=FALSE,nf=3); acp
#Datos estandarizados
acp$tab
#Plano factorial filas-individuos
planfac(acp, Tcol=F)
#Plano factorial columnas-variables
s.corcircle(acp$co)
#Ayudas a la interpretación del ACP
dudi.tex(acp,job="variables")
#Valores propios
acp$eig
sum(acp$eig)
```

ANÁLISIS EN CORRESPONDENCIAS SIMPLES

```
# Activación de paquetes

library(ade4)
library(xtable)
library(FactoClass)


# lectura de datos
datos<-read.table("acs.librodp.txt",header=TRUE);datos

           A     B     C     D     E
finca1    15    54   231   149     0
finca2    30    29   126    51     1
finca3    12    51   533   125     0

#Gráficos
score(acs)
mosaicplot(t(datos),color=rainbow(5),main="Perfiles columna")
mosaicplot(datos,color=rainbow(12),main="Perfiles fila")
#Análisis de correspondencias simples como ACP
acs<-dudi.coa(datos,scann=FALSE); acs
#Planos factoriales de filas, columnas, filas-columnas
plot.dudi(acs, Tcol=FALSE); plot.dudi(acs, Trow=FALSE)
plot.dudi(acs)
#Ayudas a la intepretación
dudi.tex(acs,job="res.libroacs")
#Datos estandarizados
acs$tab
#Suma de valores propios
sum(acs$eig)
```


ANÁLISIS EN CORRESPONDENCIAS MÚLTIPLES

```
# Activación de paquetes
library(ade4)
library(xtable)
library(FactoClass)

# lectura de datos
Datos<-read.table("spvdp.txt", header=TRUE); Datos

        FINCA    CRUCE       NP    EPOCA    SEXO
Vaca1   GA-GF    indicus     n1    INV      M
Vaca2   GA-GF    indicus     n2    VER      M
Vaca3   GA-GF    indicus     n2    VER      M
Vaca4   GA-GF    indicus     n3    INV      M
Vaca5   GA-GF    indicus     n4    VER      M
Vaca6   GA-GF    indicus     n3    INV      M
Vaca7   GA-GF    indi-crio   n4    VER      M
Vaca8   G07      indi-euro   n3    VER      M
Vaca9   G07      indi-euro   n4    VER      H
Vaca10  GA-GF    indi-euro   n1    INV      H
Vaca11  GA-GF    indi-euro   n3    INV      M
Vaca12  GA-GF    indi-euro   n4    INV      H
Vaca13  GA-GF    criollo     n2    INV      M
Vaca14  GA-GF    indi-crio   n2    INV      H
Vaca15  GA-GF    indi-crio   n4    INV      H
Vaca16  GA-GF    indi-crio   n2    INV      M
Vaca17  GA-GF    indi-euro   n3    INV      H
Vaca18  GA-GF    indicus     n2    INV      M
Vaca19  GA-GF    indi-euro   n3    VER      M
Vaca20  G07      indi-euro   n3    VER      M

# Tabla Disyuntiva Completa
acm.disjonctif(Datos)
# Tabla de Burt
acm.burt(Datos,Datos)
# Análisis de Correspondencias Múltiple como un ACP
acm<-dudi.acm(Tabla,scann=FALSE,nf=5); acm
# Valores propios
inertia.dudi(acm)
par(mfrow=c(1,2))
barplot(acm$eig);plot(acm$eig)
# gráficas
scatter.acm(acm) # subnubes en el primer plano
#Primer plano factorial de filas-columnas, columnas
plot.dudi(acm); plot.dudi(acm, Trow=F)
# Ayudas a la interpretación
dudi.tex(acm, job="ayudas.acm")
```

BIBLIOGRAFÍA

Abdessemed, L. & Escofier, B. (1992). Généralisation de l'analyse factorielle multiple a l'étude des tableaux de fréquence et comparaison avec l'analyse canonique des correspondances. *Technical Report 688, INRIA*.

