UNIVERSIDAD DE COSTA RICA

REVISTA
Actualidades Investigativas en Educación

http://revista.inie.ucr.ac.cr/
ISSN 1409-4703

INIE
Instituto de Investigación en Educación

R PROJECT: SU APLICACIÓN COMO *SOFTWARE* LIBRE PARA ANÁLISIS EN COMPONENTES PRINCIPALES

R PROJECT: ITS USE AS OPEN SOURCE FOR PRINCIPAL COMPONENTS ANALYSIS

Volumen 11, Número Especial
pp. 1-18

Este número se publicó el 30 de junio de 2011

Fabricio Bolaños Guerrero

La revista está indexada en los directorios:

LATINDEX, REDALYC, IRESIE, CLASE, DIALNET, DOAJ, E-REVIST@S,

La revista está incluida en los sitios:

REDIE, RINACE, OEI, MAESTROTECA, PREAL, HUASCARAN, CLASCO

Los contenidos de este artículo están bajo una licencia Creative Commons

R PROJECT: SU APLICACIÓN COMO *SOFTWARE* LIBRE PARA ANÁLISIS EN COMPONENTES PRINCIPALES

R PROJECT: ITS USE AS OPEN SOURCE FOR PRINCIPAL COMPONENTS ANALYSIS

Fabricio Bolaños Guerrero[1]

Resumen: *Este artículo es producto de un proyecto de investigación realizado en colaboración con profesores de la Escuela de Matemáticas de la Universidad de Costa Rica (UCR), para dar a conocer una opción de software estadístico llamado R Project. Con este paquete es posible hacer Análisis en Componentes Principales (ACP) y representar los resultados usando el Plano Principal y el Círculo de Correlaciones, como herramientas para poder realizar una mejor interpretación de los datos de la tabla (individuos y variables). El software R es de distribución libre, su implementación es sencilla y no requiere de mayores recursos informáticos. Dentro de sus diversas aplicaciones está el ACP, que es una herramienta que se utiliza para la interpretación de la información presentada en una tabla de datos cuantitativos; por lo tanto, las personas investigadoras de diferentes áreas tienen una opción económica y sencilla para realizar Análisis de Datos. Se llevan a cabo dos ejemplos de ACP, donde se muestra un posible uso de la herramienta y se dan las instrucciones sobre cómo realizarlo paso a paso.*

Palabras clave*: ANÁLISIS DE DATOS ESTADÍSTICOS, ANÁLISIS DE COMPONENTES PRINCIPALES, R PROJECT, SOFTWARE LIBRE.*

Abstract: *This article is a product of a research project made in collaboration with teachers of the Mathematics School of the University of Costa Rica, in order to show an option of statistical software called "R Project". With this software, it is possible to do an Analysis in the Principal Components (PCA) and to represent the results using the "principal plane" and the "circle of correlations", as tools to have a better interpretation of the data in the chart (individuals and variables). The software R is for free distribution, its implementation is simple and it does not require great computer resources. Among its diverse applications there is the PCA, which is a tool used to interpret the information showed in a chart with quantitative data; therefore, the researchers of different areas have a cheap and simple option to do a Data Analysis. There are two examples of PCA carried out, where it is showed a possible use of the tool and there are given the instructions of how to do it step by step.*

Key words*: STATISTICAL ANALYSIS, PRINCIPAL COMPONENTS ANALYSIS, R PROJECT, OPEN SOURCE.*

[1] *Licenciado en Matemática, Master en Administración y Dirección de Empresas ambos títulos de la Universidad de Costa Rica. Profesor Sedes del Pacífico y de Guanacaste de la Universidad de Costa Rica.*

Direcciones electrónicas: Fabriciobolanos@gmail.com y fabricio.bolanos@ucr.ac.cr

Artículo recibido*: 17 de febrero, 2011*
Aprobado*: 16 de junio, 2011*

1. Introducción

El *software* libre viene mostrando un aumento en su participación de mercado a nivel nacional e internacional y, en este sentido, las autoridades de la Universidad de Costa Rica (UCR) no se han quedado atrás, tal como lo muestra la política institucional emitida por el Consejo Universitario en La Gaceta Universitaria #43-2008, año XXXII del 5 de Diciembre de 2008, que hace referencia al software libre en la Universidad de Costa Rica; dicho documento expresa:

"El Consejo Universitario, **CONSIDERANDO QUE:**

1. El *software* libre es un conjunto de herramientas que en la actualidad goza de un uso muy amplio, debido a la expansión de Internet y a la necesidad de contar con una solución que no esté regida por intereses comerciales. Este tipo de *software* posee mucho potencial, porque es respaldado por una comunidad desarrolladora y por usuarios que robustecen esas herramientas.

2. Es importante fortalecer el aprovechamiento racional y el manejo eficiente del recurso tecnológico en la Universidad.

ACUERDA:

Solicitar a la Rectoría integrar esfuerzos para presentar al Consejo Universitario, en un plazo de un año, una estrategia de trabajo integral sobre la posibilidad de iniciar un proceso de migración de las personas usuarias de ofimática de la Institución, hacia el uso de *software* libre y *software* gratuito. Dicha estrategia debe contener, entre otros aspectos:

- Consideraciones sobre las aplicaciones informáticas de uso institucional que no pueden utilizar software libre.
- Valoración de la infraestructura de servidores y su posibilidad de uso de software libre, como parte de las etapas por desarrollar.
- Posibilidad de migrar de manera paulatina las oficinas que puedan incorporarse como plan piloto.
- Definición de etapas de aplicación de la estrategia, así como el tiempo que tardará cada una de ellas; debe incluir una estimación del tiempo de capacitación del personal. Fortalecer el presupuesto de capacitación para las personas que tendrán a cargo la migración de este primer grupo de usuarios.
- Promover la utilización de software gratuito.

