Ejemplo práctico de uso de ACP

Me tomé la libertad de sacar estadísticas de algunas comunas de Santiago, relacionadas principalmente con los ingresos y la educación (año 1998), para hacer este ejemplo:

Comuna	Habitantes	Tasa de alfabetismo adultos	Años de escolaridad	Tasa de matrícula	Ingreso Per-Cápita (pesos 1998)
Conchalí	380.016	98,7	9,26	71,2	73.950
Providencia	89.324	99,4	13,75	81,8	372.394
Las Condes	345.325	99,3	13,15	82,3	291.573
Ñuñoa	171.462	99,3	12,73	82,7	206.711
La Reina	91.310	98,4	12,61	78,2	201.666
Macul	120.204	98,8	10,85	74,4	112.642
Peñalolén	175.260	93,6	7,43	65,6	45.441
La Florida	309.140	96	9,23	66	86.019
San Joaquín	124.043	96,9	9,25	71,2	58.305
La Granja	123.518	95,5	7,91	65,4	49.056
La Pintana	144.628	96,6	7,81	62,9	43.936
San Ramón	121.286	91,5	7,2	67,4	51.089
San Miguel	181.048	99,5	10,69	72,4	86.335
La Cisterna	358.540	96,6	8,26	64,9	56.170
Est. Central	14.085	95,8	9,06	72,6	60.817
Maipú	305.140	99,5	10,54	73,4	94.061
Qta. Normal	92.834	98,3	8,91	73,5	66.505
Lo Prado	110.952	96,6	8,74	70,8	62.039
Pudahuel	132.240	95,5	7,79	65,9	52.683
Cerro Navia	155.428	93,1	7,34	65,7	48.202
Renca	151.632	97,3	9,01	69	55.373

(fuente: PNUD)

El objetivo es caracterizar e identificar a las comunas de acuerdo a un conjunto de índices que explique la mayor parte de la variabilidad de los datos. Asimismo, interesa descubrir las relaciones entre las variables originales y el agrupamiento de las comunas. En este ejemplo sólo se tomaron 22 comunas dentro de las más representativas de Santiago.

Definamos las variables:

habit: Habitantes **alfab_adult**: Tasa de alfabetismo **años_esc**: Años de escolaridad

tasa_matric: Tasa de matrícula **ipercap1998**: Ingreso Per-Cápita

El primer paso consiste en obtener las estadísticas básicas para cada variable, es decir, medias y desviaciones:

	Media	Desv. Estándar
habit	176067,381	98944,557
alfab_adult	96,962	2,216
años_esc	9,596	1,967
tasa_matric	71,300	5,850
ipercap1998	103569,857	87120,313

A continuación estandarizamos los datos (restar media y dividir por deviación estándar), lo que nos da:

comuna	**habit**	**alfab_adult**	**años_esc**	**tasa_matric**	**ipercap1998**
Conchalí	2,06	0,784	-0,171	-0,017	-0,340
Providencia	-0,877	1,100	2,112	1,795	3,086
Las Condes	1,711	1,055	1,807	1,880	2,158
Ñuñoa	-0,047	1,055	1,593	1,949	1,184
La Reina	-0,857	0,649	1,532	1,179	1,126
Macul	-0,565	0,829	0,637	0,530	0,104
Peñalolén	-0,008	-1,517	-1,101	-0,974	-0,667
La Florida	1,345	-0,434	-0,186	-0,906	-0,201
San Joaquín	-0,526	-0,028	-0,176	-0,017	-0,520
La Granja	-0,531	-0,660	-0,857	-1,008	-0,626
La Pintana	-0,318	-0,163	-0,908	-1,436	-0,685
San Ramón	-0,554	-2,465	-1,218	-0,667	-0,602
San Miguel	0,050	1,145	0,556	0,188	-0,198
La Cisterna	1,844	-0,163	-0,679	-1,094	-0,544
Est. Central	-1,637	-0,524	-0,273	0,222	-0,491
Maipú	1,304	1,145	0,480	0,359	-0,109
Qta. Normal	-0,841	0,604	-0,349	0,376	-0,425
Lo Prado	-0,658	-0,163	-0,435	-0,085	-0,477
Pudahuel	-0,443	-0,660	-0,918	-0,923	-0,584
Cerro Navia	-0,209	-1,743	-1,147	-0,957	-0,636
Renca	-0,247	0,153	-0,298	-0,393	-0,553