Aguilera, S., Quiroz, V. & Calderón, R. (1999). *Manejo de ganado bovino de doble propósito en el trópico*. Libro Técnico Núm. 5. Veracruz, México: INIFAP CIRGOC.

Alfaro, C., Aranguren, C., Clavijo, A. & Díaz, C. (2004). Prevalencia serológica de leptospirosis en ganado doble propósito del noreste de Monagas, Venezuela. *Zootecnia Tropical, 22*(2)*,* 117-124.

Alfaro, C., De Rolo, M., Clavijo, A. y Valle, A. (2006). Caracterización de la paratuberculosis bovina en ganado doble propósito de los llanos de Monagas, Venezuela. *Zootecnia Tropical, 24*(3)*,* 321-332.

Alfaro, C., Díaz, C. & Tirado, H. (1999). Caracterización sanitaria de la ganadería doble propósito en el municipio Ezequiel Zamora del estado Monagas-Venezuela. *Veterinaria Tropical, 24*(2), 103-119.

Alves, E. (2013). Recomendaciones para el desarrollo de la altillanura colombiana basado en lo hecho en Brasil. Periódico El Tiempo, 25 de enero de 2013. Bogotá, Colombia.

Anderson, T.W. (1984). *An introduction to multivariate analysis.* Nueva York: Editorial Wiley.

Arango L. (1986). La ganadería de doble propósito. Coyuntura Agropecuaria: *CEGA, 1*(2), 131-137.

Aranguren, J., Román, R., Villasmil, Y. & Yánez, F. (2007). Evaluación genética de la ganadería mestiza doble propósito en Venezuela. XX Reunión ALPA, XXX Reunión APPA-Cusco-Perú. *Arch. Latinoam. Prod. Anim., 15*(S1), 241-250.

ASODOBLE. Asociación Colombiana de Criadores de Ganado en Doble Propósito (1992). *Ganadería de doble propósito: la solución para el tercer mundo*. *Memorias I y II. Encuentro agropecuario en Cartagena*. Botero M., R., Botero A., L. M., Castañeda N., N., Montoya E., H., Botero B, J. B., Londoño, E., López, L. F. & Álvarez, O. Cartagena, Colombia.

Avilez, J., Escobar, P., Von Fabeck, G., Villagrán, K., García, F., Matamoros, R. & García, A. (2010). Caracterización productiva de explotaciones lecheras empleando metodología de análisis multivariado. *Revista Científica FCV-LUZ, 20*(1), 74-80.

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.

Blench, R. (2001). *You can't go home again. Pastoralism in the new millennium*. ODI-FAO. This version: London, 17 May [En línea] <http://www.org.odi.uk/staff/r.blench>

Botero, L. M. & Rodríguez, D. (2006). Costo de producción de un litro de leche en una ganadería del sistema doble propósito. Magangué, Bolívar. *Revista MVZ Córdoba, 11*(2), 806-815.

Botero, L. M. & Vertel, M. (2006). Modelo matemático aplicado a la curva de lactancia en ganado vacuno doble propósito. *Revista MVZ Córdoba, 11*(1), 806-815.

Botero, L. M. & Vertel, M. L. (2007). Curva de ganancia de peso del nacimiento al destete de crías macho vacunas, del sistema doble propósito. *Revista Notas Ganaderas, 1*(20), 8-10.

Botero, L. M., Vertel, M. L., Flórez, L. & Medina, J. (2012). Calidad composicional e higiénico-sanitaria de la leche cruda entregada en época seca por productores de Galeras, Sucre. *Revista Facultad de Química Farmacéutica Universidad de Antioquía (VITAE), 19*(S1), 314-316.

Botero, L. M., Vertel, M. L., Flórez, L. & Medina, J. (2012). *Caracterización multivariada de la leche cruda*. España: Editorial Académica Española.

Botero, L. M., Vertel, M. L. & Rodríguez, E. (2014). Modelos no lineales para estimación de curva de crecimiento de crías bovinas machos. *Revista Facultad Nacional de Agronomía Medellín, 67*(S2), 1012-1014.

