Reducción de la Dimensionalidad en Análisis de Datos. Análisis de Componentes Principales.

A. Jiménez, A. Murillo,
E. Piza, M. Villalobos, J. Trejos.

April 27, 2010

Contenido

Contenido

Contenido

Contenido

- ACP
  1. Objetivo.
  2. Solución.
  3. Representación gráfica. Calidad de representación.
  4. Interpretación.
- Aplicación: Rendimiento académico

Contenido

- ACP
  1. Objetivo.
  2. Solución.
  3. Representación gráfica. Calidad de representación.
  4. Interpretación.
- Aplicación: Rendimiento académico

Contenido

- ACP
  1. Objetivo.
  2. Solución.
  3. Representación gráfica. Calidad de representación.
  4. Interpretación.
- Aplicación: Rendimiento académico

ACP: Introducción

El Análisis de Componentes Principales (ACP) es una técnica multivariable mediante la cual se analiza una tabla de datos $\mathbf{X}$ que contiene p observaciones cuantitativas realizadas a n individuos.

Mediante el ACP es posible definir otras técnicas como Análisis de Correspondencias (AC) y el Análisis Factorial Discriminante (AFC).

ACP: Introducción

El objetivo del ACP es extraer información importante de la tabla y representarla, mediante una cantidad menor de variables sintéticas (nuevas) a fin de hallar la relación entre las variables originales.

Estas nuevas variables son combinación lineal de las variables originales.

ACP: Objetivo

Los objetivos del ACP son:

- Extraer la información más importante de la tabla de datos.
- Reducir el tamaño del conjunto de datos.
- Simplificar la descripción del conjunto de datos.
- Analizar la estructura y relación de las observaciones y de las variables.

Se desea una pérdida mínima de información al proyectar los n individuos sobre un espacio de dimensión q con $q < p$ tal que la dispersión en el espacio proyectado $\mathbb{R}^q$ debe ser máxima.

ACP: Objetivo

Los objetivos del ACP son:

- Extraer la información más importante de la tabla de datos.
- Reducir el tamaño del conjunto de datos.
- Simplificar la descripción del conjunto de datos.
- Analizar la estructura y relación de las observaciones y de las variables.

Se desea una pérdida mínima de información al proyectar los n individuos sobre un espacio de dimensión q con $q < p$ tal que la dispersión en el espacio proyectado $\mathbb{R}^q$ debe ser máxima.

ACP: Objetivo

Los objetivos del ACP son:

- Extraer la información más importante de la tabla de datos.
- Reducir el tamaño del conjunto de datos.
- Simplificar la descripción del conjunto de datos.
- Analizar la estructura y relación de las observaciones y de las variables.

Se desea una pérdida mínima de información al proyectar los n individuos sobre un espacio de dimensión q con $q < p$ tal que la dispersión en el espacio proyectado $\mathbb{R}^q$ debe ser máxima.

ACP: Objetivo

Los objetivos del ACP son:

- Extraer la información más importante de la tabla de datos.
- Reducir el tamaño del conjunto de datos.
- Simplificar la descripción del conjunto de datos.
- Analizar la estructura y relación de las observaciones y de las variables.

Se desea una pérdida mínima de información al proyectar los n individuos sobre un espacio de dimensión q con $q < p$ tal que la dispersión en el espacio proyectado $\mathbb{R}^q$ debe ser máxima.

ACP: Objetivo

A fin de lograr estos objetivos, mediante el ACP se construyen nuevas variables llamadas *componentes principales*, tal que:

- Cada componente principal debe ser combinación lineal de las variables originales. La información contenida en la variable se ve reflejada en la componente.
- Las componentes principales son no correlacionadas dos a dos, eliminándose información redundante.
- Las componentes principales deben tener varianza máxima, conteniendo el máximo de información posible.

ACP: Objetivo

A fin de lograr estos objetivos, mediante el ACP se construyen nuevas variables llamadas *componentes principales*, tal que:

- Cada componente principal debe ser combinación lineal de las variables originales. La información contenida en la variable se ve reflejada en la componente.
- Las componentes principales son no correlacionadas dos a dos, eliminándose información redundante.
- Las componentes principales deben tener varianza máxima, conteniendo el máximo de información posible.

ACP: Objetivo

A fin de lograr estos objetivos, mediante el ACP se construyen nuevas variables llamadas *componentes principales*, tal que:

- Cada componente principal debe ser combinación lineal de las variables originales. La información contenida en la variable se ve reflejada en la componente.
- Las componentes principales son no correlacionadas dos a dos, eliminándose información redundante.
- Las componentes principales deben tener varianza máxima, conteniendo el máximo de información posible.

ACP: Solución

Sea $\mathbf{X}$ la tabla de datos definida con p variables cuantitativas tal que $\mathbb{R}^p$ es el espacio de individuos y $\mathbb{R}^n$ es el espacio de variables.

	x^1	$\ldots$	x^p
x_1	x_1^1	$\ldots$	x_1^p
x_2	x_2^1	$\ldots$	x_2^p
$\vdots$	$\vdots$		$\vdots$
x_n	x_n^1	$\ldots$	x_n^p

ACP: Solución

Si $\mathbf{M}$ es la métrica sobre $\mathbb{R}^p$ y $\mathbf{D}_p$ la métrica de pesos sobre $\mathbb{R}^n$, entonces:

$$N = (\mathbf{X}, \mathbf{M}, \mathbf{D}_p)$$

representa la nube de n puntos ponderados del espacio vectorial $\mathbb{R}^p$, junto con las medidas de proximidad y angular definidas por $\mathbf{M}$, y las medidas de tendencia central y de dispersión asociadas a $\mathbf{D}_p$.
El concepto de nube de puntos N es geométrico, cuya forma se busca describir y sintetizar mediante métodos estadísticos.

ACP: Solución

Definición

Se denomina ACP normado al análisis de componentes principales realizado a una tabla de datos $\boldsymbol{X}$ en la cual **las variables** *están centradas y estandarizadas.*

De esta forma todas tienen media cero y varianza 1.

ACP: Solución

En el ACP normado:

- La matriz de covarianzas $\mathbf{V}$ y la matriz de correlaciones $\mathbf{R}$ coinciden.
- La inercia de N es igual a la traza de $\mathbf{V}$, la cual es igual a la suma de las varianzas. En este caso, esta suma es exactamente p.

ACP: Solución

Solución matemática:

Dada la tabla de datos $\mathbf{X}$, la solución del ACP se obtiene al diagonalizar su matriz de correlaciones $\mathbf{R}$.

En el caso general, para una métrica cualquiera, la solución es la diagonalización de una matriz $\mathbf{VM}$, producto de la matriz de covarianzas $\mathbf{V}$ de $\mathbf{X}$ y de la métrica $\mathbf{M}$.

