ANÁLISIS ESTADÍSTICO MULTIVARIADO

Análisis Multivariado

Introducción

- El objetivo de este procedimiento es agrupar elementos en grupos homogéneos en función de las similitudes entre ellos.
- Estos métodos se suelen denominar de clasificación automática o no supervisada.
- El análisis de conglomerados o clusters estudia 3 tipos de problemas:

1. Partición de los datos
2. Construcción de jerarquías
3. Clasificación de las variables

Método de las k-medias

- Consideremos una muestra de n elementos y p variables. El objetivo es dividir esta muestra en k grupos.

El algoritmo de las k-medias (*kmeans*) requiere las siguientes etapas:

1. Seleccionar k puntos como centros de los grupos iniciales.
2. Calcular las distancias euclídeas de cada elemento con respecto a los k centros y asignar cada elemento al grupo cuyo centro esté más próximo.
3. Definir un criterio de optimalidad y comprobar si reasignando alguno de los elementos mejora el criterio.
4. Si no es posible mejorar el criterio de optimalidad, fin del proceso.

Método de las k-medias

- El criterio de homogeneidad u optimalidad que se utiliza en el algoritmo de medias es minimizar la suma de cuadrados dentro de los grupos (SCDG) para todas las variables:

$$SCDG = \sum_{g=1}^{k} \sum_{j=1}^{p} \sum_{i=1}^{n_g} (x_{ijg} - \bar{x}_{jg})^2$$

donde x_{ijg} es el valor de la variable j en el elemento i del grupo g y $\bar{x}_{jg}$ es la media de ese grupo.

- Este criterio es equivalente a la suma ponderada de las variancias de las variables en los grupos ya que puede escribirse como:

$$SCDG = \sum_{g=1}^{k} \sum_{j=1}^{p} n_g s_{jg}^2$$

donde n_g es el número de elementos en el grupo g y s_{jg}^2 es la variancia de la variable j en dicho grupo.

Método de las k-medias

- Las variancias de las variables en los grupos son claramente una medida de la heterogeneidad de la clasificación y al minimizarlas obtendremos grupos más homogéneos.
- Un criterio alternativo de homogeneidad sería minimizar las distancias al cuadrado entre los puntos y el centro de su grupo.
- Si consideramos distancias euclídeas, este criterio resulta:

$$\sum_{g=1}^{k}\sum_{i=1}^{n_g}(x_{ig}-\bar{x}_g)'(x_{ig}-\bar{x}_g)=\sum_{g=1}^{k}\sum_{i=1}^{n_g}d^2(i,g)$$

donde $d^2(i,g)$ es el cuadrado de la distancia euclídea entre el elemento i del grupo g y el centro de su grupo.

Método de las k-medias

Ambos criterios son idénticos:

$$\begin{aligned}
\sum_{g=1}^{k}\sum_{i=1}^{n_g} d^2(i,g) &= \sum_{g=1}^{k}\sum_{i=1}^{n_g} tr[d^2(i,g)] \\
&= \sum_{g=1}^{k}\sum_{i=1}^{n_g} tr[(x_{ig}-\bar{x}_g)'(x_{ig}-\bar{x}_g)] \\
&= \sum_{g=1}^{k}\sum_{i=1}^{n_g} tr[(x_{ig}-\bar{x}_g)(x_{ig}-\bar{x}_g)'] \\
&= tr[W]
\end{aligned}$$

donde $W = \sum_{g=1}^{k}\sum_{i=1}^{n_g}(x_{ig}-\bar{x}_g)(x_{ig}-\bar{x}_g)'$.

Método de las k-medias

- La minimización de la $SCDG$ requeriría calcularla para todas las posibles particiones de los elementos en k grupos, lo cual es imposible salvo para n pequeño.
- El algoritmo de k-medias busca la partición óptima con la restricción de que en cada iteración solo se permite mover un elemento de un grupo a otro.

El algoritmo funciona de la siguiente manera:

1. El proceso comienza a partir de una asignación inicial.
2. Comprobar si reasignando algún elemento se reduce la $tr(W)$.
3. Si es posible reducir la $tr(W)$ moviendo un elemento hacerlo, recalcular las medias de los dos grupos afectados por el cambio y volver al paso (2). Si no es posible reducir la $tr(W)$, terminó el proceso.