ACUERDO FIRME."

Además, existe un proyecto de ley denominado "UTILIZACION DE SOFTWARE LIBRE EN LAS INSTITUCIONES DEL ESTADO, Expediente No 16.912", actualmente en trámite en la Asamblea Legislativa, del cual se tomó el siguiente cuadro que representa los gastos de algunas instituciones públicas en el año 2007:

Cuadro N° 1
Gasto en software y licencias Instituciones Públicas – 2007

Institución	**Presupuesto (miles de colones)**	**Presupuesto (dólares)****
Instituto Costarricense de Electricidad	3.747.861	7.071.436
Instituto Mixto de Ayuda Social	6.613	12.478
Instituto Nacional de Seguros	651.502	1.229.248
Acueductos y Alcantarillados	100.000	188.679
Instituto Costarricense de Turismo	10.362	19.551
Instituto Nacional de Aprendizaje	142.364	268.612

Elaboración propia, a partir de la información obtenida en http://www.conare.ac.cr/proyectos/16912.htm el 11 de marzo de 2011

** Proyección del dólar para el año 2007 utilizada por la Oficina de Administración Financiera de la U.C.R

Por lo anterior, en el presente trabajo se desarrolla un algoritmo clásico de Análisis de Datos utilizando una herramienta de software libre llamada *R Project* o simplemente R; esto supone la utilización de *software* libre no sólo en ofimática (*software* de oficina como los procesadores de textos, hojas de cálculo, presentaciones), sino en actividades especializadas de investigación, para lo cual la Institución o el interesado sólo deben preocuparse por la capacitación en el uso de la herramienta. Cabe considerar que los costos que deben asumir las universidades por la utilización de programas que cumplen funciones similares a *R Project*, podrían ser evitados si se utiliza *software* libre. En caso de que dichos costos se reduzcan o eliminen del todo, será posible destinar más recursos institucionales a otras actividades que las autoridades estimen más productivas o estratégicas.

Ahora bien, la participación de varios profesores de la Escuela de Matemática de la Universidad de Costa Rica en eventos internacionales (algunos realizados en Europa, por ejemplo), los elevados costos de adquisición del *software* especializado -en este caso para Análisis de Datos- y la tendencia mundial de hacer migraciones a *software* libre, motivaron la necesidad de introducir herramientas de código abierto para realizar una de las técnicas básicas de Análisis de Datos, el *Análisis en Componentes Principales* (en adelante ACP) utilizando para esto el *Programa R* (Project R Development Core Team, 2009)[2].

Es importante mencionar que también existe toda una plataforma de Análisis de Datos que se llama *Facto mine* que trabaja sobre R y que permite resolver una gran cantidad de problemas de Análisis de Datos, en particular ACP de una forma un poco más amigable, es decir, con una mejor interfaz gráfica, sin necesidad de dar instrucciones paso a paso para llevar a cabo el ACP; el mismo programa R tiene incorporada una función que realiza el ACP (*princomp*) sin pasar por toda la descripción del método (operaciones matriciales, diagonalización de matrices y proyección de los individuos sobre un subespacio de menor dimensión) que se realiza en el presente trabajo.

2. Referente teórico

La descripción del método es primordial, pues es precisamente lo que genera conocimiento, ya que de lo contrario, nos convertiríamos en simples usuarios finales de herramientas tecnológicas; esto limitaría las posibilidades del investigador de generar nuevo conocimiento al ignorar la justificación del cómo y por qué funciona el método. La justificación teórica matemática no es el objetivo del presente trabajo, sino más bien dar a conocer la herramienta de uso libre R mediante la implementación de un método clásico de Análisis Multivariado de Datos que se utiliza con variables de tipo cuantitativo. Dicho método es el ya mencionado ACP[3], el cual se irá desarrollando paso a paso a lo largo del trabajo, con la

[2] Por ser libre, el *software* funciona en ambiente Linux, MAC y, por supuesto, Windows que fue donde se utilizó en el presente trabajo. El software se puede conseguir en la siguiente dirección electrónica http://www.r-project.org/ donde se pueden encontrar gran variedad de documentación como referencias a libros, los manuales (en inglés) y publicaciones relativas a R. En español se recomienda **R para principiantes** de Emmanuel Paradis de la Universidad de Montpellier II traducido por Jorge Ahumada citado en las Referencias.

[3] Si el lector lo desea puede consultar el desarrollo de este método en particular, detalles específicos y limitaciones en Castillo, González y Trejos (2007), Bolaños (1996) y Guevarra (1980); además, puede consultar la página del Profesor Oldemar Rodríguez Rojas, Ph.D. en http://www.oldemarrodriguez.com, sección de cursos.

intención de ilustrar la forma en la que puede ser utilizado por investigadores de diferentes áreas.[4]

Si consideramos los datos de un individuo como un punto en un espacio de dimensión "p", el cual no es fácil de representar gráficamente, la tabla de datos sería una colección de "n" puntos; el objetivo del método es encontrar el subespacio de menor dimensión (generalmente dos) sobre el cual se pueden proyectar los datos, de manera que resulte más fácil hacer su interpretación, para que el investigador especializado pueda hacer mejores conclusiones en cuanto a los individuos o las variables.