ACP: Solución

Al diagonalizar $\mathbf{R}$ se obtienen p valores propios denotados $\lambda_1, \lambda_2, ...\lambda_p$ (ordenados en orden decreciente) a partir de los cuales se obtienen los llamados vectores principales $u_1, u_2, ..., u_p$, donde u_i es un vector propio normado de $\mathbf{R}$ asociado al valor propio λ_i.

Siendo $\mathbf{R}$ simétrica y positiva, tiene p valores propios reales. Como es semidefinida positiva, estos valores propios son mayores o iguales que cero, pero su suma es p.

ACP: Solución

Las componentes principales serán las variables asociadas a estos ejes principales, tales que

$$c^i = \mathbf{X}u_i$$

será llamada la $i-$ésima componente principal.

Por definición, las componentes principales son combinación lineal de las variables originales (que son las columnas de $\mathbf{X}$), por lo que su media también es cero.

ACP: Solución

Al vector u_1 asociado al mayor valor propio λ_1 de $\mathbf{R}$ se le llama el primer eje del ACP de la nube N.

ACP: Solución

En general, los ejes principales del ACP de la nube N son los vectores propios de la matriz de correlaciones $\mathbf{X}$ asociados a los valores propios de ésta, dados en orden decreciente:

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$$

asociados respectivamente a $u_1, u_2, ..., u_p$.

Seleccionando cualquier par de ejes principales, se puede generar un plano principal.

Se llama *espacio principal* a un espacio generado por un cierto número de componentes principales.

ACP: Solución

¿ Cuántas componentes?

Existen algunas guías para escoger la cantidad de componentes principales

- Graficar los valores propios de acuerdo a su magnitud (gráfica continua) y observar si existe un punto en la gráfica (*codo*) a partir del cual la tasa de decrecimiento disminuye considerablemente. Escogiéndose así los valores propios ubicados antes del *codo*.
  Este procedimiento es llamado Prueba del codo.

ACP: Solución

- Fijar un porcentaje de inercia P considerado como mínimo para que el ACP sea satisfactorio, entonces tomará el número de ejes que sea necesario hasta que la inercia explicada por ellos sobrepase P.

ACP: Solución

- Otro método es escoger las componentes asociadas los valores propios mayores a la media de los valores propios.
  Esto es escoger la componente c^k si

$$\lambda_k > \frac{1}{p}\sum_{i}^{p}\lambda_i = \frac{\text{Inercia}}{p}$$

ACP: Representación gráfica

Las componentes principales permiten hacer una representación en pocas dimensiones de los hechos más sobresalientes de una tabla de datos.
Se obtienen dos tipos de representaciones gráficas:

- Los planos principales.
- Los círculos de correlaciones

ACP: Representación gráfica

Los planos principales.

Están formados por las coordenadas de los individuos en las componentes principales

En ellos se pueden apreciar las principales agrupaciones y dispersiones de los individuos.

El plano definido por c^1, c^2 es llamado el *primer plano principal*. En general, cualquier plano definido por dos componentes principales es llamado un *plano principal*.

ACP: Representación gráfica

Los círculos de correlaciones

Son obtenidos a partir de las correlaciones entre las variables originales y las componentes principales normalizadas.

Aquí se pueden apreciar las agrupaciones de variables y su comportamiento respecto de las componentes principales.

ACP: Representación gráfica

Las dos gráficas son complementarias.

El círculo de correlaciones permite interpretar las posiciones relativas de los individuos de la misma forma que se puede apreciar para qué individuos las variables tienen grandes valores (por encima del promedio).

La construcción se obtiene calculando el coeficiente de correlación lineal entre cada variable x^j y la componente principal c_k correspondiente:

$$\text{Coordenada de la variable } \mathrm{x}^j \text{ en } c^k \text{ es } r(\mathrm{x}^j, c^k)$$

ACP: Representación gráfica

En cualquier interpretación de los gráficos, siempre debe tenerse presente que éstos no son más que simplificaciones de los hechos observados.

Cualquier hipótesis o conclusión debe ser examinada a la vista de los datos originales para verificarla o descartarla.

ACP: Calidad de representación

Primer plano:

La calidad de la representación de la nube de puntos sobre el primer plano principal se puede medir, en forma de porcentaje, por el cociente de la inercia de la nube proyectada entre la inercia total:

$$\frac{\lambda_1 + \lambda_2}{\text{Inercia}(N)}$$

ACP: Calidad de representación

Individuos:

Se tienen dos puntos a y b en la nube original. Estos se proyectan sobre el plano principal en los puntos $\mathrm{Pr}(a)$ y $\mathrm{Pr}(b)$, respectivamente.
El valor del coseno del ángulo entre la proyección y el punto mide la representación en dicho plano. A menor distancia entre el punto y el plano, mayor será el valor del coseno cuadrado del ángulo α entre a y $\mathrm{Pr}(a)$:

$$\cos^2 \alpha = \frac{|| \mathrm{Pr}(a) ||^2}{|| a ||^2}$$

ACP: Calidad de representación

Variables:

La calidad de la representación de una variable sobre el círculo de correlaciones, será también medida con el cuadrado del coseno del ángulo la variable y su proyección.

Entre variables, el valor del coseno es igual a una correlación, por lo que serán las correlaciones las que midan la calidad de la representación de las variables:

$$|| \Pr(\mathrm{x}) ||^2 = r^2(\mathrm{x}, c^1) + r^2(\mathrm{x}, c^2)$$

Esta suma de correlaciones al cuadrado se conoce con el nombre de comunalidad.

ACP: Interpretación de resultados

El arte y la experiencia juegan un papel importante en la interpretación de resultados.

Inicialmente se debe etiquetar a las componentes principales, usando las medidas de calidad de representación de los individuos y de las variables.
Para esto se usan generalmente dos criterios:

- Un eje tendrá mucha relación con aquellos individuos cuyo coseno cuadrado sea superior o igual a 0,5. Estos individuos están particularmente bien representados sobre ese eje.
- Una componente principal puede ser interpretada a partir de las variables originales que tengan con ella una correlación mayor o igual a 0,7.

Aplicación

El siguiente análisis pretende hallar relación entre el rendimiento académico en primaria y secundaria con la cantidad de graduados de universidades estatales y privadas, principalmente en el área de educación. Además se analiza su posible relación con el rendimiento académico en el curso Matemática General del ITCR, como muestra de un curso básico universitario.

Aplicación: Datos

Obtenidos en:

- Programa Estado de la Nación
  Programa de investigación y formación sobre desarrollo humano sostenible creado, en 1994, con el propósito de dotar a la sociedad costarricense de instrumentos de fácil acceso para conocer su evolución.
- Departamento de Admisión y Registro, ITCR.

Aplicación: Datos

Observaciones:

Años: 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008

Aplicación: Datos

Variables y etiquetas: medidas por porcentajes

- A12: Aprobación el I y II ciclo de educación (primaria).
- A3EDI: Aprobación el III ciclo y Educación diversificada (secundaria).
- AMGITCR: Aprobación en el curso Matemática General, ITCR.
- R1c: Repitencia en el I ciclo de educación (primaria).
- R2c: Repitencia en el II ciclo de educación (primaria).
- R3c: Repitencia en el III ciclo de educación (secundaria).
- RED: Repitencia en el ciclo de Educación Diversificada (secundaria).