Método de las k-medias

- Se han propuesto diferentes métodos para seleccionar el número de grupos.
- Un procedimiento que se utiliza bastante (sin mucha justificación) es realizar un test F de reducción de variabilidad, comparando la $SCDG$ con k grupos con respecto a la suma obtenida con $k+1$ grupos y calculando la reducción relativa de variabilidad al agregar un grupo adicional.
- El estadístico es:

$$F = \frac{SCDG(k) - SCDG(k+1)}{SCDG(k)/(n-k-1)}$$

  En la práctica, el valor obtenido de F se compara con el valor de la distribución F con p y $p(n-k-1)$ grados de libertad.
- Una regla empírica que puede encontrarse en los manuales es agregar un grupo más si este cociente es mayor que 10.

Método de las k-medias

- Otra alternativa para determinar el número adecuado de clusters es comparar los resultados del estadístico de Calinsky-Harabasz:

$$C - H = \frac{tr(B)/(k-1)}{tr(W)/(n-k)}$$

En la práctica, se elige la cantidad de clusters que da por resultado el valor más grande del estadístico.

Métodos Jerárquicos

- Los métodos jerárquicos se plantean a partir de una matriz de distancias o similaridades entre los elementos de la muestra y construyen una jerarquía basada en estas distancias.
- Si todas las variables son continuas, la distancia más utilizada es la euclídea entre las variables estandarizadas de manera univariada.
- En general no es recomendable utilizar la distancia de Mahalanobis ya que la única matriz de variancias y covariancias disponible es la de toda la muestra, que puede mostrar correlaciones entre las variables muy distintas de las que existen dentro de los grupos.
- Cuando la información disponible corresponde tanto a variables continuas como atributos es posible trabajar con una matriz de similaridades (ver Anexo).

Métodos Jerárquicos

- Dada una matriz de distancias o similitudes se desea clasificar los elementos en una jerarquía. Los algoritmos existentes funcionan de manera tal que los elementos son sucesivamente asignados a los grupos pero la asignación es irrevocable.
- Existen dos tipos de algoritmos:
  - De aglomeración: parten de los elementos individuales y los van agregando en grupos.
  - De división: parten del conjunto total de elementos y lo van dividiendo sucesivamente hasta llegar a los elementos individuales.

Métodos Jerárquicos

Los diversos algoritmos de este tipo se diferencian en la forma de calcular las distancias entre grupos pero tienen la misma estructura:

1. Comenzar con tantas clases como elementos, n. las distancias entre clases son las distancias entre los elementos originales.
2. Seleccionar los dos elementos mas próximos en la matriz de distancias y formar con ellos una clase.
3. Sustituir los dos elementos seleccionados en 2 para definir la clase por un nuevo elemento que los represente. Las distancias entre este nuevo elemento y los anteriores se calculan con alguno de los criterios que veremos a continuación.
4. Volver al segundo paso y repetir hasta que todos los elementos hayan sido agrupados en una sola clase.

Métodos de cálculo de distancias - Encadenamientos

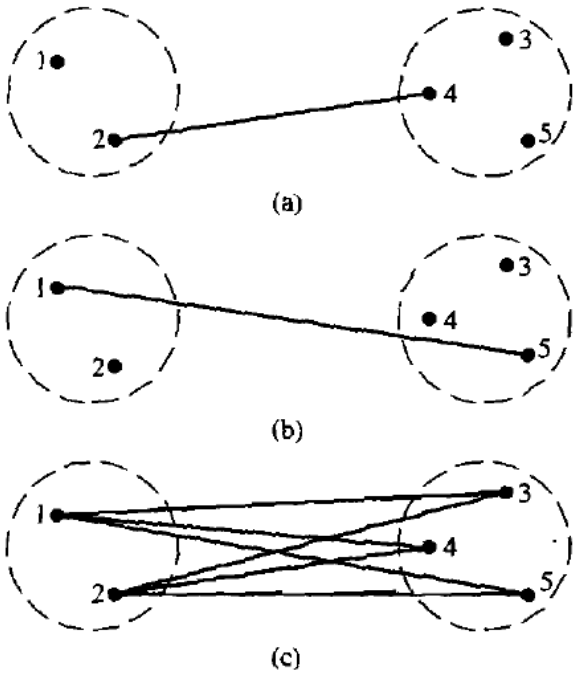
Supongamos que tenemos un grupo A con n_a elementos y un grupo B con n_b elementos. Ambos grupos se fusionan para crear un grupo (AB) con $n_a + n_b$ elementos. La distancia entre el nuevo grupo (AB) y otro grupo C con n_c elementos se puede calcular por algunas de las 4 reglas siguientes:

- Encadenamiento simple o vecino más cercano: la distancia entre los dos grupos es la menor de las distancias entre los grupos antes de la fusion:

$$d(C, AB) = min[d(C, A); d(C, B)]$$

Este criterio solo depende del orden de las distancias por lo tanto sera invariante ante transformaciones monotonas. Este criterio tiende a producir grupos alargados que pueden incluir elementos muy diferentes en los extremos.

Cálculo de distancias - encadenamientos



Cluster distance

d_{24}

d_{15}

$$\frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{6}$$

Figure 12.2 Intercluster distance (dissimilarity) for (a) single linkage, (b) complete linkage, and (c) average linkage.

Métodos de cálculo de distancias - Encadenamientos

- Encadenamiento completo o vecino más alejado: la distancia entre los dos grupos es la mayor de las distancias entre los grupos antes de la fusión:

$$d(C, AB) = max[d(C, A); d(C, B)]$$

Este criterio también será invariante ante transformaciones monótonas y tiende a producir grupos esféricos.

- Media de los grupos: la distancia entre los dos grupos es la media ponderada de las distancias entre los grupos antes de la fusión:

$$d(C, AB) = \frac{n_a}{n_a + n_b} d(C, A) + \frac{n_b}{n_a + n_b} d(C, B)$$

Este criterio no es invariante ante transformaciones monótonas.

Métodos de cálculo de distancias - Encadenamientos

- Método del centroide: se aplica cuando todas las variables son continuas. La distancia entre los dos grupos es igual a la distancia euclídea entre sus centros:

$$d(C, AB) = d(\bar{x}_C, \bar{x}_{AB})$$

donde se toman como centros los vectores de medias de las observaciones que pertenecen al grupo. Se puede demostrar que el cuadrado de la distancia euclídea de un grupo C con respecto a la fusión (AB) es:

$$d^2(C, AB) = \frac{n_a}{n_a + n_b} d^2(C, A) + \frac{n_b}{n_a + n_b} d^2(C, B) - \frac{n_a n_b}{n_a + n_b} d^2(A, B)$$

Método de Ward

- Es un procedimiento diferente para construir un agrupamiento jerárquico. La diferencia con los métodos anteriores es que ahora se parte directamente de la información de los elementos (en lugar de partir de una matriz de distancias) y se define una medida global de la heterogeneidad de una agrupación de observaciones en grupos.
- Esta medida representada por W es la suma de las distancias euclídeas al cuadrado entre cada elemento y la media de su grupo:

$$W = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{ig} - \bar{x}_g)(x_{ig} - \bar{x}_g)'$$

- El criterio comienza suponiendo que cada observación forma un grupo, $g = n$ y por lo tanto W es igual a 0.
- Luego, se fusionan los elementos que produzcan el incremento mínimo de W, lo cual implica tomar los elementos más próximos considerando la distancia euclídea.

Método de Ward

- En la segunda etapa tenemos $(n-1)$ grupos, $(n-2)$ de ellos contienen un solo elemento y el grupo restante tiene 2 elementos. Decidimos nuevamente unir aquellos dos grupos que produzcan el mínimo incremento en W, y repetimos el procedimiento hasta que obtenemos un único grupo de n elementos.
- Los valores de W van indicando el crecimiento del criterio al formar los grupos y pueden utilizarse para decidir cuantos grupos naturales contienen nuestros datos.
- En cada etapa los grupos de deben unirse para minimizar W son aquellos que satisfacen:

$$min \frac{n_a n_b}{n_a + n_b} (\bar{x}_a - \bar{x}_b)'(\bar{x}_a - \bar{x}_b)$$

Dendrogramas

- El dendrograma, o gráfico jerárquico, es una representación gráfica del resultado del proceso de agrupamiento en forma de árbol.
- Los criterios para definir distancias que hemos presentado tienen la propiedad de que si consideramos tres grupos A, B y C se verifica:

$$d(A, C) \leq max[d(A, B); d(B, C)]$$

y una medida de distancia que tiene esta propiedad se denomina ultramétrica. Esta es una propiedad más fuerte que el cumplimiento de la desigualdad triangular, una ultramétrica siempre es una distancia.