Acto seguido obtenemos la matriz de correlaciones R, que es el resultado de multiplicar la última matriz obtenida, traspuesta, por ella misma (X^tX) y dividir por el número de observaciones, lo que nos da:

	habit	alfab_adult	años_esc	tasa_matric	ipercap1998
habit	1,000	0,234	0,074	-0,052	0,045
alfab_adult	0,234	1,000	**0,792**	**0,703**	**0,563**
años_esc	0,074	**0,792**	1,000	**0,933**	**0,906**
tasa_matric	-0,052	**0,703**	**0,933**	1,000	**0,848**
ipercap1998	0,045	**0,563**	**0,906**	**0,848**	1,000

Se puede observar que en negrita están las correlaciones más significativas, lo que en cierto modo es esperable, por ejemplo en el caso de las variables alfabetización, años de escolaridad y tasa de matrículas, junto con el ingreso Per-Cápita las correlaciones esperadas son altas.

El paso siguiente es resolver el sistema:

$$Ru = \lambda u$$

y obtener los valores y vectores propios asociados:

	U1	U2	U3	U4	U5
Valor propio	3,395	1,059	0,410	0,110	0,027
% varianza de cada componente	67,902	21,171	8,199	2,195	0,533
% acumulado	67,902	**89,074**	97,273	99,467	100,000

Vectores propios:

	U1	U2	U3	U4	U5
habit	0,059	0,951	-0,263	0,151	0,000
alfab_adult	0,450	0,231	0,775	-0,280	-0,255
años_esc	0,537	-0,037	-0,047	-0,061	0,839
tasa_matric	0,514	-0,172	-0,054	0,790	-0,282
ipercap1998	0,492	-0,105	-0,570	-0,520	-0,389

Debe notarse que como usamos la matriz de correlaciones, la suma de los valores propios debe ser igual a....5 (el número de variables), que es igual a la suma de las varianzas de la matriz de correlaciones (solo unos en la diagonal).

Notemos que la varianza aportada por los dos primeros valores propios acumula el 89,074% de toda la varianza de los datos, por lo que nos quedaremos con los dos primeros vectores propios para construir nuestras componentes principales.