Aplicación: Datos

Variables y etiquetas: medidas por porcentajes

- DUCR: Diplomas otorgados por la Universidad de Costa Rica.
- DUNA: Diplomas otorgados por la Universidad Nacional.
- DITCR: Diplomas otorgados por el Instituto Tecnológico de Costa Rica.
- DUNED: Diplomas otorgados por la Universidad Nacional Estatal a Distancia.
- DUP: Diplomas otorgados por universidades privadas.
- DEEd: Diplomas otorgados por universidades estatales en área educación.
- DPEd: Diplomas otorgados por universidades privadas en área educación.

Aplicación: Datos

Año	A12	A3EDI	R1c	R2c	R3c	RED	AMG
1998	88.3	78	12.4	7.2	12.4	6.5	51.98
1999	90.2	83.4	11.8	6.7	12.4	5.5	55.27
2000	90.6	82.7	10.6	5.6	10.3	4.9	56.73
2001	90.9	82.2	10.7	5.9	10.9	4.9	48.98
2002	91.2	81.1	9.8	5.2	11.4	7.4	52.17
2003	90.7	81.9	9.6	5.1	11.7	7.3	58.36
2004	90.5	80	9.6	5	11.7	6.5	51.29
2005	88.8	79	9.5	5.2	12.9	7.6	53.48
2006	88.7	78.4	9.6	5.5	12.9	8.1	51.43
2007	89.3	79.4	10.1	5.5	13.9	8.2	53.30
2008	93	82	9.7	4.9	13.2	7.9	54.70

Aplicación: Datos

Año	DUCR	DUNA	DITCR	DUNED	DUP	DEEd	DPEd
1998	19.6	10.6	3.9	9.8	56	17.6	14.3
1999	18.8	10.9	3.2	9.1	58	17.6	15.7
2000	14.9	9.8	3.3	8.2	63.8	14	21
2001	17.7	10.3	3.6	7.1	61.3	13	20.6
2002	15.8	10.9	4.2	8.5	60.7	15.4	20
2003	17	11.1	4.3	8.5	59.2	15.4	17.3
2004	15.5	11.9	3.8	8	60.8	14.8	19
2005	15.8	12.2	4.8	8.9	58.4	16.4	19
2006	14.7	9.9	4.7	8.1	62.5	13.7	22.4
2007	14.4	4	8.1	7.6	65.9	12	21.6
2008	13	4	7.5	6.6	68.9	10.2	22.1

Aplicación: ACP

Dada la tabla de datos $\mathbf{X}$
con variables centradas y estandarizadas, se desea proyectar la nube

$$N = (\mathbf{X}, \mathbf{I}_{14}, \mathbf{D}_{14})$$

en un espacio menor a 14, donde la distancia utilizada es la euclídea y todas las variables tienen el mismo peso.
La matriz de covarianzas $\mathbf{V}$ y la matriz de correlaciones $\mathbf{R}$ coinciden.

Aplicación: Matriz de correlaciones $\mathbf{R}$

	A12	A3EDI	R1c	R2c	R3c	RED	AMG
A12	1	0.7302157	-0.2979858	-0.4876033	-0.2768998	-0.1200279	0.2225193
A3EDI	0.7302157	1	0.0808882	-0.08271	-0.5059072	-0.569827	0.4419449
R1c	-0.2979858	0.0808882	1	0.9681292	-0.1197498	-0.5642049	-0.0670889
R2c	-0.4876033	-0.08271	0.9681292	1	-0.0246041	-0.4625257	-0.1534929
R3c	-0.2768998	-0.5059072	-0.1197498	-0.0246041	1	0.770649	-0.0387869
RED	-0.1200279	-0.569827	-0.5642049	-0.4625257	0.770649	1	0.0372285
AMG	0.2225193	0.4419449	-0.0670889	-0.1534929	-0.0387869	0.0372285	1
DUCR	-0.3960094	0.0305913	0.7710044	0.8089422	-0.2669321	-0.520111	-0.112853
DUNA	-0.3424162	-0.0402573	0.1161159	0.1919157	-0.5521634	-0.4075473	-0.0978212
DITCR	0.1739767	-0.2417302	-0.3777372	-0.3915439	0.7646247	0.7264642	0.0541238
DUNED	-0.6651999	-0.308309	0.5203676	0.6027752	-0.0570045	-0.119906	0.1658473
DUP	0.5663861	0.2026279	-0.4699321	-0.5634573	0.2414739	0.3027472	0.0783417
DEEd	-0.5452145	-0.1358592	0.4991686	0.57027	-0.2021367	-0.2838559	0.0886665
DPEd	0.3408279	0.0249519	-0.6802781	-0.6788733	0.0940716	0.3092304	-0.1773175

Aplicación: Matriz de correlaciones $\mathbf{R}$

	DUCR	DUNA	DITCR	DUNED	DUP	DEEd	DPEd
A12	-0.3960094	-0.3424162	0.1739767	-0.6651999	0.5663861	-0.5452145	0.3408279
A3EDI	0.0305913	-0.0402573	-0.2417302	-0.308309	0.2026279	-0.1358592	0.0249519
R1c	0.7710044	0.1161159	-0.3777372	0.5203676	-0.4699321	0.4991686	-0.6802781
R2c	0.8089422	0.1919157	-0.3915439	0.6027752	-0.5634573	0.57027	-0.6788733
R3c	-0.2669321	-0.5521634	0.7646247	-0.0570045	0.2414739	-0.2021367	0.0940716
RED	-0.520111	-0.4075473	0.7264642	-0.119906	0.3027472	-0.2838559	0.3092304
AMG	-0.112853	-0.0978212	0.0541238	0.1658473	0.0783417	0.0886665	-0.1773175
DUCR	1	0.5835195	-0.6507959	0.6939719	-0.8666987	0.7908047	-0.8743309
DUNA	0.5835195	1	-0.888827	0.6352133	-0.8401614	0.7928313	-0.5232377
DITCR	-0.6507959	-0.888827	1	-0.5372382	0.7218638	-0.7006103	0.5084538
DUNED	0.6939719	0.6352133	-0.5372382	1	-0.8659387	0.9475678	-0.8091744
DUP	-0.8666987	-0.8401614	0.7218638	-0.8659387	1	-0.9520258	0.841687
DEEd	0.7908047	0.7928313	-0.7006103	0.9475678	-0.9520258	1	-0.8526425
DPEd	-0.8743309	-0.5232377	0.5084538	-0.8091744	0.841687	-0.8526425	1