Botero, R. (2010). Desaparecerá el doble propósito. *Infortambo Andina, 24*. Octubre. Bogotá, Colombia.

Botero, R. (2010a). La fertilidad es la clave. *Carta Fedegan*, (117). Bogotá, Colombia.

Botero, R. (2011). La selección en el doble propósito debe ser integral. *Carta Fedegan*, (123), marzo-abril. Bogotá, Colombia.

Botero, R. (2011). Sin olvidar la base. *Infortambo Andina*, (33). Julio. Bogotá, Colombia.

Bourges, H., Bengoa, J. & O'Donnell, A. (2001). *Historia de la nutrición en América Latina*. Buenos Aires, Argentina: SLAN (Sociedad Latinoamericana de Nutrición) –Fundación Cavendes– CESNI (Centro de Estudios sobre Nutrición Infantil)-INCMNSZ.

Cabarcas, G. & Pardo, C. E. (2001). *Métodos estadísticos multivariados en investigación social*. Cursillo, Simposio de Estadística-Santa Marta: Universidad Nacional. Departamento de Estadística. *http://es.geocities.com/socccadcolombia/documentos/cepardot/MetEstMulInvSoc.html

Cabrera, K. R. (2002). Aplicaciones en ciencias ambientales y del suelo utilizando el lenguaje estadístico R. En *Memorias Simposio de Estadística 2002: Estadística Aplicada a las Ciencias Ambientales*. Bogotá: Universidad Nacional de Colombia. Facultad de Ciencias. Departamento de Estadística.

Cámara Gremial de la Leche (2011). Modelos *de Producción Competitivos sostenibles en Ganadería Bovina Lechería Especializada*. Segunda reunión anual. Bogotá: FEDEGAN.

Campo, D., Discuviche, M., Blanco, P., Montero, Y., Orozco, K. & Assia, Y. (2014). Detección de *toxoplasma gondii* por amplificación del gen B1 en carnes de consumo humano. *Infectio, 18*(3), 93-99.

Canavos, G. (1987). *Probabilidad y estadística: Aplicaciones y métodos*. México: Editorial McGraw-Hill.

CEGA –Centro de Estudios Ganaderos y Agrícolas– (1998). *Del proteccionismo a la apertura, ¿el camino a la modernización agropecuaria?* Balcázar V., Á., Vargas, A. & Orozco A., M. L. Bogotá: Editorial Finagro en coedición con Tercer Mundo Editores.

CEPAL, FAO, IICA (2013). *Perspectivas de la agricultura y del desarrollo rural en las Américas: una mirada hacia América Latina y el Caribe*. Santiago, Instituto Interamericano de Cooperación para la Agricultura (IICA), San José de Costa Rica.

CEPAL-Naciones Unidas (2010). *Panorama del desarrollo territorial en América Latina y el Caribe*. Santiago de Chile: Documento proyecto.

CEPAL- Naciones Unidas (2011). *Transformaciones rurales en América Latina y sus relaciones con la población rural*. Reunión de expertos sobre población, territorio y desarrollo sostenible. Santiago de Chile.

Contexto ganadero (2013). Conozca cuál es la ventaja competitiva de las vacas del país. FEDEGAN. Bogotá. [En línea] http://www.contextoganadero.com/ganaderia-sostenible/conozca-cual-es-la-ventaja-competitiva-de-las-vacas-del-pais.

Correa, J. C. & Salazar, J. C. (2000). *R: Un lenguaje estadístico*. Medellín, Colombia: Editorial Universidad Nacional de Colombia.

Crivisqui, E. (1993). *Análisis Factorial de Correspondencias*. Asunción, Paraguay: Editorial de la Universidad Católica de Asunción.

Cuadrado, H., Mejía, F., Contreras, A., Romero, A. & García, F. (2003). *Manejo agronómico de algunos cultivos forrajeros y técnicas para su conservación en la región Caribe colombiana.* Córdoba, Colombia: CORPOICA.