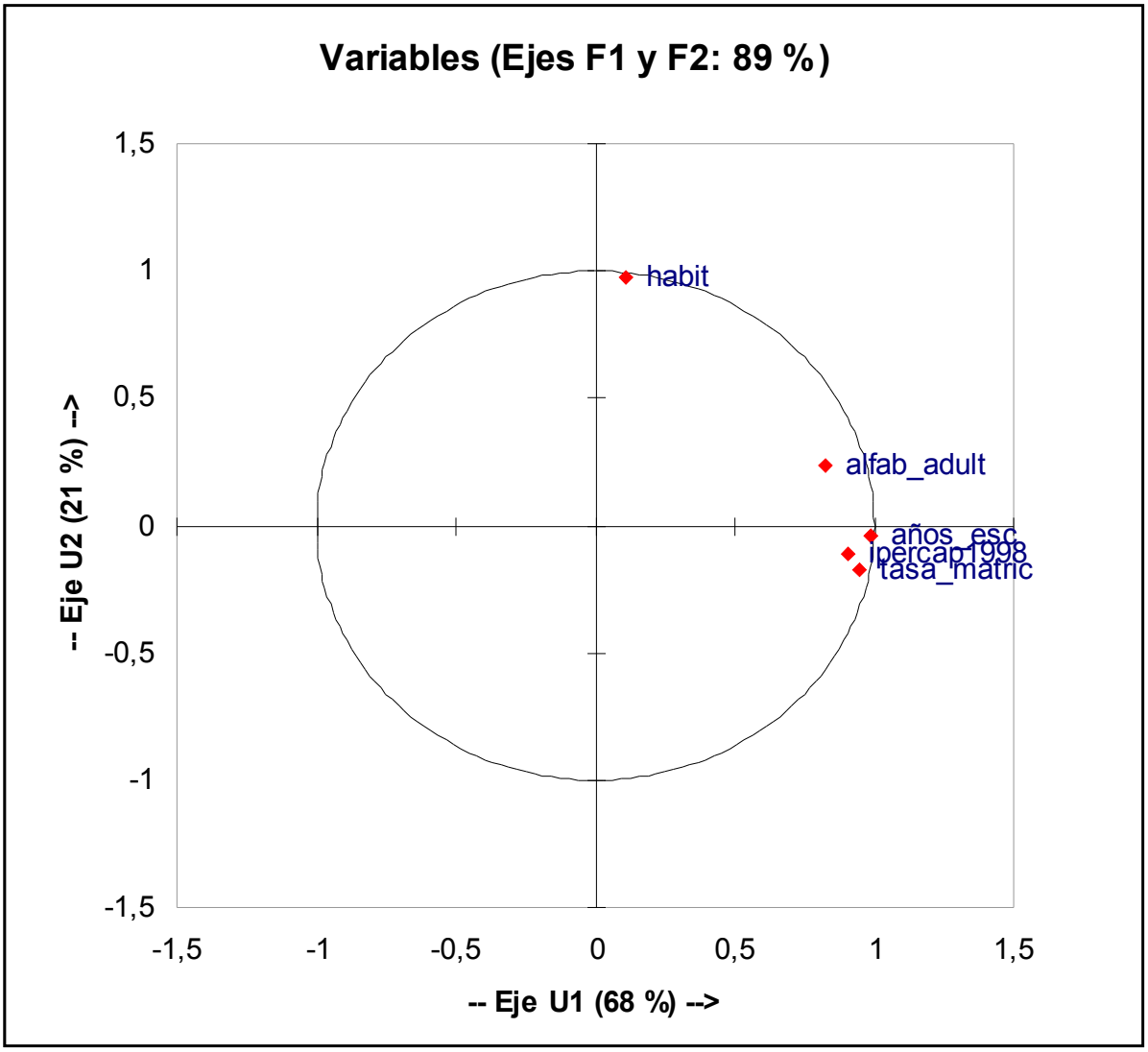
Antes de hacer interpretaciones, debemos analizar las nuevas variables creadas a través del círculo de correlaciones, el cual se contruye usando las fórmulas vistas en clase. Por ejemplo, para la variable **habit**:

$$\rho_{habit,U1} = \frac{u_{habit,1}\sqrt{\lambda_1}}{\sigma_{habit}} = \frac{0,059 \cdot \sqrt{3,395}}{1} = 0,108$$

$$\rho_{habit,U2} = \frac{u_{habit,2}\sqrt{\lambda_2}}{\sigma_{habit}} = \frac{0,951 \cdot \sqrt{1,059}}{1} = 0,978$$

IMPORTANTE: Como estamos trabajando con la matriz de correlaciones, y las observaciones están estandarizadas, las varianzas de las variables estndarizadas son 1.

Asimismo, podemos escribir las correlaciones de cada una de las variables estandarizadas y construir el círculo de correlaciones:



El círculo de correlaciones es clarificador, en efecto, las variables matrícula,ingreso, escolaridad y alfabetización se encuentran muy concentradas en el primer eje, al cual podemos denominar "nivel educativo" o "capacidad de educación", mientras que el segundo eje queda casi completamente explicado por la cantidad de personas que viven en una comuna y un poco por el nivel de alfabetización. A este eje le podríamos denominar "población escolarizada"[1].