Aplicación: Valores propios

λ_1	7.053977
λ_2	2.9300494
λ_3	1.6828729
λ_4	1.3619466
λ_5	0.4941486
λ_6	0.188833
λ_7	0.1425764
λ_8	0.08297
λ_9	0.0549225
λ_{10}	0.0078047
λ_{11}	0.0000064
λ_{12}	0.0000026
λ_{13}	0.0000021
λ_{14}	0.0000008

Aplicación: Valores propios

Análisis de la inercia

Eje (i)	Inercia explicada $\frac{\lambda_i}{14} \cdot 100$ % $(\approx)$	Inercia acumulada $\sum_i (\frac{\lambda_i}{14} \cdot 100\%)$
1	50.38 %	50.38%
2	20.92%	71.31 %
3	12.02%	83.33%
4	9.72 %	93.06%
5	3.52 %	96.59%
6	1.34 %	97.94%
7	1.01 %	98.96%

Aplicación: Valores propios

Análisis de la inercia

Eje (i)	Inercia explicada $\frac{\lambda_i}{14} \cdot 100$ % $(\approx)$	Inercia acumulada $\sum_i (\frac{\lambda_i}{14} \cdot 100\%)$
8	0.59 %	99.55%
9	0.39 %	99.94%
10	0.05 %	99.945%
11	4.5 E^{-5}%	$\approx$100%
12	1.8 E^{-5}%	$\approx$100%
13	1.5 E^{-5}%	$\approx$100%
14	5.9 E^{-6}%	$\approx$100%

Aplicación: Elección de valores propios

Análisis de la inercia

Se escogen los 4 primeros valores propios:

- La media de los valores propios es 1, tal que

$$\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > 1$$

- La Prueba del codo respalda esta elección.
- Con las primeras 4 componentes, asociadas a los valores escogidos, se conserva un 93.06 % de la información de la nube N, lo cual se considera aceptable.

Aplicación: Vectores propios

Los cuatro vectores propios asociados a los valores propios escogidos son:

u_1	u_2	u_3	u_4
0.1950192	-0.4103434	0.0951015	-0.1365133
0.0205577	-0.517292	0.1884779	-0.2360762
-0.2656391	-0.0061629	0.5276775	0.1000856
-0.2843867	0.0906117	0.4563619	0.1469822
0.1342916	0.4717404	0.2138428	-0.1640941
0.2001268	0.4205637	-0.1632118	-0.2385828
0.0139826	-0.1386449	0.053155	-0.7874035
-0.3482507	-0.0090759	0.1665449	0.0482269
-0.2858706	-0.0985611	-0.4734735	0.0484497
0.2991357	0.267903	0.2441861	-0.125622
-0.3239821	0.1935498	-0.0878214	-0.2471155
0.3535932	-0.0871713	0.1829816	0.0444584
-0.3502404	0.0830298	-0.1483417	-0.1882279
0.3266208	-0.0601702	-0.1370712	0.2668931

Aplicación: Componentes principales

Las 4 componentes principales se obtienen de la forma

$$c^1 = \mathbf{X}u_1$$

$$c^2 = \mathbf{X}u_2$$

$$c^3 = \mathbf{X}u_3$$

$$c^4 = \mathbf{X}u_4$$

Aplicación: Componentes principales

	c^1	c^2	c^3	c^4
1998	-4.7569189	2.2167177	1.49159	0.4392034
1999	-3.5613873	-0.9264556	1.6122211	-1.0437983
2000	-0.0586433	-2.7841847	0.0496728	-0.2668964
2001	-0.3716294	-2.1903815	0.5172282	2.5092399
2002	0.066545	-0.5709851	-1.201512	0.0474773
2003	-0.5671641	-0.7840192	-0.9697949	-2.2570422
2004	0.0065034	-0.404749	-1.5556818	0.7411759
2005	-0.6844087	1.7655735	-1.7591507	-0.4973921
2006	1.2220688	1.9083076	-1.1113629	1.0076565
2007	3.3625302	2.3798122	1.4704193	-0.0307685
2008	5.3424902	-0.6096058	1.456363	-0.6488417

Aplicación: Calidad de representación

Individuos: Cosenos cuadrados

A menor distancia entre un punto de la nube N y el plano en que se proyecta, mejor será su representación y mayor será el valor del coseno cuadrado del ángulo entre el punto y su proyección.

Aplicación: Calidad de representación

Así, si un individuo x_1 se relaciona con las componentes principales c^1, c^2, ..., c^q de la forma:

	c^1	c^2	...	c^q
x_1	x_1^1	x_1^2	...	x_1^q

La calidad de representación de x_1 por la primera componente es

$$\frac{(x_1^1)^2}{(x_1^1)^2 + (x_1^2)^2 + ... + (x_1^q)^2}$$

Aplicación: Calidad de representación

Individuos: representación por componente

	$\cos^2\alpha$ en c^1	$\cos^2\alpha$ en c^2	$\cos^2\alpha$ en c^3	$\cos^2\alpha$ en c^4
1998	0.739795468	0.160650082	0.072737622	0.006306546
1999	0.697366936	0.047192411	0.142913134	0.05990404
2000	0.000308324	0.6949716	0.000221212	0.006386404
2001	0.01138702	0.395575602	0.022057402	0.519127866
2002	0.001450505	0.106791784	0.472873001	0.000738347
2003	0.04183394	0.079940101	0.122312569	0.662507281
2004	$1.11E^{-5}$	0.043047105	0.63593675	0.144349392
2005	0.06457478	0.429737739	0.426616828	0.034105911
2006	0.179194445	0.436948236	0.148198914	0.121831124
2007	0.575957926	0.288498623	0.110138907	$4.82E^{-5}$
2008	0.887371937	0.011553575	0.065941299	0.013088676

Aplicación: Calidad de representación

Individuos en los 3 planos principales:

	plano12	plano13	plano14
1998	90.04	81.25	74.61
1999	74.46	84.03	75.73
2000	69.53	0.05	0.67
2001	40.70	3.34	53.05
2002	10.82	47.43	0.22
2003	12.18	16.41	70.43
2004	4.31	63.59	14.44
2005	49.43	49.12	9.87
2006	61.61	32.74	30.10
2007	86.45	68.61	57.60
2008	89.89	95.33	90.05

Aplicación: Calidad de representación

Variables: Comunalidades

Entre variables el valor del coseno es igual a una correlación, a menor valor del ángulo, mayor será la calidad de representación de las variables por medio de las componentes principales.

Aplicación: Calidad de representación

Dada una variable x^1 que se relaciona con las componentes principales c^1, c^2, ..., c^q de la forma:

	c^1	c^2	...	c^q
x^1	$r(\mathrm{x}^1, c^1)$	$r(\mathrm{x}^1, c^2)$	...	$r(\mathrm{x}^1, c^q)$

La calidad de representación de x^1 por la primera componente es $r^2(\mathrm{x}^1, c^1)$

La suma de correlaciones al cuadrado se conoce con el nombre de comunalidad.