- En efecto, si $d(A, C)$ es menor o igual que el máximo entre $d(A, B)$ y $d(B, C)$ forzosamente será menor o igual que la suma $d(A, B) + d(B, C)$.
- El dendrograma es la representación de una ultramétrica.

Dendrograma

- El dendrograma se construye como sigue:
  - En la parte inferior del grafico se disponen los n elementos iniciales.
  - Las uniones entre elementos se indican por lineas rectas, dos dirigidas a los elementos que se unen y que son perpendiculares al eje de los elementos y una paralela a este eje que se situa al nivel en que se unen.
  - El proceso se repite hasta que todos los elementos estan conectados por lineas rectas.
- Si cortamos el dendrograma a un nivel de distancia dado obtenemos una clasificacion de los elementos en los grupos existentes a ese nivel y los elementos que los forman.

Ejemplo: Clusters jerárquicos

La matriz de distancias de los elementos es la siguiente:

Distancias	A	B	C	D
A	0	1	4	2.5
B		0	2	3
C			0	4
D				0

- ✓ **Método 1: encadenamiento simple**

Partimos de una clasificación en 4 clases o grupos (tantas como elementos diferentes) y unimos las clases más cercanas. Los elementos / grupos más próximos son A y B, por lo tanto formamos una clase uniendo ambos elementos. En la siguiente etapa tendremos 3 grupos de elementos. Calculamos las distancias entre (A,B) y C y entre (A,B) y D:

$$d(AB,C) = min\{d(A,C)\,;\, d(B,C)\} = min\{4\,,2\} = 2$$

$$d(AB,D) = min\{d(A,D)\,;\, d(B,D)\} = min\{2.5\,,3\} = 2.5$$

Distancias	(A,B)	C	D
(A,B)	0	2	2.5
C		0	4
D			0

Ejemplo: Clusters jerárquicos

Distancias	(A,B)	C	D
(A,B)	0	2	2.5
C		0	4
D			0

Los grupos más próximos son (A,B) y C, por lo tanto formamos una clase uniendo ambos grupos. En la siguiente etapa tendremos 2 grupos de elementos. Debemos calcular las distancias entre (A,B,C) y D:

$$d(ABC, D) = min\{d(AB, D)\,;\, d(C, D)\} = min\{2.5\,, 4\} = 2.5$$

Distancias	(A,B,C)	D
(A,B,C)	0	2.5
D		0

Finalmente unimos las dos grupos y obtenemos el conglomerado (A,B,C,D).

Ejemplo: Clusters jerárquicos

La matriz de distancias de los elementos es la siguiente:

Distancias	A	B	C	D
A	0	1	4	2.5
B		0	2	3
C			0	4
D				0

- ✓ **Método 2: encadenamiento completo**

Al igual que en el caso anterior partimos de una clasificación en 4 clases o grupos (tantas como elementos diferentes) y unimos las clases mas cercanas. Los elementos / grupos las próximos son A y B, por lo tanto formamos una clase uniendo ambos elementos. En la siguiente etapa tendremos 3 grupos de elementos. Calculamos las distancias entre (A,B) y C y entre (A,B) y D:

$$d(AB, C) = max\{d(A, C)\,;\, d(B, C)\} = max\{4\,, 2\} = 4$$

$$d(AB, D) = max\{d(A, D)\,;\, d(B, D)\} = max\{2.5\,, 3\} = 3$$

Distancias	(A,B)	C	D
(A,B)	0	4	3
C		0	4
D			0

Ejemplo: Clusters jerárquicos

Distancias	(A,B)	C	D
(A,B)	0	4	3
C		0	4
D			0

Los grupos más próximos son (A,B) y D, por lo tanto formamos una clase uniendo ambos grupos. Calculamos la distancia entre (A,B,D) y C:

$$d(ABD, C) = max\{d(AB, C)\,;\, d(D, C)\} = max\{4\,, 4\} = 4$$

Distancias	(A,B,D)	C
(A,B,D)	0	4
C		0

Finalmente unimos las dos grupos y obtenemos el conglomerado (A,B,C,D).