Cuadrado, H., Torregroza, L. & Garcés, J. (2005). Producción de carne con machos de ceba en pastoreo de pasto híbrido mulato y *Brachiaria Decumbens* en el valle del Sinú. *Revista MVZ Córdoba, 10*(1), 573-580.

Chatfields, C. & Collins, A.J. (1980). *Introduction to multivariate analysis.* New York: Editorial Chapman & Hall.

Chernoff, H. (1973). Using faces to represent points in k-dimensional space graphically. *Journal of the American Statistics Association, 68*, 361-368.

Chessel, D. (1992). *Echanges interdisciplinaires en analyse des données écologiques*. Mémoire d'Habilitation a Dirigé des Recherches. Lyon: Université Lyon I.

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.

Chessel, D., Dufour, A. B. & Thioulouse, J. (2004), The ade4 Package - I: One-table Methods. *R News, 4*(1), 5-10.

Departamento Administrativo Nacional de Estadística-DANE-SISAC (2000). *Encuesta Nacional Agropecuaria - Resultados 1999*. Bogotá.

Departamento Administrativo Nacional de Estadística-DANE-SISAC-DNP (2003). *Encuesta Nacional Agropecuaria 2002*. Bogotá.

Departamento Nacional de Planeación –DNP– (2005). *Portal web del Departamento Nacional de Planeación*. Web. *http://www.dnp.gov.co.

Díaz, L. G. & Morales, M. (2012). *Análisis de Datos Multivariados.* Bogotá, Colombia: Editorial Universidad Nacional de Colombia.

Díaz, L. R. (2002). Planes de desarrollo local: enfoques y tendencias en América Latina. *Rev. Inst. Investig. Fac. Minas Metal Cienc. Geog, 5*(10), 58-64.

Dolédec, S. & Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biology* (31), 277-294.

Dray, S. & Chessel, D. (2003). *Elèments d'interface entre analyses multivariées, systèmes d'information geographique et observations écologiques*. PhD thesis, Universite Claude Bernard - Lyon 1.

Dray, S., Chessel, D. & Thioulouse, J. (2003). Co-inertia analysis and the linking of ecological data tables. *Ecology, 84*(11), 3078-3089.

Dray, S. & Dufour, A-B. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software, 22*(4), 1-20.

Encuesta Nacional Agropecuaria –ENA– (2009). Sistema de información de la oferta agropecuaria, forestal, pesquera y acuícola. Colombia. ISBN 2027-3959.

Escofier, B. & Pagès, J. (1988-1998). *Análisis factoriales simples y múltiples. Objetivos, métodos e interpretación*. Bilbao: Editorial Universidad del País Vasco.

FAO - Departamento de Desarrollo Económico, Banco Mundial (2001). Sistemas de producción agropecuaria y pobreza. Cómo mejorar los medios de subsistencia de los pequeños productores en un mundo cambiante. ISBN 92-5-104627-1.

FAO (2010). La producción láctea a pequeña escala; una vía para salir de la pobreza. CP-56-2010. Comunicado de prensa. España.

Faye, B., Lescourret, F., Dorr, N., Tillard, E., MacDermott, B. & McDermott, J. (1997). Interrelationships between herd management practices and udder health status using canonical correspondence analysis. *Preventive Veterinary Medicine, 32*(1), 171-192.

FEDEGAN - Fondo de Estabilización de Precios –FEP– (2010*). Lo que usted necesita saber sobre la leche en Colombia*. Bogotá.

FEDEGAN - FNG (2012). *Plan Estratégico de la Ganadería Colombiana 2019*. Bogotá.

FEDEGAN - FNG (2013). *Análisis del inventario ganadero colombiano*. *Comportamiento y variables explicativas*. Bogotá.

Fernández, F. (2002). "El uso del Análisis de Correspondencia Simple (ACS) como ayuda en la interpretación del dato en arqueología. Un caso de estudio". *Boletín Antropológico, 20*(55), 687-713.