Aplicación: Calidad de representación

Correlaciones: Variables-Componentes

	c^1	c^2	c^3	c^4
A12	0.5179579	-0.7024007	0.1233709	-0.1593143
A3EDI	0.0545999	-0.8854688	0.2445041	-0.2755068
R1c	-0.7055195	-0.0105493	0.6845327	0.1168024
R2c	-0.7553119	0.1551036	0.5920181	0.1715318
R3c	0.3566694	0.8074964	0.2774089	-0.1915019
RED	0.5315232	0.7198952	-0.2117275	-0.278432
AMG	0.0371369	-0.2373239	0.0689556	-0.9189193

Aplicación: Calidad de representación

Correlaciones: Variables-Componentes

	c^1	c^2	c^3	c^4
DUCR	-0.9249302	-0.0155355	0.2160513	0.056282
DUNA	-0.7592531	-0.1687109	-0.6142163	0.0565419
DITCR	0.7944841	0.4585799	0.3167719	-0.146604
DUNED	-0.8604747	0.3313066	-0.1139269	-0.2883899
DUP	0.9391197	-0.1492145	0.237374	0.051884
DEEd	-0.9302149	0.1421253	-0.1924371	-0.2196666
DPEd	0.8674827	-0.1029957	-0.1778164	0.3114708

Aplicación: Calidad de representación

Variables en los 3 planos (círculos) principales:

	circ.12	circ.13	circ.14
A12	0.76164713	0.283500765	0.293661432
A3EDI	0.787036145	0.062763404	0.078885146
R1c	0.497869053	0.966342782	0.511400566
R2c	0.594553193	0.920981497	0.599919225
R3c	0.779263497	0.204168759	0.163886039
RED	0.800766011	0.327345446	0.360041291
AMG	0.057701783	0.006134024	0.845791829

Aplicación: Calidad de representación

Variables en los 3 círculos de correlaciones:

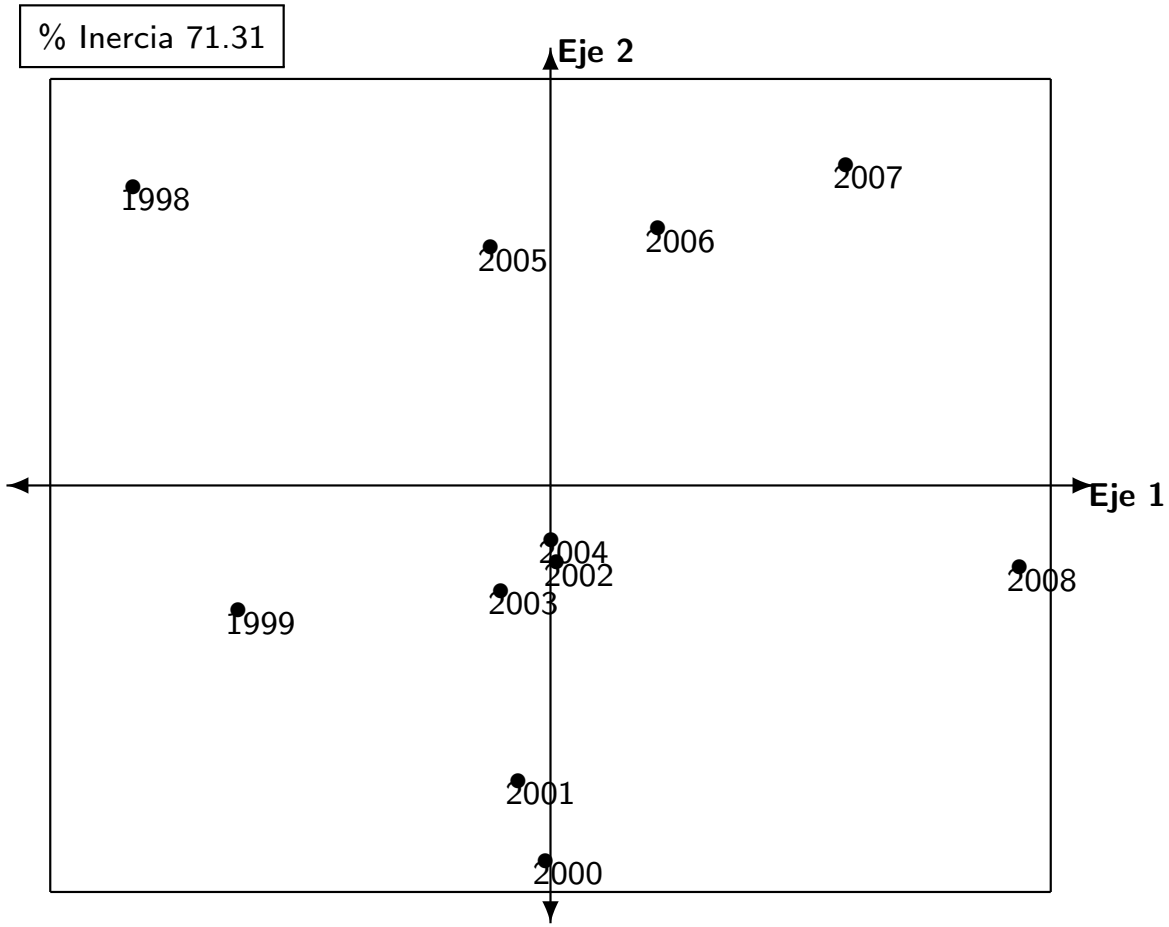
	circ.12	circ.13	circ.14
DUCR	0.855737227	0.902174039	0.858663538
DUNA	0.604928638	0.953726933	0.579662256
DITCR	0.84150051	0.731549422	0.652697718
DUNED	0.850180773	0.753396048	0.823585444
DUP	0.904210778	0.938292227	0.88463776
DEEd	0.885499361	0.902331798	0.913553175
DPEd	0.763134349	0.784144907	0.849540294