En este sentido, se definen el *Plano Principal* como la representación gráfica de los individuos del análisis en un plano (que es la proyección sobre las componentes principales) y el *Círculo de Correlaciones*, como la representación gráfica de las variables sobre un círculo que es indispensable en un análisis multivariado de datos.

Un detalle importante en la implementación del ACP es que la solución del problema no es única, dado que los vectores propios que diagonalizan la matriz pueden quedar multiplicados por -1, pero son siempre ortogonales; es decir, la representación gráfica puede variar según se hayan tomado los vectores propios, lo que, por supuesto, no representa ningún problema, porque sería como "ver" los resultados reflejados en un espejo[5].

3. Metodología

Se presentarán dos tablas de datos a las cuales se les realizarán los respectivos Análisis en Componentes Principales. En una, se presentan los *Gastos del Estado de Francia* y, en la otra, las *Distancias de algunas ciudades de Costa Rica*. Se detallará cómo se lleva a cabo lo anterior con el Lenguaje R, ya que las instrucciones son básicamente operaciones matriciales y la respectiva representación gráfica de este análisis en el *Plano Principal* y el *Círculo de Correlaciones*.

[4] Hay que recordar que esta aplicación (A.C.P.) de R es un caso particular muy sencillo de esta poderosa herramienta y que los campos de aplicación son muy variados por ejemplo biotecnología, demografía, historia, ciencias de la salud entre otras disciplinas que requieran estudiar tablas de datos cuantitativos y necesiten realizar una análisis multivariado de datos.

[5] La comparación de los datos realizados por Bouroche y Saporta (1989) y los obtenidos por el autor mediante el programa se presentan en el Anexo No. 1, para el caso del plano principal.

3.1 La tabla de Datos de Gastos del Estado de Francia

Como primer ejemplo, se presenta el cuadro de *Gastos del Estado de Francia*, donde se presentan, en porcentaje, los gastos que se hicieron en ese país durante algunos años (individuos) y en varios sectores económicos (variables), almacenados en columnas; los datos fueron obtenidos de Bouroche y Saporta (1989, p. 18).

Cuadro No. 2

La estructura funcional de gastos del Estado de Francia

1872-1971 (en porcentajes)

AÑO	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ACO	DEF	DET	DIV
1872	18	0,5	0,1	6,7	0,5	2,1	2	0	26,4	41,5	2,1
1880	14,1	0,8	0,1	15,3	1,9	3,7	0,5	0	29,8	31,3	2,5
1890	13,6	0,7	0,7	6,8	0,6	7,1	0,7	0	33,8	34,4	1,7
1900	14,3	1,7	1,7	6,9	1,2	7,4	0,8	0	37,7	26,2	2,2
1903	10,3	1,5	0,4	9,3	0,6	8,5	0,9	0	38,4	27,2	3
1906	13,4	1,4	0,5	8,1	0,7	8,6	1,8	0	38,5	25,3	1,9
1909	13,5	1,1	0,5	9	0,6	9	3,4	0	36,8	23,5	2,6
1912	12,9	1,4	0,3	9,4	0,6	9,3	4,3	0	41,1	19,4	1,3
1920	12,3	0,3	0,1	11,9	2,4	3,7	1,7	1,9	42,4	23,1	0,2
1923	7,6	1,2	3,2	5,1	0,6	5,6	1,8	10	29	35	0,9
1926	10,5	0,3	0,4	4,5	1,8	6,6	2,1	10,1	19,9	41,6	2,3
1929	10	0,6	0,6	9	1	8,1	3,2	11,8	28	25,8	2
1932	10,6	0,8	0,3	8,9	3	10	6,4	13,4	27,4	19,2	0
1935	8,8	2,6	1,4	7,8	1,4	12,4	6,2	11,3	29,3	18,5	0,4
1938	10,1	1,1	1,2	5,9	1,4	9,5	6	5,9	40,7	18,2	0
1947	15,6	1,6	10	11,4	7,6	8,8	4,8	3,4	32,2	4,6	0
1950	11,2	1,3	16,5	12,4	15,8	8,1	4,9	3,4	20,7	4,2	1,5
1953	12,9	1,5	7	7,9	12,1	8,1	5,3	3,9	36,1	5,2	0
1956	10,9	5,3	9,7	7,6	9,6	9,4	8,5	4,6	28,2	6,2	0
1959	13,1	4,4	7,3	5,7	9,8	12,5	8	5	26,7	7,5	0
1962	12,8	4,7	7,5	6,6	6,8	15,7	9,7	5,3	24,5	6,4	0,1
1965	12,4	4,3	8,4	9,1	6	19,5	10,6	4,7	19,8	3,5	1,8
1968	11,4	6	9,5	5,9	5	21,1	10,7	4,2	20	4,4	1,9
1971	12,8	2,8	7,1	8,5	4	23,8	11,3	3,7	18,8	7,2	0

Fuente: Bouroche y Saporta (1989, p. 18).

Los gastos de cada año, de este periodo en estudio, se realizaron (descripción de las variables), según los siguientes sectores económicos:

PVP: Pouvoirs publics (Poderes Públicos).

AGR: Agriculture (Agricultura).

CMI: Commerce et industrie (Comercio e Industria).

TRA: Transports. (Transportes).

LOG: Logement et aménagement du territoire (Vivienda y ordenación del terrritorio).

EDU: Education et culture. (Educación y cultura).

ACS: Action sociale. (Acción social).

ACO: Anciens combattants (Veteranos combatientes).

DEF: Defense (Defensa).

DET: Dette (Deuda).