Ejemplo: Clusters jerárquicos

La matriz de distancias de los elementos es la siguiente:

Distancias	A	B	C	D
A	0	1	4	2.5
B		0	2	3
C			0	4
D				0

- ✓ **Método 3: media de los grupos**

Como primer paso unimos los elementos / grupos mas próximos, A y B. En la siguiente etapa tendremos 3 grupos de elementos. Calculamos las distancias entre (A,B) y C y entre (A,B) y D:

$$d(AB, C) = ½\, d(A, C) + ½.\, d(B, C) = 3$$

$$d(AB, D) = ½.\, d(A, D) + ½.\, d(B, D) = 2.75$$

Distancias	(A,B)	C	D
(A,B)	0	3	2.75
C		0	4
D			0

Ejemplo: Clusters jerárquicos

Distancias	(A,B)	C	D
(A,B)	0	3	2.75
C		0	4
D			0

Los grupos más próximos son (A,B) y D, por lo tanto formamos una clase uniendo ambos grupos. Calculamos la distancia entre (A,B,D) y C:

$$d(ABD, C) = 2/3.d(AB, C) + 1/3.d(D, C)\} = 3.33$$

Distancias	(A,B,D)	C
(A,B,D)	0	3.33
C		0

Finalmente unimos las dos grupos y obtenemos el conglomerado (A,B,C,D).

Ejemplo: Algunos indicadores socioeconómicos en centros urbanos de Argentina

- Fuente: Encuesta Permanente de Hogares, cuarto trimestre y segundo semestre de 2020.

Lista de variables

X1	Tasa de actividad
X2	Tasa de empleo
X3	Tasa de desocupación
X4	Porcentaje de hogares por debajo de la línea de pobreza
X5	Porcentaje de personas por debajo de la línea de pobreza
X6	Porcentaje de hogares con acceso a computadora
X7	Porcentaje de hogares con acceso a internet

Estadísticos descriptivos

Vars	Promedio	Desv. Std	CV	Asimetría	Kurtosis
X1	44.11	3.63	0.08	-0.84	2.74
X2	40.60	3.02	0.07	-0.71	3.09
X3	7.84	3.10	0.40	0.42	-0.80
X4	28.81	5.81	0.20	-0.24	1.73
X5	38.25	7.24	0.19	-0.61	2.19
X6	63.90	8.40	0.13	0.59	-0.32
X7	89.95	4.39	0.05	-0.41	-0.52