Aplicación: Calidad de representación

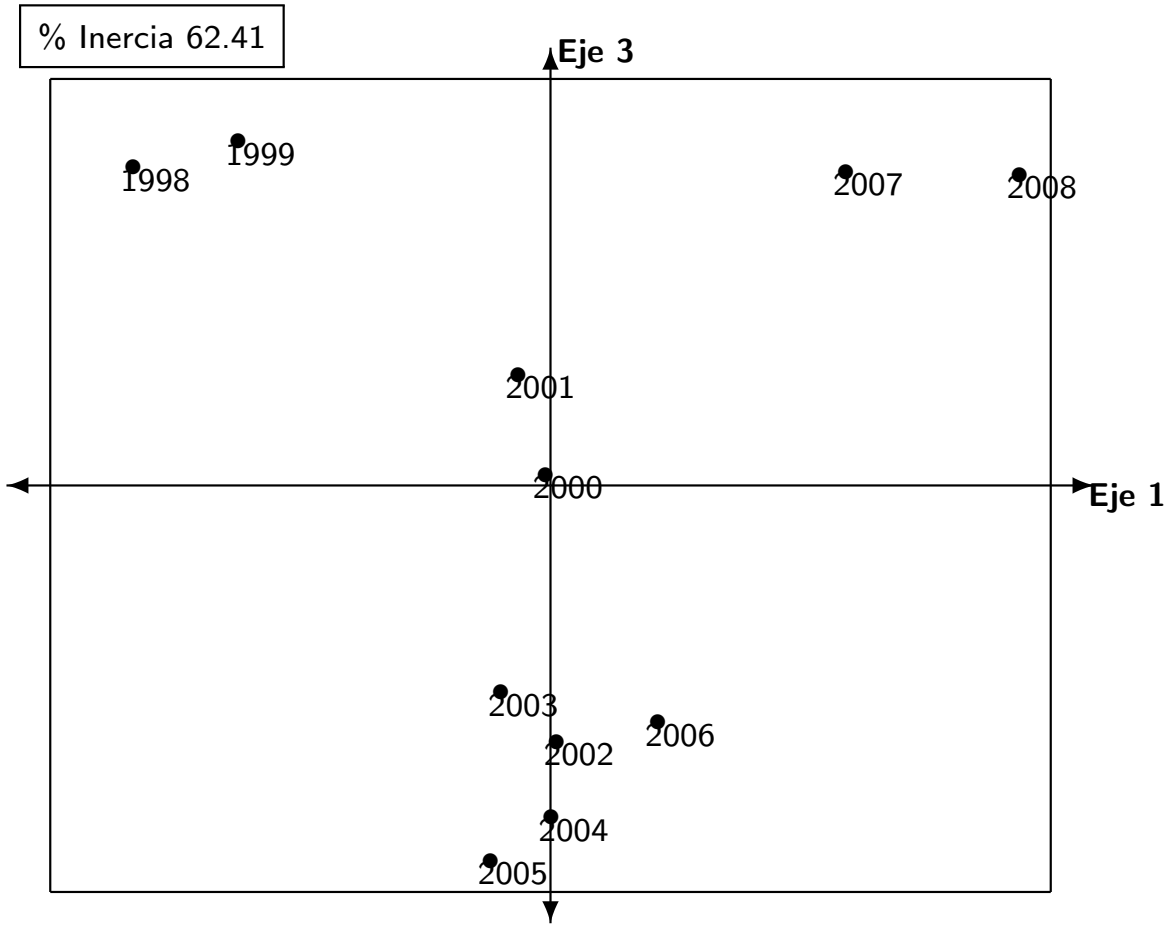
Individuos: relación ejes-años / componente-variable

Eje / comp.	Años	Variables
1	1998, 1999, 2007, 2008	$R1c^-$, $R2c^-$, $DUCR^-$, DNA^-, $DUNED^-$, $DITCR^+$, $DEEd^-$, DUP^+, $DPED^+$
2	2000, 2005	$A12^-$, $A3ED^-$, $R3c^+$, RED^+
3	2002, 2004	$R1c^+$*, $DUNA^-$*
4	2001, 2003	AMG^-

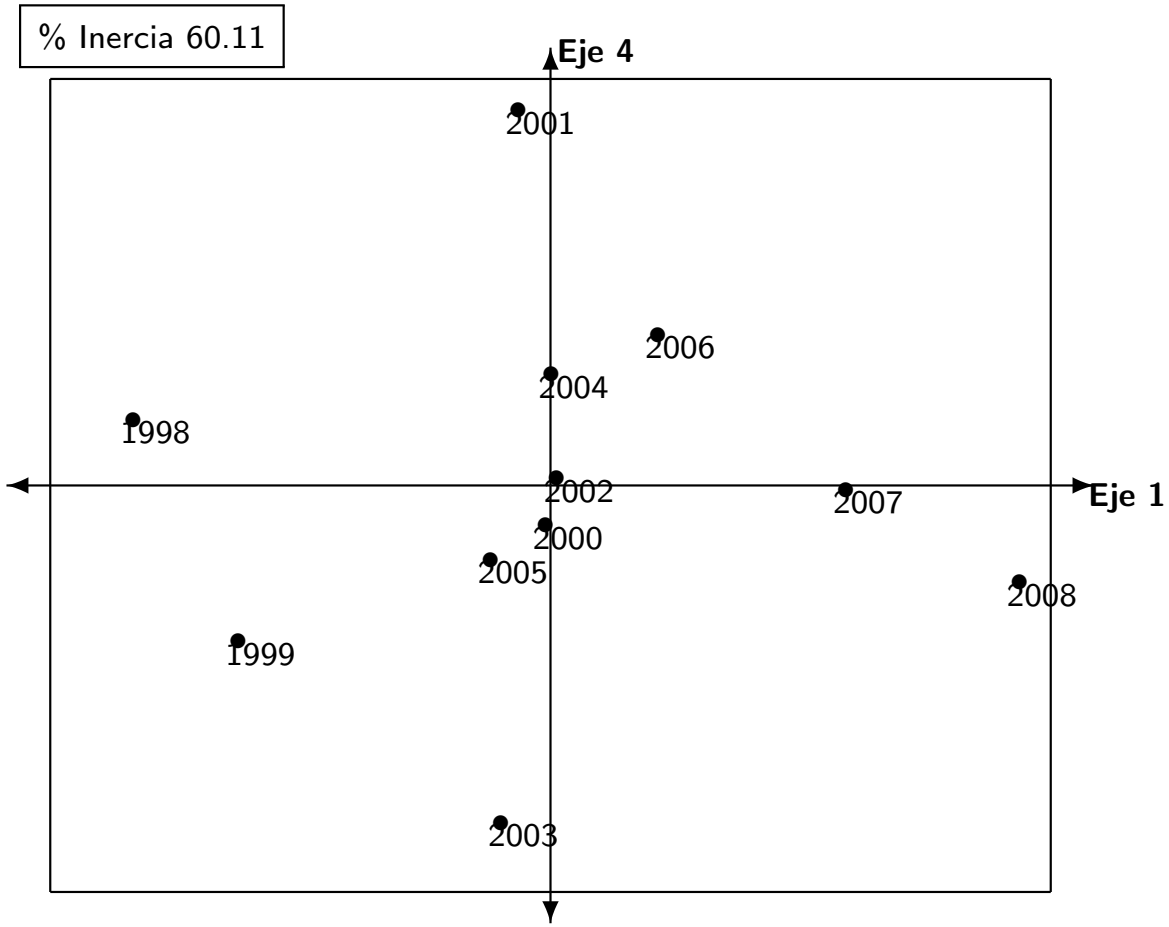
Aplicación: Representación gráfica, plano12