Fine, J. (1996). Iniciación a los análisis de datos multidimensionales a partir de ejemplos. Folleto: PRESTA: Programme de recherche et a enseignement en statistique appliquée, Sao Carlos.

Fisher, R. (1940). The precision of discriminant functions. *Annals of Eugenics, 10*, 422-429.

Freitas, A. (2005). Curvas de crescimento na produção animal. *Rev. Bras. Zootecn., 34*, 786-795.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika, 58*, 453-467.

García, J., Cuesta, M. & Pedroso, R. (2005). Administración de sulfato de cobre sobre la hemoquímica, hematología y bioactividad del líquido ruminal en vacas. *Revista MVZ Córdoba*, *10*(2), 639-647.

Greenacre, M. (2007). *Correspondence Analysis in Practice*, 2 edn. London: Chapman & Hall/CRC.

Gross, J. (2012). Nortest: Tests for Normality. [Internet]. Polonia: R package 1.0-2 [citado 2012]. Disponible en: http://cran.r-project.org/web/packages/nortest/index.html

GTZ - Deutsche Gesellschaft Fuer Teechnische Zusammenarbeit- Centro Internacional de Agricultura Tropical (CIAT) (1985). *Informe Técnico N5. Sistemas de Producción de Leche y Carne en Fincas Ganaderas en la Costa Atlántica de Colombia*. Alemania: Eschborn.

Harris, R. (1967). *A primer of multivariate statistics.* Nueva York: Ed. Academic.

Herrera, M., Soto, Á., Urrego, V., Rivera, G., Zapata, M. & Ríos, L. (2008). Frecuencia de hemoparásitos en bovinos del bajo Cauca y alto San Jorge, 2000-2005. *Revista MVZ Córdoba*, *13*(3), 1486-1494.

Herrero, M., Solano, C., Bernues, A., Ugarteche, J. & Rojas, F. (1998). *Caracterización Preliminar de los Sistemas de Producción de Leche y Doble Propósito en la Región de Sara e Ichilo*. pp. 86-95. Metodologías de Investigación Pecuaria en Sistemas de Producción de Pequeños Productores, el CIAT, Bolivia. Proyecto CIAT-IERM Edimburgo.

Hess, D., Matsumoto, A., Kim, S., Marshall, H. & Stamler, J. (2005). Protein S-nitrosylation: purview and parameters. *Nature Review of Molecular Cell Biology,* 6, 150-166.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Rev. J. Educ. Psychol.*, 24, 417-441.

Husson, F., Lê, S. & Pagés, J. (2011). Exploraty multivariate analysis by example using R. London, UK: Ed. Chapman & Hall/CRC, Computer Science and Data Analysis Series.

Jiménez, J. A. (2012). *Álgebra matricial con aplicaciones en estadística*. 2ª edición. Bogotá, Colombia: Editorial Universidad Nacional.

Johnson, D. (2000). *Métodos multivariados aplicados al análisis de datos*. México: Editorial Thomson.

Kendall, M.G. (1980). *Multivariate analysis*. Londres: Ed. Griffin.

Koger, M., Peacock, E., Kirk, W. & Crockett, J. (1975). Heterosis effects on Weaning performance of Brahman-Shorthorn calves. *J. Anim Sci. 40.* doi:10.2134/jas1975.405826x

Komsta, L. & Novomestky, F. (2012*). Moments: Moments, cumulants, skewness, kurtosis and related tests*. [Internet]. Polonia: R package 0.13 [citado 2012]. Disponible en: http://www.komsta.net/

Koppel, R., Ortiz, O., Ávila, A., Lagunes, L., Castañeda, M., López, G., Aguilar, B., Román, P., Villagómez, C., Aguilera, S., Quiroz, V. & Calderón, R. (1999). *Manejo de ganado bovino de doble propósito en el trópico*. Libro técnico Num. 5. Veracruz, México: INIFAP. CIRGOC.