DIV: Divers (Varios).

Este ACP clásico aparece en Bouroche y Saporta (1989, pp. 40-45), donde puede consultarse la proyección de los individuos en el plano principal, el círculo de correlaciones y las interpretaciones de los resultados.

3.2 La tabla de Distancias de algunas ciudades de Costa Rica

El segundo ejemplo que se presenta es sobre un cuadro de *Distancias entre algunas ciudades de Costa Rica* tomadas de Instituto Costarricense de Turismo (ICT) (2005). Dicho cuadro se encuentra en el Anexo N° 2.[6]

3.3. Las instrucciones que se deben digitar en R

Los comandos que se deben digitar en R son de dos tipos:

- Primero: operaciones de matrices para realizar el ACP de la tabla de datos deseada.
- Segundo: para realizar los gráficos el Plano Principal y el Círculo de Correlaciones, estos son simplemente instrucciones para hacer gráficos en dos dimensiones:

[6] La limitación que se presenta en este caso es que la distancia entre ciudades se obtiene por medio de la medición por carretera y no en línea recta, por lo que las mismas dependen de la apertura de nuevas rutas, como ocurrió recientemente con la nueva -y polémica- carretera a Caldera.

3.3.1 Instrucciones para realizar el ACP en R[7].

1. Introducir los datos como un documento de texto y guardarlo como extensión txt. Donde R pueda leerlos con el comando definido en el paso siguiente; en este caso, los datos están almacenados en un archivo que se llama **Gastos.txt** y en el mismo directorio donde se encuentra R, si no es, así la siguiente instrucción no se podría ejecutar, ahora, también se puede cambiar de directorio con una pestaña del programa hacia el directorio donde se encuentran los datos; los nombres de las variables así como los individuos se leerán después, de la misma forma, ya que estos se usaran únicamente en la representación gráfica del análisis.

2. En R digitar:
   - Z<-read.table ("gastos.txt")

3. El promedio de cada variable lo realiza la función interna del programa mean(X), por lo que se puede guardar en un vector de promedios de las variables usando la única instrucción:
   - centro<-mean(Z)

4. Para saber cuántas filas y columnas tiene la matriz, definir la matriz de (n x m) y realizar el proceso de centrar la matriz de datos se pueden ejecutar las siguientes instrucciones:
   - filas<-nrow(Z)
   - columnas<-ncol(Z)
   - X<- matrix(data=0,nr=filas,nc=columnas)
   - for (j in 1:columnas) {for (i in 1:filas) X[i,j]<-Z[i,j]-centro[j] }

   La variable que almacena la matriz con los datos centrados es X.

5. Para definir la matriz de pesos, primero la inicializamos como una matriz de ceros y después le ponemos los pesos en la diagonal que queremos, esto por cuanto la función diag que crea una matriz diagonal no sirve, pues aunque crea la misma matriz diagonal,

[7] Puede consultarse el manual elemental sobre R preparado por Paradis (2002).

esta se guarda como un vector y, por lo tanto, no se pueden realizar las operaciones matriciales necesarias.

- D<- matrix(data=0,nr=filas,nc=filas)
- for (i in 1:filas) {D[i,i]=1/filas}

6. Luego, para calcular la matriz de varianzas, simplemente se realiza la multiplicación de matrices respectiva.
   - V1<- t(X) %*% D
   - V<-V1 %*% X

7. La métrica usual que se utiliza en A.C.P. es la matriz diagonal de las inversas de las varianzas de las variables D_s que se almacena en la variable Ds
   - Ds<- matrix(data=0,nr=columnas,nc=columnas)
   - for (i in 1:columnas){Ds[i,i]=(1/V[i,i])^(1/2)}

8. Luego, calculamos la matriz de correlaciones que está dada por R = D_s V D_s y se guarda en la variable R.
   - R1<- Ds %*% V
   - R<-R1 %*% Ds

9. Para diagonalizar esta matriz R cuenta con una función incorporada que se llama *eigen,* por lo que solamente realizamos las siguientes instrucciones para calcular y almacenar los valores propios y los vectores propios respectivos en las respectivas variables.
   - propios<-eigen(R)
   - valpropios<-propios$values
   - vecpropios<-propios$vectors

10. Las coordenadas de los individuos en el plano principal se obtienen haciendo la multiplicación de las siguientes matrices C <- X %*% D_s %*% vecpropios. El plano principal se obtiene de escoger las dos primeras columnas de C, que se pueden accesar en R de la siguiente forma: C[,1], C[,2].

- C1<-X %*% Ds
- C<-C1 %*% vecpropios

11. Después, las coordenadas de las variables en el círculo de correlaciones se obtienen tomando las dos primeras columnas de la matriz CC calculada de la siguiente forma:
    - Matvalpropios<-matrix(data=0,nr=columnas, nc=columnas)
    - for (i in 1:columnas){Matvalpropios[i,i]=(valpropios[i])^(1/2)}
    - CC<-vecpropios %*% Matvalpropios

12. El vector de inercias se calcula por la instrucción:
    - Inercias<-vector(mode="double", columnas)
    - for (i in 1:columnas){ Inercias [i]=(valpropios[i])/ columnas}

3.3.2 Instrucciones para hacer los gráficos en R

Si el lector desea una forma más depurada de realizar los gráficos puede consultar R para principiantes de Emmanuel Paradis (2002, pp. 35-40); en este caso, se presenta una forma muy simple de realizar los gráficos sin mucho detalle.