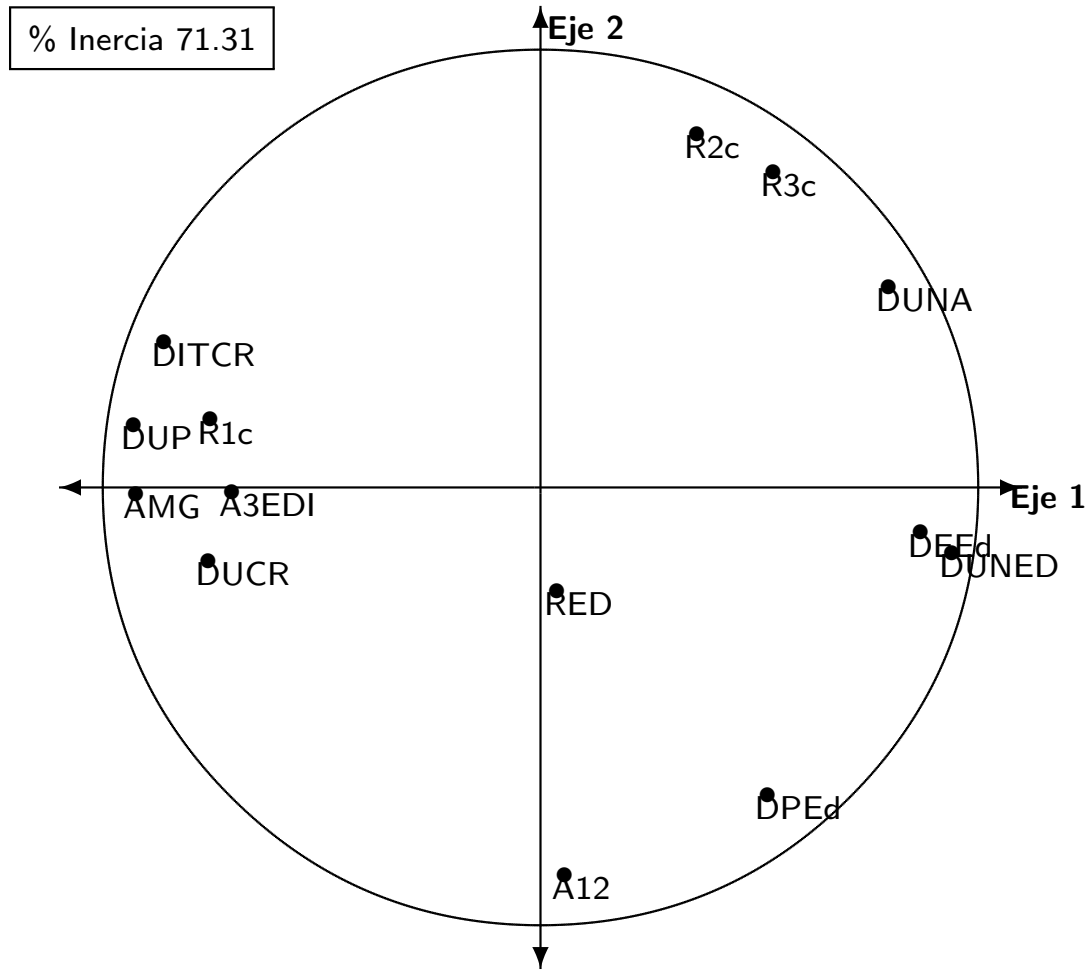
Aplicación: Representación gráfica, plano13



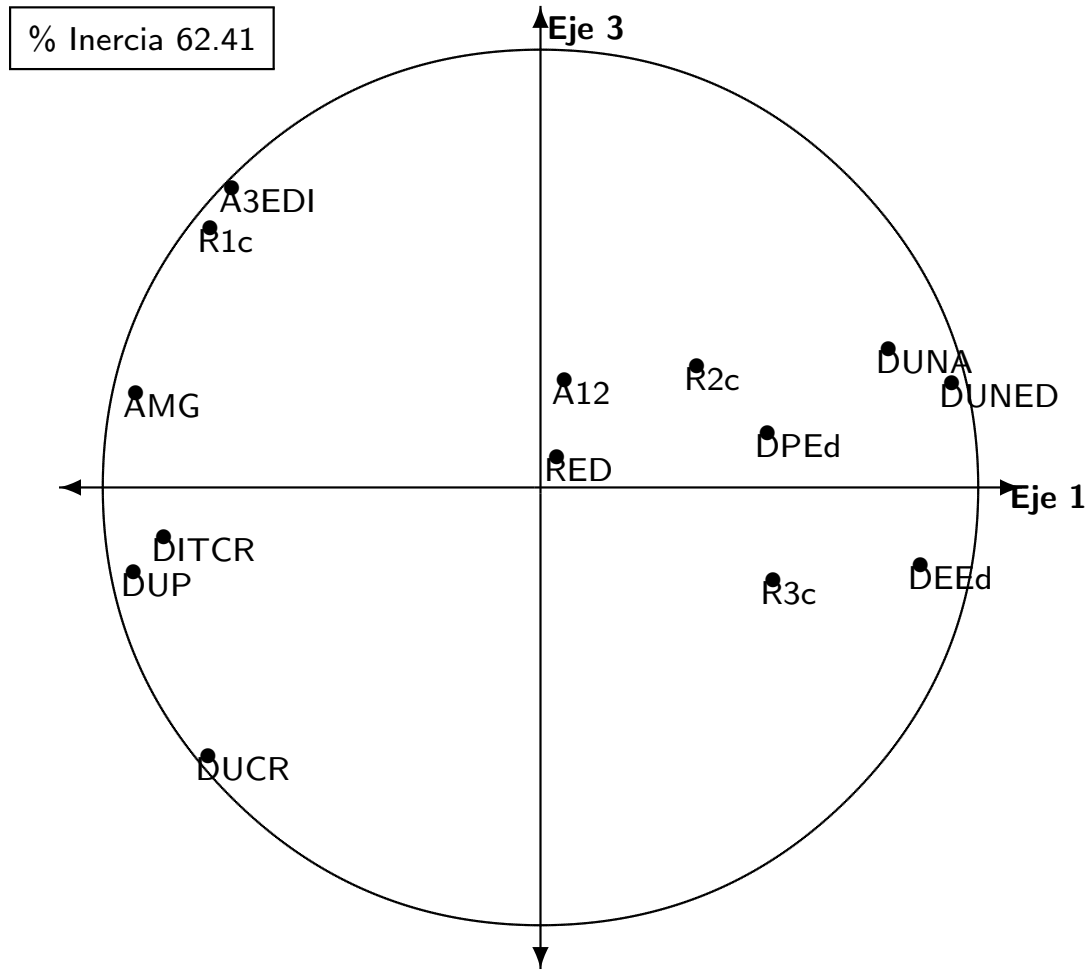
Aplicación: Representación gráfica, plano14