Matriz de datos

Aglomerados urbanos	id	X1	X2	X3	X4	X5	X6	X7
Ciudad Autónoma de Buenos Aires	1	51.4	47.7	7.2	12.2	16.5	82.8	96.0
Partidos del GBA	2	43.4	37.3	14.1	40.9	51.0	59.1	88.6
Gran Mendoza	3	49.5	44.3	10.6	32.6	44.0	59.2	93.1
Gran San Juan	4	44.1	41.9	5.2	24.9	34.8	52.8	80.5
Gran San Luis	5	44.8	42.6	4.9	32.4	40.6	77.6	91.8
Corrientes	6	40.1	37.4	6.7	32.2	42.9	55.2	83.1
Formosa	7	32.1	30.7	4.2	25.7	36.4	52.0	84.6
Gran Resistencia	8	42.3	40.0	5.3	40.3	53.6	57.5	92.9
Posadas	9	46.0	43.1	6.4	27.6	37.7	59.9	90.1
Gran Catamarca	10	43.2	40.6	6.1	28.7	35.7	56.3	86.2
Gran Tucumán - Tafí Viejo	11	43.3	39.2	9.5	33.8	43.5	51.4	92.6
Jujuy - Palpalá	12	43.1	41.4	4.0	27.4	37.7	67.2	93.1
La Rioja	13	41.4	39.6	4.3	25.3	35.3	70.9	93.4
Salta	14	44.8	40.7	9.0	31.2	41.7	61.9	91.3
Santiago del Estero - La Banda	15	41.1	39.5	3.9	31.4	39.4	60.1	92.6
Bahía Blanca - Cerri	16	48.1	43.4	9.7	18.7	24.0	63.5	86.5
Concordia	17	39.7	36.3	8.6	39.3	49.5	54.3	84.5
Gran Córdoba	18	47.9	41.7	13.0	29.5	40.8	62.7	90.5
Gran La Plata	19	48.0	43.6	9.1	24.0	31.7	69.9	91.0
Gran Rosario	20	46.7	40.3	13.6	29.1	38.3	61.0	83.8
Gran Paraná	21	41.4	39.8	4.0	30.4	40.9	76.4	93.2
Gran Santa Fe	22	44.9	41.4	7.8	28.0	39.8	61.5	91.7
Mar del Plata	23	47.8	42.5	11.1	30.5	41.1	63.8	87.7
Río Cuarto	24	46.3	42.0	9.2	27.2	39.2	61.8	91.3
Santa Rosa - Toay	25	45.6	40.5	11.2	24.9	33.5	63.8	87.1
San Nicolás - Villa Constitución	26	41.1	37.2	9.5	32.4	43.6	54.5	82.1
Comodoro Rivadavia - Rada Tilly	27	41.4	40.0	3.3	24.0	31.7	71.9	94.7
Neuquen - Plottier	28	44.8	41.1	8.4	32.1	40.4	63.3	86.2
Río Gallegos	29	43.1	40.1	6.8	26.0	33.2	69.8	95.7
Ushuaia - Río Grande	30	44.4	38.7	12.8			80.9	98.4
Rawson - Trelew	31	48.0	45.4	5.4	25.2	32.0	74.8	92.3
Viedma - Carmen de Patagones	32	41.9	39.4	6.0	25.2	35.1	67.5	92.0