- individuos<-read.table("gastos ind.txt")
- plot(C[,1], C[,2],pch=19,cex=0.5,ylim=c(min(C[,2])-0.5,max(C[,2])+0.5),
  xlim=c(min(C[,1])-0.5,max(C[,1])+1),
  xlab="Factor 1: 45.3%",
  ylab="Factor 2: 18.6%",
  main="Plano Principal",cex.main=0.8)
- for (i in 1:filas){text(C[i,1],C[i,2],label=individuos[i,1],adj=c(0,1),cex=0.7)}

13. El caso con las variables es prácticamente análogo con el de los individuos, por lo que los pasos para leerlos y después representarlos en el Círculo de Correlaciones se realizan de la misma forma:
    - variables <-read.table("gastos var.txt")
    - plot(CC[,1],CC[,2],pch=19,cex=0.5,ylim=c(-1,1), xlim=c(-1,1),
    - main="Las correlaciones entre las variables ",cex.main=0.8)

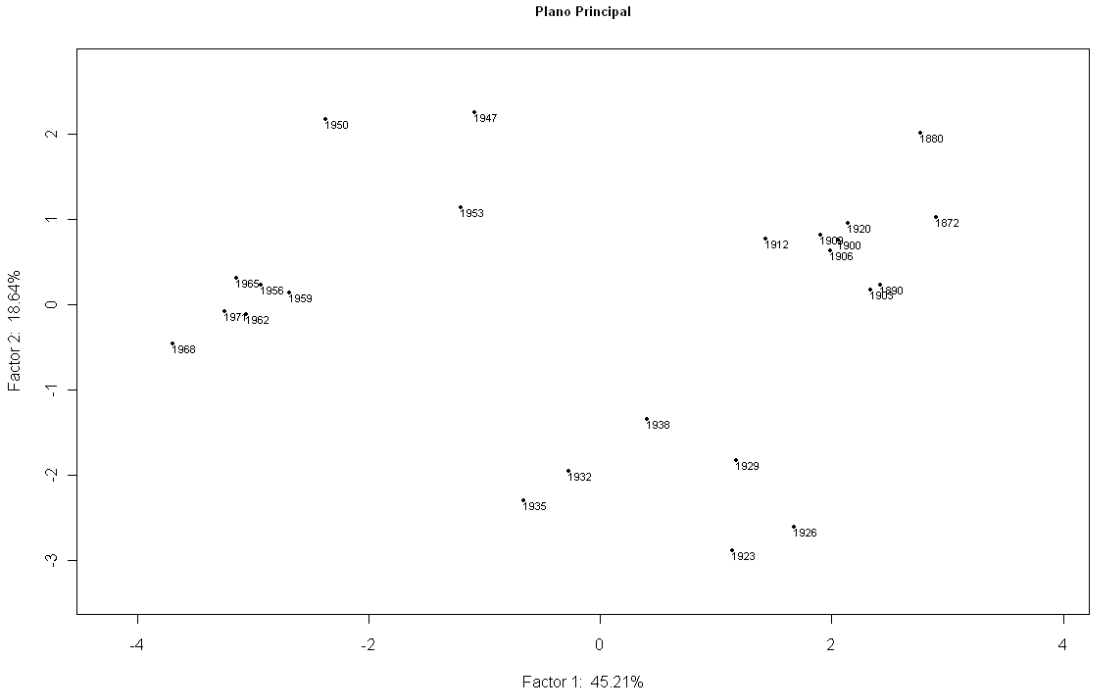
- for (j in 1:columnas){text(CC[j,1],CC[j,2],label=variables[j,1],adj=c(0,1),cex=0.7)}

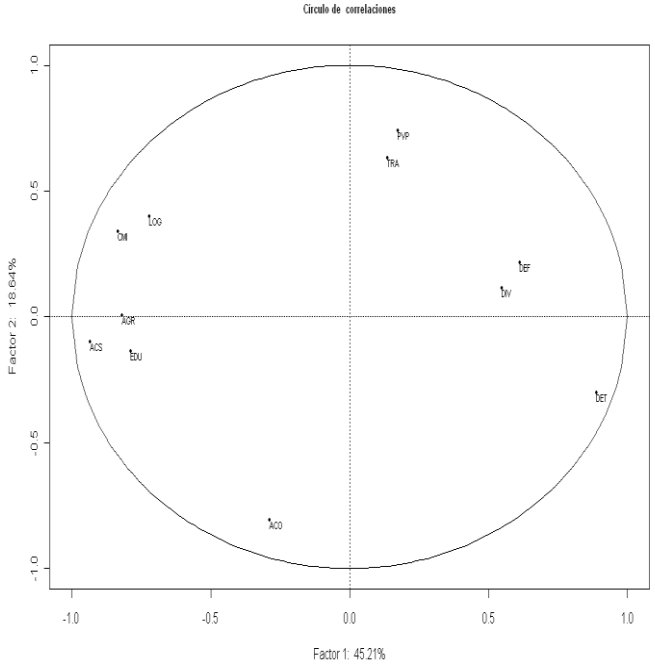
14. Si se desea hacer los círculos de radio 1, se deben digitar las dos primeras instrucciones siguientes y para implementar los ejes de coordenadas se hace con las últimas instrucciones; puede además, si el usuario lo desea, hacer los ejes continuos o punteados
    - curve(sqrt(1-x^2),add = TRUE)
    - curve(-sqrt(1-x^2),add = TRUE)
    - abline(h=0,lty=1)
    - abline(h=0,lty=3)

4. Resultados y discusión

4.1 Caso de la tabla de Gastos de Estado de Francia

Los resultados que se obtuvieron utilizando el software R, con las instrucciones anteriores para realizar el *Análisis en Componentes Principales* en el caso de la tabla de *Gastos del Estado de Francia* se presentan en los siguientes gráficos del *Plano Principal* y el *Círculo de Correlaciones*.