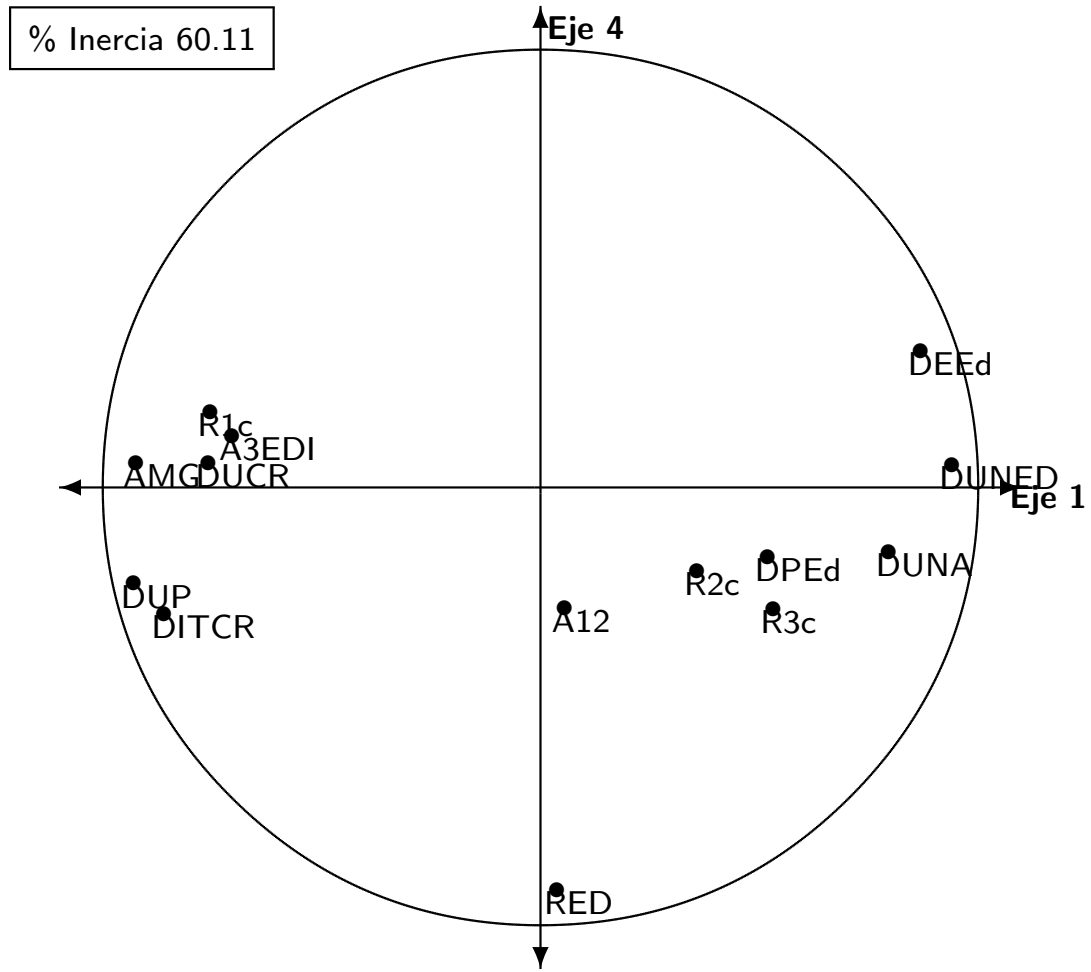
Aplicación: Representación gráfica, círculo 12



Aplicación: Representación gráfica, círculo 13



Aplicación: Representación gráfica, círculo 14



Aplicación: Interpretación de resultados

Caracterización de los ejes

- Eje 1: Diplomas otorgados por universidades privadas.
- Eje 2: Reprobación en nivel de primaria y secundaria.
- Eje 3: Diplomas otorgados en el área de educación.
- Eje 4: Reprobación en curso universitario.

Aplicación: Interpretación de resultados

Caracterización de los ejes

- Eje 1: Diplomas otorgados por universidades privadas.
- Eje 2: Reprobación en nivel de primaria y secundaria.
- Eje 3: Diplomas otorgados en el área de educación.
- Eje 4: Reprobación en curso universitario.

Aplicación: Interpretación de resultados

Caracterización de los ejes

- Eje 1: Diplomas otorgados por universidades privadas.
- Eje 2: Reprobación en nivel de primaria y secundaria.
- Eje 3: Diplomas otorgados en el área de educación.
- Eje 4: Reprobación en curso universitario.

Aplicación: Interpretación de resultados

Caracterización de los ejes

- Eje 1: Diplomas otorgados por universidades privadas.
- Eje 2: Reprobación en nivel de primaria y secundaria.
- Eje 3: Diplomas otorgados en el área de educación.
- Eje 4: Reprobación en curso universitario.

Aplicación: Interpretación de resultados

Observaciones

- Plano 12: Se observa cierta disminución en la repitencia en primaria hasta el año 2004, relacionada al aumento en la cantidad de diplomas otorgados por las universidades privadas, esta reprobación aumenta del 2005 al 2007. Se observa una mejoría en la aprobación.

Aplicación: Interpretación de resultados

Observaciones

- Plano 13: Se observa un aumento en la cantidad de diplomas obtenidos en educación mediante universidades privadas para los años 2007 y 2008. Durante el período 2002-2006 la cantidad de diplomas otorgados en educación disminuye considerablemente y su aumento esta relacionado al aumento de diplomas otorgados en universidades privadas.

Aplicación: Interpretación de resultados

Observaciones

- Plano 14: No se observa mejoría en la aprobación del curso de nivel universitario con respecto al aumento en la cantidad de diplomas otorgados por universidades privadas. El porcentaje de aprobación no supera el 58 %.

Aplicación: Interpretación de resultados

Observaciones

- Círculo 12: Existe gran relación entre la cantidad de diplomas otorgados por la Universidad Estatal a Distancia y la cantidad de diplomas otorgados en al área de educación.
  Se observa relación entre la cantidad de diplomas otorgados por universidades privadas y por el Instituto Tecnológico de Costa Rica, esto debido quizá a su constante aumento.
  Se percibe una importante relación entre el porcentaje de aprobación en el curso universitario Matemática General con la cantidad de diplomas otorgados por universidades privadas.

Aplicación: Interpretación de resultados

Observaciones

- Círculo 13: Se observa una relación inversa entre el porcentaje de diplomas otorgados por universidades estatales en el área de educación y el porcentaje de aprobación ene l curso universitario Matemática General.

Aplicación: Interpretación de resultados

Observaciones

- Círculo 14: Es mínima la relación que existe entre el porcentaje de repitencia del nivel Educación diversificada con el porcentaje de aprobación del curso universitario Matemática General.
  Existe una importante relación entre el porcentaje de aprobación del III nivel y educación diversificada con el porcentaje de aprobación del curso universitario Matemática General.

Referencias

- Abdi, Hervé. Multivariate Analysis. The University of Texas at Dallas.
- Abdi, Hervé. Williams, Lynne J. Principal Component Analysis. The University of Texas at Dallas.
- Castillo E., William. González V., Jorge. Trejos Z., Javier. Anális Multivariado de Datos: Métodos y Aplicaciones. Escuela de Matemática, Universidad de Costa Rica.