Algunos indicadores socioeconómicos en centros urbanos de Argentina. Año 2020

```
. pca X1-X7, cov

Principal components/covariance                 Number of obs    =         31
                                                Number of comp.  =          7
                                                Trace            =   198.7411
    Rotation: (unrotated = principal)           Rho              =     1.0000
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	126.058	86.5361	0.6343	0.6343
Comp2	39.5217	19.7814	0.1989	0.8331
Comp3	19.7404	11.4437	0.0993	0.9325
Comp4	8.2967	4.05079	0.0417	0.9742
Comp5	4.24591	3.3702	0.0214	0.9956
Comp6	.875705	.872834	0.0044	1.0000
Comp7	.0028702	.	0.0000	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Unexplained
X1	0.1611	0.0785	0.7063	0.1001	-0.1877	-0.0243	-0.6505	0
X2	0.1791	0.0913	0.4211	0.1853	-0.5202	0.0104	0.6908	0
X3	-0.0676	-0.0338	0.5382	-0.2140	0.7482	-0.0037	0.3148	0
X4	-0.4521	0.4271	0.0587	-0.1235	-0.0873	0.7658	-0.0191	0
X5	-0.5789	0.4976	0.0484	-0.0349	-0.0826	-0.6378	0.0116	0
X6	0.6022	0.6173	-0.1295	-0.4852	0.0366	-0.0529	-0.0003	0
X7	0.1878	0.4163	-0.1078	0.8112	0.3444	0.0571	0.0029	0

Algunos indicadores socioeconómicos en centros urbanos de Argentina. Año 2020

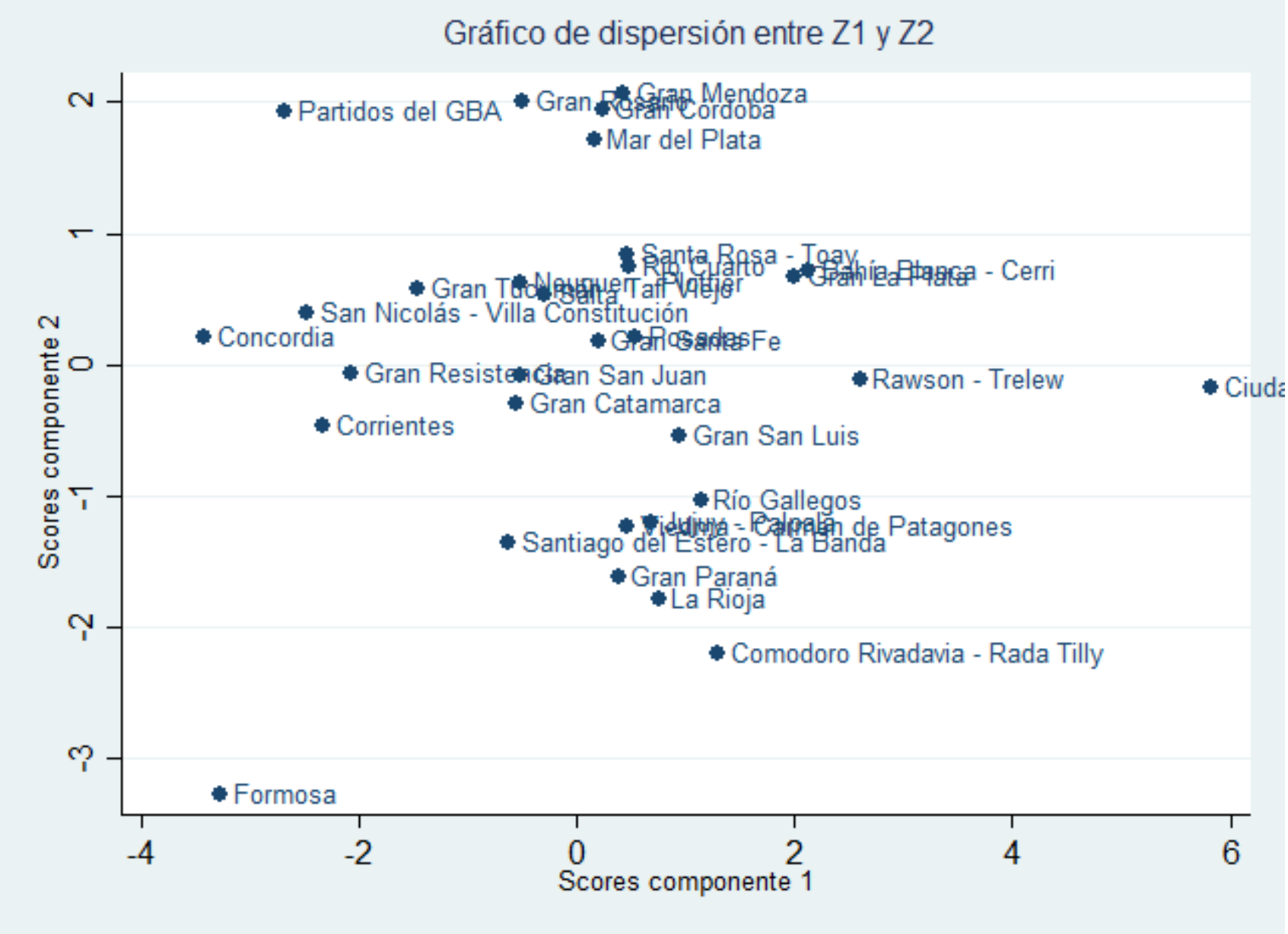
```
. pca X1-X7

Principal components/correlation                  Number of obs    =         31
                                                  Number of comp.  =          7
                                                  Trace            =          7
    Rotation: (unrotated = principal)             Rho              =     1.0000
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.48805	1.81883	0.4983	0.4983
Comp2	1.66922	.507693	0.2385	0.7368
Comp3	1.16153	.768751	0.1659	0.9027
Comp4	.392778	.125786	0.0561	0.9588
Comp5	.266991	.245821	0.0381	0.9969
Comp6	.0211705	.0209123	0.0030	1.0000
Comp7	.000258176	.	0.0000	1.0000