Plano Principal
1950
1947
1880
1953
1872
1920
1912
1909
1900
1906
1965
1956
1959
1890
1905
1971
1962
1968
1938
1929
1932
1935
1926
1923
Factor 2: 18.64%
Factor 1: 45.21%


Circulo de correlaciones
PVP
TRA
COG
DIM
BEF
DIV
AGR
ACS
EDU
BEF
ACO
Factor 2: 18.64%
Factor 1: 45.21%

En el análisis efectuado en Bouroche y Saporta (1989, pp. 40-45), los datos que se obtienen son los mismos, salvo los signos, lo que no afecta la interpretación de los resultados, ya que, por ejemplo, en el caso del cambio de signo en el primer eje equivale a pasar los puntos del lado positivo del eje "x" al lado negativo (de derecha a izquierda) y lo mismo para el signo del segundo eje, esta transformación simplemente trasladaría los puntos que están arriba del eje para abajo y viceversa. Esta comparación se puede observar en el Anexo No. 1.

En cuanto al *Plano Principal* se pueden ver cuatro grupos de individuos claramente separados:

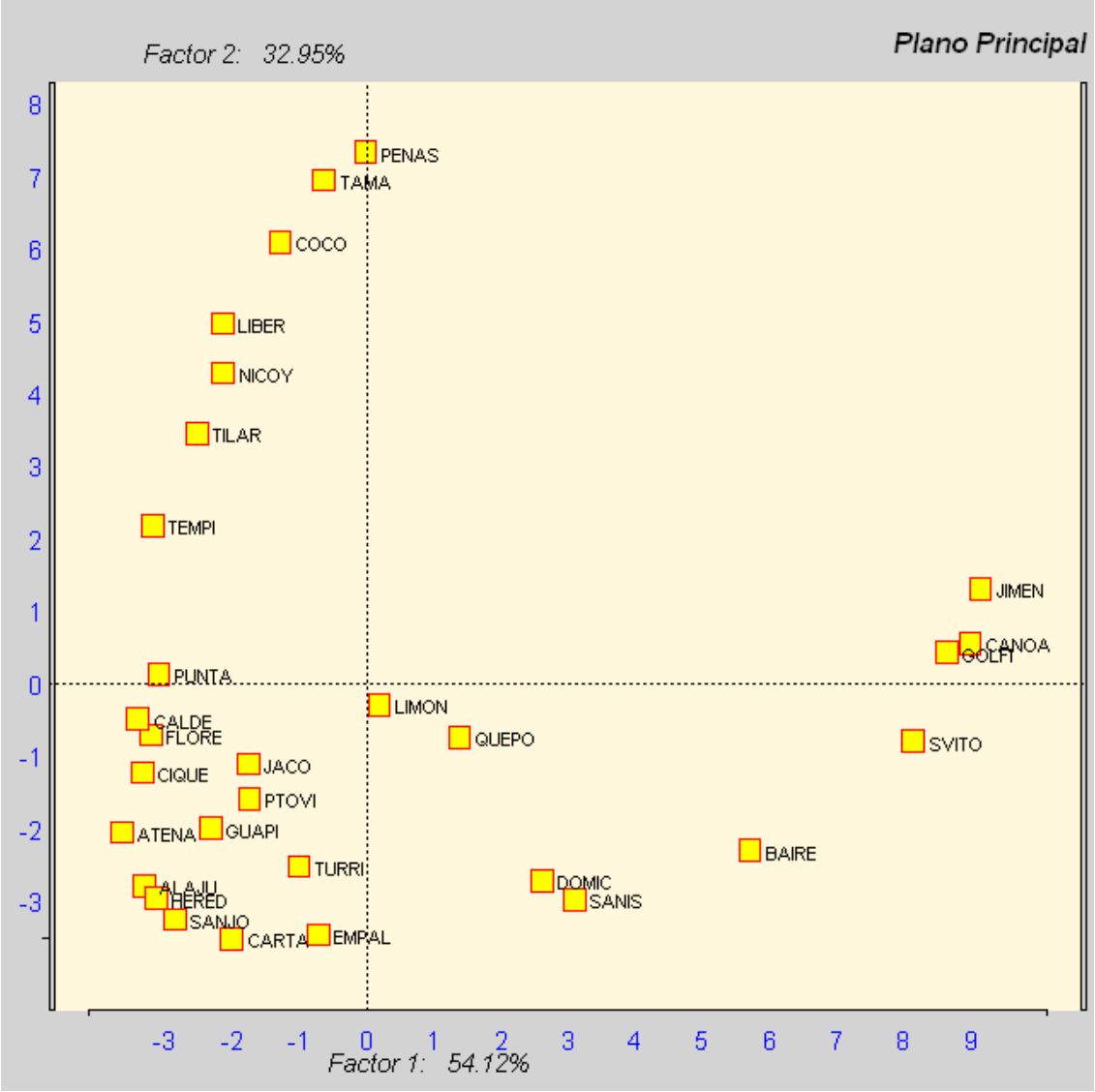
a) El periodo de la posguerra 1947, 1950 y 1953.
b) El periodo de antes de la Primera Guerra Mundial 1872 a 1920.
c) El periodo posterior de la Segunda Guerra Mundial 1956 a 1971.
d) El periodo entre las dos guerras mundiales de 1923 a 1936.