Kuelh, R. O. (2001). *Diseño de experimentos*. México: Thomson Editores.

Lebart, L., Morineau, A. & Piron, M. (1995). *Statisitique exploratoire multidimensionnelle*. Dunod, Paris.

Lebreton, J., Chessel, D., Prodon, R. & Yoccoz, N. (1988). L'analyse des relations espèces milieu par l'analysés canonique des correspondances; i.- variables de milieu quantitatives. *Acta Ecologique, 9*(1), 53-67.

Lebreton, J., Sabatier, R., Banco, G. & Bacou, A. M. (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. Dans Karcher W., ed. *Applied Multivariate Analysis in SAR and Environmental Studies* (pp.85-114). Kluwer Academic Publishers.

Madelena, F. (1986). Utilización de recursos genéticos en programas de mestizaje en el trópico. En: Seminario internacional sobre sistemas de producción bovina de doble propósito en el trópico. Bogotá, Colombia.

Magaña, J. (1995). *Genetic effects on productive efficiency of dual purpose cattle*. Procceding of the International Workshop. Facultad de Medicina Veterinaria y Zootecnia, Universidad Autónoma de Yucatán/Internacional Foundation for Science. Eds: Anderson, S. and J. Wadsworth. Dual Purpose Cattle Production Reseach. Mérida, México.

Maldini, P. & Peixoto, M. (2011). ¿Bueno de leche y bueno de carne es posible? *Rev. Venezuela Bovina*. [en línea] http://www.ganadof1.com/articulos_publicados.htm.

Mardia, K., Kent, J. & Bibby, J. (1982). *Multivariate analysis*. Londres: Ed. Academic.

Martínez, G., & Brandi, M. (1997). Factores que afectan las pérdidas en un rebaño doble propósito. *Rev. Arch. Latinoam. Prod. Anim., 5*(S1), 488-490.

Martínez, J. (2013). Director Ejecutivo de ASOLECHE, entrevista al periódico *El Colombiano*, Medellín, Colombia.

Medina, J. (2005). Caracterización bovino-métrica de hembras cebú y cruces con blanco Orejinegro, Romosinuano y Angus. *Revista MVZ Córdoba, 10*(1), 581-588.

Mendiburu F. (2010). *Agricolae: This package contains functionality for the statistical analysis of experimental designs applied specially for field experiments in agriculture and plant breeding* [Internet]. Lima, Perú: R package 1.07. 2009 [citado 2010]. Disponible en: http://tarwi.lamolina.edu.pe/~fmendiburu

Ministerio de Agricultura y Desarrollo Rural (MADR) y Corporación Colombia Internacional (CCI) (2009). *Oferta Agropecuaria, Encuesta Nacional Agropecuaria 2009*. [En línea] http://www.agronet.gov.co/www/docs_agro-net/201046112648_ Resultados_Ena_2009.pdf

Molinuevo, H. (2003). Productividad del sistema y potencial genético en rodeos de cría. *Rev. Hereford, 67*(631), 14-25.

Montero, E. (2013). *Situación actual y perspectivas del sector lácteo a nivel mundial*. Congreso Nacional Lechero. Cámara Nacional de Productores de Leche. Costa Rica.

Morrison, D. (1976). *Multivariate statistical methods*. Nueva York: Editorial McGraw.

Murcia, L. & Martínez, G. (2013). Factores que afectan la vida útil de vacas doble propósito. *Revista MVZ Córdoba, 18*(2), 3459-3466.

Noy-Meir, I. (1973). Data transformation in ecological ordination. I. Some advantages of non-centring. *Journal of Ecology, 61*, 329-341.

Noy-Meir, I., Walker, D. & Williams, W. (1975). Data transformation in ecological ordination. II. On the meaning of data standardization. *Journal of Ecology, 63*, 779-800.

Observatorio del Caribe Colombiano (2012). *Programa de asistencia técnica a la Comisión Regional de Competitividad del departamento de Sucre*.