¿Se acuerdan de las altas correlaciones mostradas por las variables originales al principio?. Bueno, esto queda ratificado por la proximidad que muestran unas con otras en el gráfico anterior.

¿Qué pasa con las observaciones?

1 Esta es quizá la parte más subjetiva del análisis, ya que depende de como uno interprete la relación y agrupamiento de las variables.

Para analizarlas, debemos construir las componentes principales para cada observación. Por ejemplo, para Conchalí, las primera componente principales es:

$$C_{conchali,1} = U_1^t \cdot R_{conchali}$$

Donde U_1 es el primer eje principal, y $R_{conchali}$ es el vector que contiene los valores de las variables para Conchalí en la matriz estandarizada, esto es:

$C_{conchali,1}$ = 0,059*2,061+0,45*0,784+0,537*(-0,171)+0,514*(-0,017)+0,492*(-0,34) = 0,206

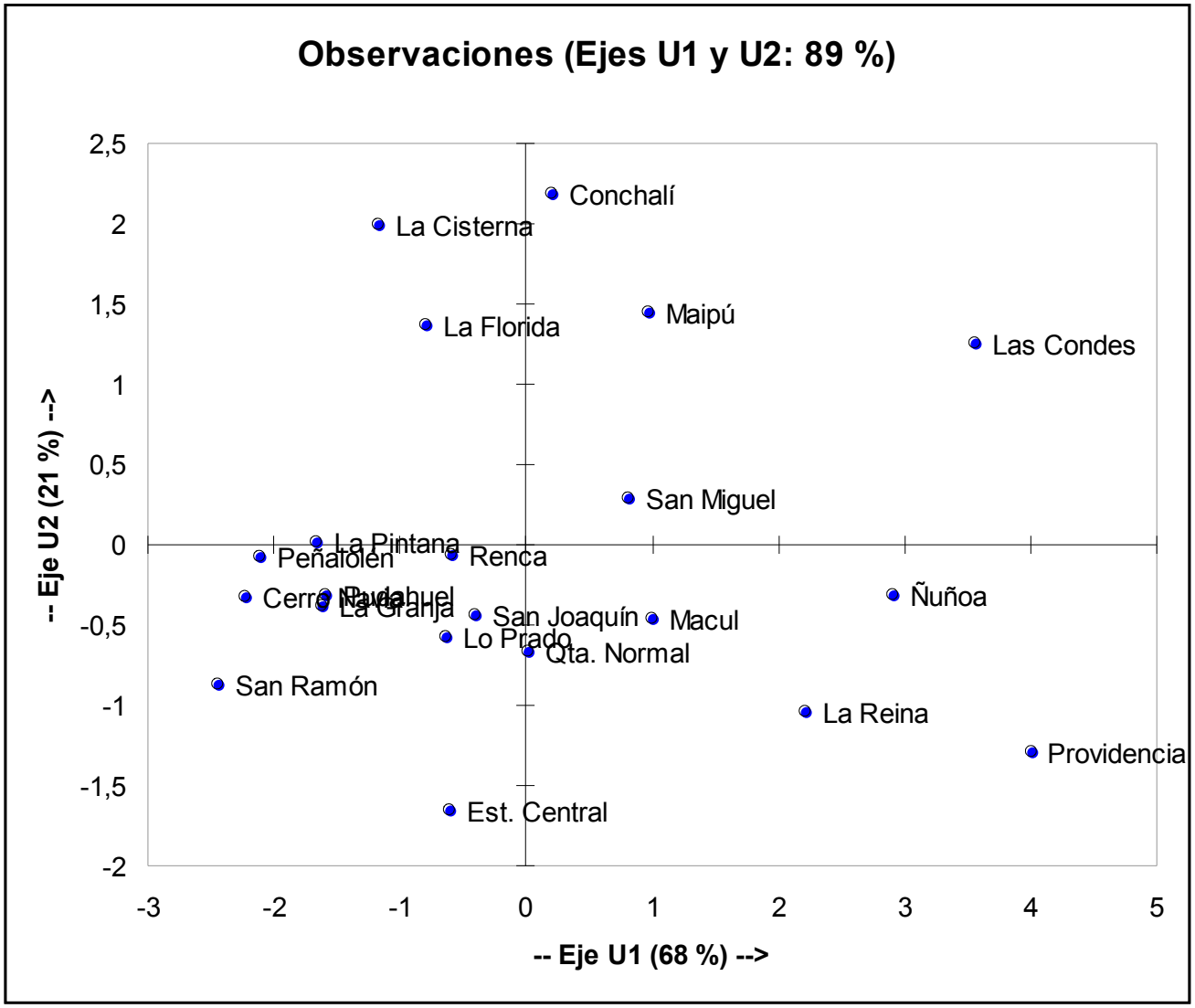
Asimismo, su segunda componente principal es:

$C_{conchali,2}$ = 0,951*2,061+0,231*0,784+(-0,037)*(-0,171)+(-0,172)*(-0,017)+(-0,105)*(-0,34) = 2,187

Luego el vector de componentes principales para Conchalí es:

$$C_{conchali} = \begin{pmatrix} 0,206 \\ 2,187 \end{pmatrix}$$

Habiendo calculado las componentes principales para cada observación, podemos hacer un gráfico con los dos ejes:



Habiendo realizado la identificación de la componentes, podemos distinguir que las comunas con mayor acceso o capacidad de educación son principalmente las comunas del Barrio Alto o de mejores ingresos (Las Condes, Providencia, La Reina, Ñuñoa) lo cual es intuitivamente esperable. Lo que las diferencia, sin embargo, es la cantidad de población escolarizada presente en ellas, siendo la más alta la comuna de Las Condes, seguida por Ñuñoa (pero de lejos). En el otro lado del primer eje se encuentran las comunas más pobres y con menos acceso a educación: San Ramón, Cerro Navia, La Pintana y Peñalolén. En general, las comunas más pobres se caracterizan por tener un bajo nivel de ingresos y un bajo nivel de escolaridad entre los adultos. El gráfico casi habla por si solo una vez que se ha comprendido el significado de cada eje.

Adicionalmente, podemos analizar las correlaciones al cuadrado para estudiar el ajuste de cada variable respecto de la construcción de cada eje. De hecho, la distancia de cada variable al origen del círculo representa la bondad del ajuste de la variable en el plano conformado por las componentes principales,

es decir, es un coeficiente de determinación, o sea un símil a R^2.

Cuadrados de las correlaciones:

	U1	U2	U3	U4	U5
habit	0,012	0,957	0,028	0,003	0,000
alfab_adult	0,687	0,057	0,246	0,009	0,002
años_esc	0,979	0,001	0,001	0,000	0,019
tasa_matric	0,897	0,031	0,001	0,068	0,002
ipercap1998	0,821	0,012	0,133	0,030	0,004

Contribución de cada variable al factor (%):

	U1	U2	U3	U4	U5
habit	0,345	90,444	6,923	2,289	0,000
alfab_adult	20,226	5,353	60,079	7,864	6,477
años_esc	28,821	0,135	0,221	0,371	70,453
tasa_matric	26,417	2,963	0,288	62,384	7,949
ipercap1998	24,192	1,106	32,489	27,09	15,121

¿Qué sucede si introducimos una nueva comuna, como por ejemplo, Puente Alto?

Los valores de Puente Alto son:

habit	alfab_adult	años_esc	tasa_matric	ipercap1998
386536	98	10,18	73,6	94078

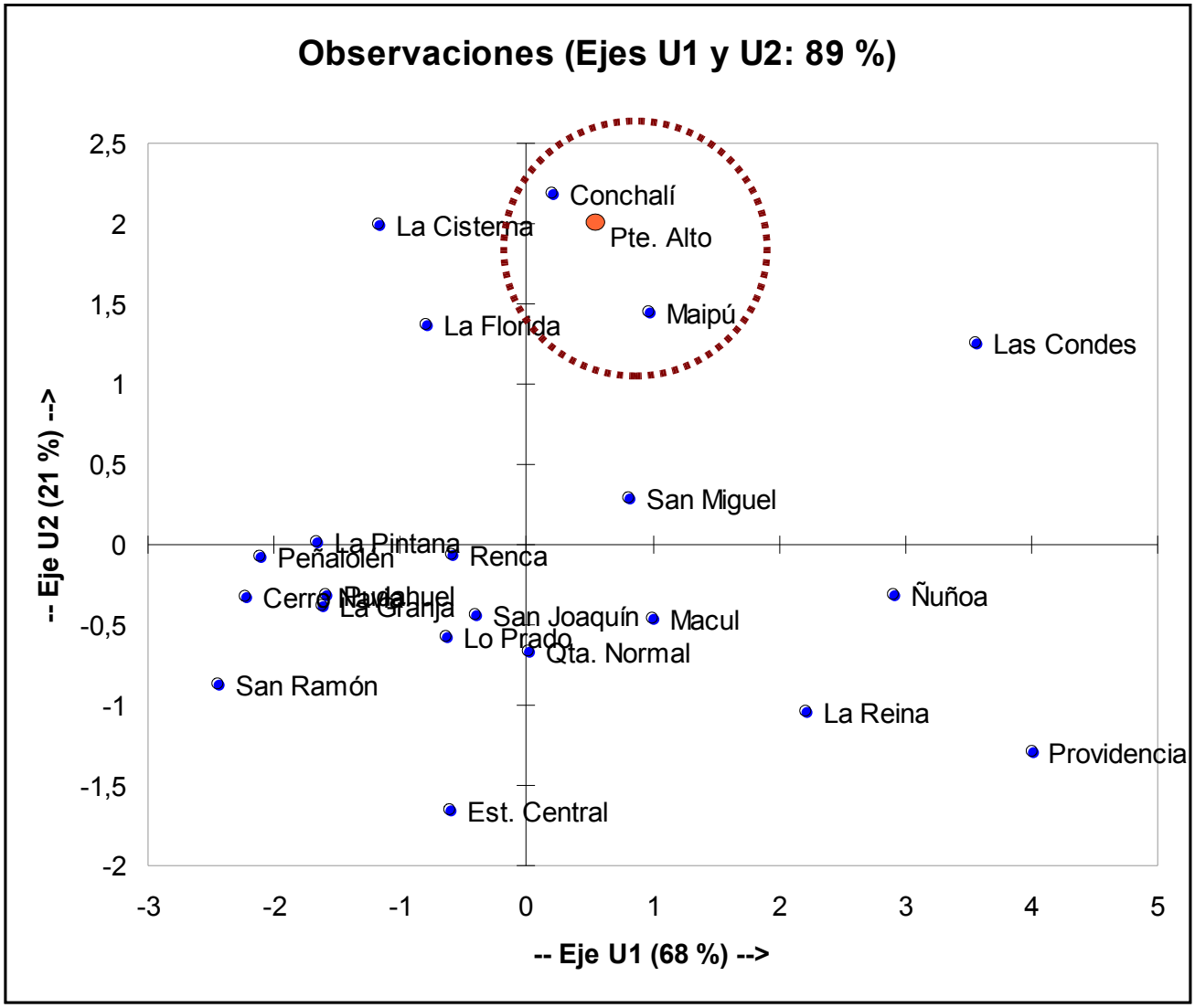
Las que estandarizadas dan:

habit	alfab_adult	años_esc	tasa_matric	ipercap1998
2,13	0,47	0,3	0,39	-0,11

Las componentes principales asociadas son:

$$C_{pte.Alto,1} = 0,125 \qquad C_{pte.Alto,2} = 2,023$$

Lo que ubica a Pte. Alto en el gráfico de los ejes cerca de Conchalí y Maipú:



Se puede observar con esto que el ACP puede ser útil, además, como herramienta de clasificación.

Eso es todo.

Suerte. Rodrigo Abt B.

26/11/03