Del *Círculo de Correlaciones* se pueden notar las oposiciones entre algunas variables más correlacionadas ACS, CMI, AGR, EDU contra DET y DEF[8].

4.2 Caso de la tabla de Distancias entre ciudades de Costa Rica

El resultado de este ACP, para el caso del cuadro de distancias de Costa Rica, se presenta a continuación, realizando unos cambios de signos en los vectores propios, para obtener un gráfico similar al de un mapa político de Costa Rica.

[8] Si el lector desea conocer más interpretaciones puede consultar Bouroche y Saporta (1989).



Se pueden observar tres grandes grupos de ciudades y unas cuantas que aparecen solas:

a) Las que se ubican en la zona norte de la carretera interamericana: Peñas Blancas, Tamarindo, Playas del Coco, Liberia y otras.
b) Un sector central que comprende el Gran Área Metropolitana: Alajuela, San José, Heredia, Cartago, Atenas, Turrialba y otras.
c) Las que se ubican al sur del país: Paso Canoas, Golfito, Puerto Jiménez, San Vito y otras.

d) El caso de Limón y Quepos, que se ubican al centro del gráfico y muy cercanas entre sí, se puede explicar atendiendo al hecho de que en la tabla de datos (Anexo N° 2) se observa que la distancia de Quepos a San Isidro de El General es de 326 kilómetros lo cual no es correcto, a menos que se recorra primero de Quepos a San José (192 km) y después de San José a San Isidro de El General (134 km). La justificación de este error, tal vez, sea que el hecho que para algunos turistas que no rentan automóvil la ruta que deben realizar es esta -y no la más corta de Quepos a Dominical y de Dominical a San Isidro de El General-.

De hecho, se debe recordar que estos datos se tomaron de un mapa turístico y, en consecuencia, las distancias están medidas por las carreteras existentes a esa fecha, por lo que no se contempla la nueva ruta a Caldera, inaugurada en 2010. Si se tomaran en cuenta estas distancias la representación gráfica cambiaría

Como se desprende de estos ejemplos, la facilidad con la que se pueden interpretar los datos utilizando el *Plano Principal* y el *Círculo de Correlaciones* versus las tablas de datos iniciales, gracias a las herramientas que tiene incorporadas el *software* R, en este caso de operaciones con matrices y gráficos, es una de las fortalezas del *Análisis en Componentes Principales*, que es una herramienta de Análisis Multivariado de Datos necesaria para ser utilizada por investigadores de diferentes áreas que requieran realizar estos tipos de análisis.

5. Conclusiones

En esta nueva era de migración hacia el *software* libre, el principal aporte de este trabajo es dar a conocer la herramienta de *software* R Project mediante dos tablas de datos a las que se les aplicó el método de Análisis en Componentes Principales. Dado que el *software* R es de distribución libre, su implementación es sencilla y no requiere de mayores recursos informáticos y el hecho que el *Análisis en Componentes Principales* es una herramienta que se utiliza para la interpretación de la información presentada en una tabla de datos, los investigadores de diferentes áreas tienen una opción económica y sencilla para analizar sus datos; luego, las instituciones de educación superior cuentan con una poderosa herramienta para realizar investigaciones en varias áreas del conocimiento con el único gasto de capacitación (que se hace sólo una vez) sin incurrir en los costosos gastos de licencias de *software* (que hay que hacerlas cada año), ya que no debería de darse el hecho de que países del tercer mundo

tengan que depender de la adquisicion de estas herramientas para poder llevara cabo estos tipos de análisis de datos.

Finalmente, una consecuencia adicional que se puede establecer es que el uso de *software* de distribución libre, tales como R, propician una verdadera democratización del conocimiento, pues permiten que los investigadores mismos puedan realizar este tipo de análisis en particular y otros más, sin las limitaciones que acarrean los factores económicos o la hiperespecialización en el manejo de plataformas de *software* más sofisticadas, en las que se desconoce el código utilizado lo que acarrea la dependencia de conocimiento e imposibilita generar nuevos aportes.

Una última conclusión es que el ahorro que se obtiene al utilizar el *software* libre se puede destinar a abrir nuevas ofertas académicas o más cupos, donde las autoridades de la institución estimen pertinentes.

6. Referencias

Bolaños, Fabricio. (1996). **Contribución al estudio de la métricas en Análisis de Datos**. Tesis de Licenciatura no publicada. Universidad de Costa Rica. San José, Costa Rica.

Bouroche, Jean Marie y Saporta, Gilbert. (1989). **L'analyse des données**. París: Presses Universitaires de France.

Castillo, William, González, Jorge y Trejos, Javier. (2007). **Análisis Multivariado de Datos**. San José: Universidad de Costa Rica.

Guevara, Rolando. (1980). **Tópicos de Análisis de Datos**. Tesis de Licenciatura no publicada. Universidad de Costa Rica. San José, Costa Rica.

Instituto Costarricense de Turismo. (2005). **Mapa turístico** [**de Costa Rica**]. San José: autor.