[Internet]. Disponible en: http://ocaribe.org/departamentos.php?la=vh-qificl&id=8

OCDE/FAO (2013). *OCDE-FAO Perspectivas Agrícolas 2013-2022*. Texcoco, Estado de México, Universidad Autónoma Chapingo. [En línea] http://dx.doi.org/10.1787/agr_outlook-2013-es

Osorio, F. (2013a). *Costos de producción de leche bajo el nuevo escenario del sector lácteo nacional-análisis comparativo con Brasil.* Medellín Colombia: Empresa FINCA.

Osorio, F. (2013b). *Competitividad de las cuencas lecheras especializadas en Colombia y Brasil*. Medellín, Colombia: Concentrados FINCA S.A.

Páez, L., López, N., Salas, K., Spaldiliero, A. & Verde, O. (2002). Características físico-químicas de la leche cruda en las zonas de Aroa y Yaracal, Venezuela. *Rev.Científica, FCV-LUZ, 12*(2), 113-120.

Pardo, C-E. (2009). *Geometría euclidiana en estadística: métodos en ejes principales.* [Internet]. UNAL sede Bogotá. Disponible en: http://www.docentes.unal.edu.co/cepardot/docs/Conferencias/ACPgeometriaEuclidiana.pdf

Pardo, C-E. & Del Campo, P-C. (2007). Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass. *Revista Colombiana de Estadística, 30*(2), 231-245.

Pardo, C-E. & Ortiz, J. (2004). *Análisis multivariado de datos en R*. Simposio de Estadística 2004. Universidad Nacional de Colombia, Bogotá.

Pardo, C-E., Becuè, M. & Ortiz, J. (2012). *Métodos en ejes principales para tablas de contingencia con estructuras de partición en filas y columnas*. Tesis de Doctorado en Ciencias Estadísticas, Universidad Nacional de Colombia, sede Bogotá.

Pearson, K. (1901). On lines and planes of closed fit to system of point in space. *Phil. Mag., 6*, 559-572.

Peña, D. (2002). *Análisis de datos multivariados*. España: Editorial Mc-Graw-Hill/Interamericana.

Pérez, P., Rojo, A., García J., Ávila, C. & López, S. (2003). *Necesidades de investigación y transferencia de tecnología de la cadena de bovinos de doble propósito del estado de Veracruz*. México: Fundación Produce Veracruz.

Pérez, R., Pérez, A. & Vertel, M. (2010). Caracterización nutricional, físico-química y microbiológica de tres abonos orgánicos para uso en agro ecosistemas de pasturas en la subregión Sabanas del departamento de Sucre. *Revista Tumbaga, 5*, 27-37.

Pla, L. (1986). *Análisis multivariado: método de componentes principales*. Monografía 27, Serie de Matemática. OEA, Washington.

Plasse, D. (1992). *Cruzamiento en bovinos de carne en América Latina Tropical: qué sabemos y qué nos falta saber*. III Simposio Nacional de Melhoramento Animal. Universidad Central de Venezuela. Facultad de Ciencias Veterinarias. Maracay, Venezuela. 165-179. [En línea] http://sbmaonline.org.br/anais/iii/palestras/pdfs/iiip20.pdf

Preston, T. (1976). Prospets for the intensification of cattle production in developing countries. A.J. Smith (De.), *Beef Cattle Production in Developing Countries* (pp.242-257). Scotland: University of Edimburgh Press.

Puricelli, E. (2011). Las carnes en el mundo. *Rev. Brangus, 33*(63), 60-64.

R Development Core Team (2014). *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.

Rearte, D. (2007). *Situación de la ganadería argentina en el contexto mundial. Programa Nacional de Carnes.* [En línea] http://www.inta.gov.ar/balcarce/infoindices/tematica/ganad/bovi/carne.htm

Ritchie, D., Neves, C., Támara, A., Begazo, O., Igor, V. & Uribe, J. (2013). *Ganadería de doble propósito: propuesta para pequeños productores colombianos*. Lima: Universidad ESAN (Serie Gerencia para el Desarrollo, 33).