Paradis, Emmanuel. (2002). **R para Principiantes**. Consultado el 14 de marzo de 2009, disponible en http://cran.r-project.org/doc/contrib/rdebuts_es.pdf

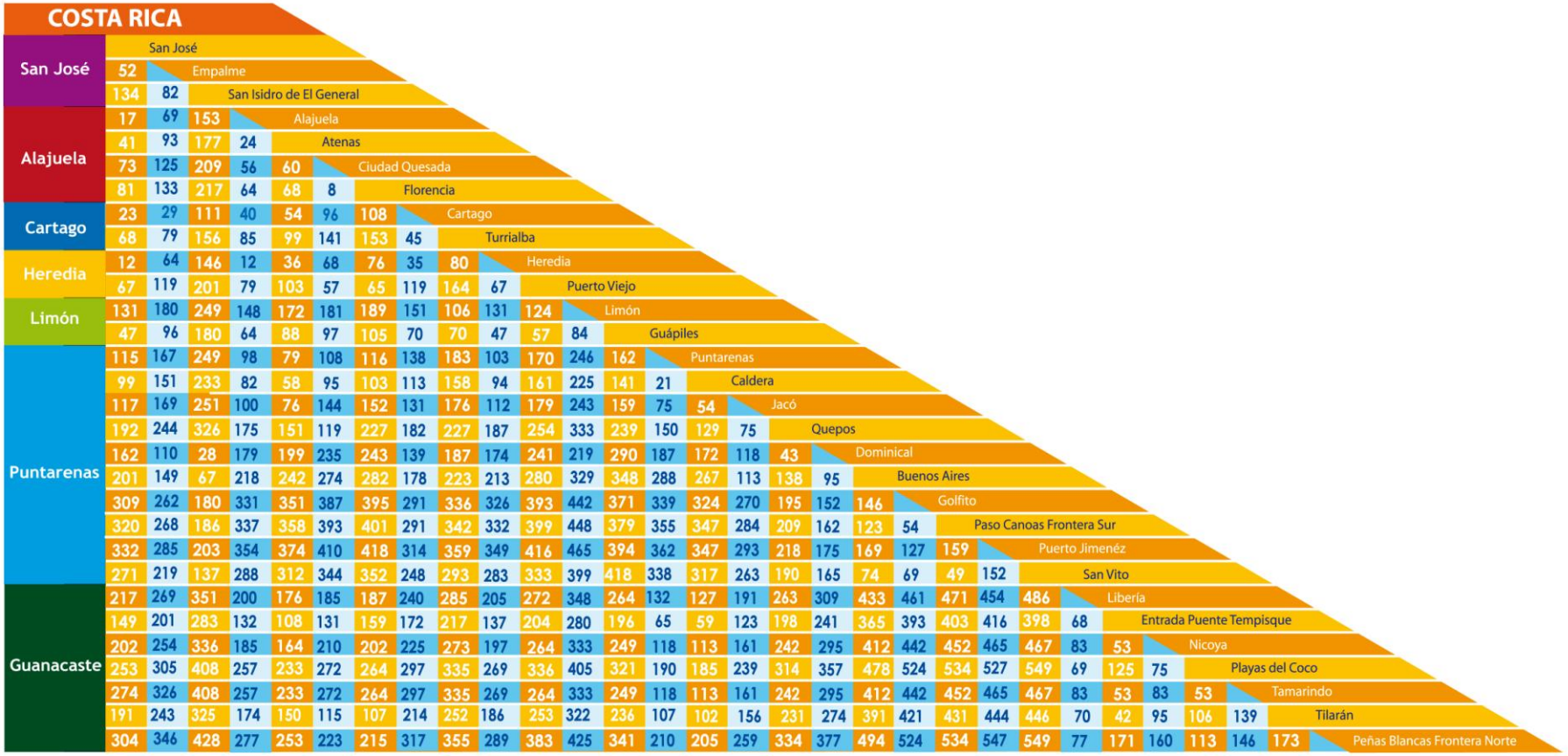
R Development Core Team. (2009). **R: A language and environment for statistical computing [R Foundation for Statistical Computing]**. Vienna, Austria. Recuperado el 14 de marzo de 2009, de http://www.R-project.org

ANEXO N° 1

	Comparación de resultados teóricos vs el programa				
	Bouroche p 39			**C[,1]**	**C[,2]**
Individuo	C_1	C_2		C_1	C_2
1872	-2,90	-1,02		2,9005	1,0244
1880	-2,77	-2,01		2,7674	2,0120
1890	-2,42	-0,22		2,4163	0,2240
1900	-2,06	-0,75		2,0566	0,7552
1903	-2,34	-0,17		2,3379	0,1672
1906	-1,98	-0,63		1,9851	0,6261
1909	-1,91	-0,81		1,9074	0,8122
1912	-1,43	-0,77		1,4311	0,7684
1920	-2,14	-0,96		2,1392	0,9559
1923	-1,14	2,88		1,1429	-2,8839
1926	-1,67	2,61		1,6741	-2,6110
1929	-1,12	1,83		1,1734	-1,8312
1932	0,27	1,96		-0,2706	-1,9593
1935	0,66	2,30		-0,6590	-2,2962
1938	-0,40	1,34		0,4024	-1,3430
1947	1,08	-2,25		-1,0813	2,2512
1950	2,37	-2,17		-2,3728	2,1754
1953	1,20	-1,13		-1,2038	1,3432
1956	2,93	-0,23		-2,9280	0,2307
1959	2,69	-0,14		-2,6862	0,1402
1962	3,06	0,11		-3,0547	-0,1108
1965	3,14	-0,31		-3,1430	0,3112
1968	3,70	0,47		-3,6956	-0,4667
1971	3,24	0,09		-3,2392	-0,0864
Elaboración propia a partir de Bouroche y datos obtenidos con R					

ANEXO N° 2

Tabla de distancias entre algunas ciudades de Costa Rica



Fuente: Mapa de Costa Rica. Instituto Costarricense de Turismo