Román, S., Ruiz, F., Montalvo, H., Rizzi, R. & Román, H. (2013). Efectos de cruzamiento para producción de leche y características de crecimiento en bovinos de doble propósito en el trópico húmedo. *Rev. Mex. Cienc Pecu., 4*(4), 405-416.

Shapiro, S. & Wilk, M. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Rev. Biometrika, 52*(3-4), 1-591.

Shneichel & Sebert (1990). Alternativas de Alimentación. Primer curso sobre ganado doble propósito ICA – GTZ. Montería, Colombia. Publicado en el libro *Génesis y consolidación del sistema vacuno doble propósito*. ASODOBLE.

Silva, J., Pulido, J., Ballesteros, H., Abuabara, Y., Benavides, E., Rodríguez, G., Roncallo, B., Abadía, B. & Molina, J. (2011). *Modelos tecnológicos y calidad de la leche en sistemas bovinos de doble propósito de la región Caribe*. Bogotá, Colombia: Ed. Corpoica.

Smith, R., Moreira, V. & Latrille, L. (2002). Caracterización de sistemas productivos lecheros en la X región de Chile mediante Análisis Multivariable. *Rev. Agric. Téc*. [Online], *62*(3). http://dx.doi.org/10.4067/S0365-28072002000300004

Tatis, R. & Botero, L.M. (2005). *Génesis y consolidación del sistema vacuno en doble propósito*. Bogotá, Colombia: ASODOBLE (Asociación Colombiana de Criadores de Ganado en Doble Propósito).

Tenenhaus, M. & Young, F. (1985). An analysis and synthesis of multiple correspondence analyses, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika, 50*, 91-119.

Ter-Braak, C. (1986). Canonical correspondence analysis: A new technique for multivariate direct gradient analysis. *Ecology, 67*(5), 1167-1179.

Thioulouse, J., Chessel, D., Doledec, S. & Olivier, J. (1997). Ade-4: a multivariate analysis andgraphical display software. *Stat. Comp., 7*, 75-83. *http://pbil.univ-lyon1.fr/ADE-4/ADE-4F.html

Thioulouse, J., Dufour, A. & Chessel, D. (2004). *ADE4: Analysis of Environmental Data: Exploratory and Euclidean method Multivariate data analysis and graphical display*. Lyon, Francia. * http://pbil.univ-lyon1.fr/JTHome/ref/ade4-Rnews.pdf

Vejarano, A., Sanabria, R. & Trujillo, G. (2005). Diagnóstico de la capacidad reproductiva de toros en ganaderías de tres municipios del alto Magdalena. *Revista MVZ Córdoba*, *10*(2), 648-662.

Vertel, M. & Pardo, C.E. (2010). *Comparación entre el análisis canónico de correspondencias y el análisis factorial múltiple en tablas de frecuencias-variables continuas*. Master's thesis, Universidad Nacional de Colombia, sede Bogotá.

Vertel, M. (2005). *Diseño y análisis de experimentos en Ciencias Agroindustriales*. Trabajo de Promoción para Ascenso Categoría Docente, Universidad de Sucre, Sincelejo (Sucre).

Vertel, M. (2012). *Complementariedad de técnicas multivariadas para análisis de datos.* España: Editorial Académica Española.

Weimer (1999). *Estadística*. México: Editorial CECSA.

Wettemann, R. & Bossis, I. (2000). Nutritional regulation of ovarian function in beef cattle. *J. Anim. Sci*. [En línea] http://www.asas.org/jas/symposia/proceedings/0934.pdf

Williams, E. (1952). Use of scores for the analysis of association in contingency tables. *Biometrika, 39*, 274-289.

Williams, J., Kubelik, A., Livak, K., Rafalski, J. & Tingey, S. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res., 18*, 6531-6535.