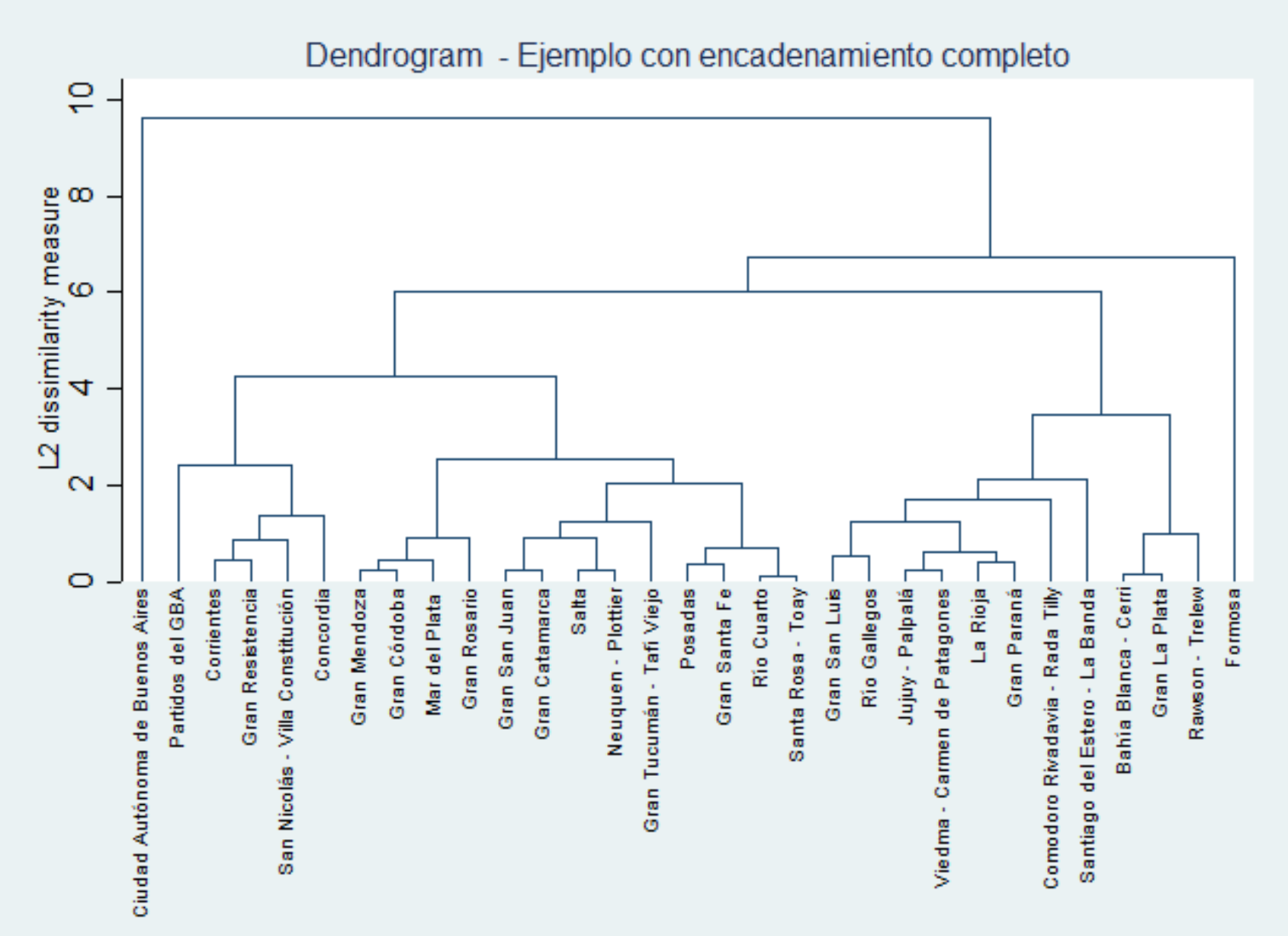
Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Unexplained
X1	0.3754	0.5402	0.0528	-0.2108	-0.0347	0.0350	-0.7194	0
X2	0.4516	0.3075	0.1273	-0.5338	-0.0181	-0.0220	0.6321	0
X3	-0.0679	0.6910	-0.1414	0.6452	-0.0217	-0.0077	0.2847	0
X4	-0.4237	0.2049	0.4844	-0.1438	0.2214	-0.6877	-0.0336	0
X5	-0.4348	0.1926	0.4693	-0.1482	0.1085	0.7205	0.0254	0
X6	0.4355	-0.1645	0.2858	0.3118	0.7744	0.0682	-0.0007	0
X7	0.3039	-0.1729	0.6515	0.3383	-0.5809	-0.0378	0.0036	0

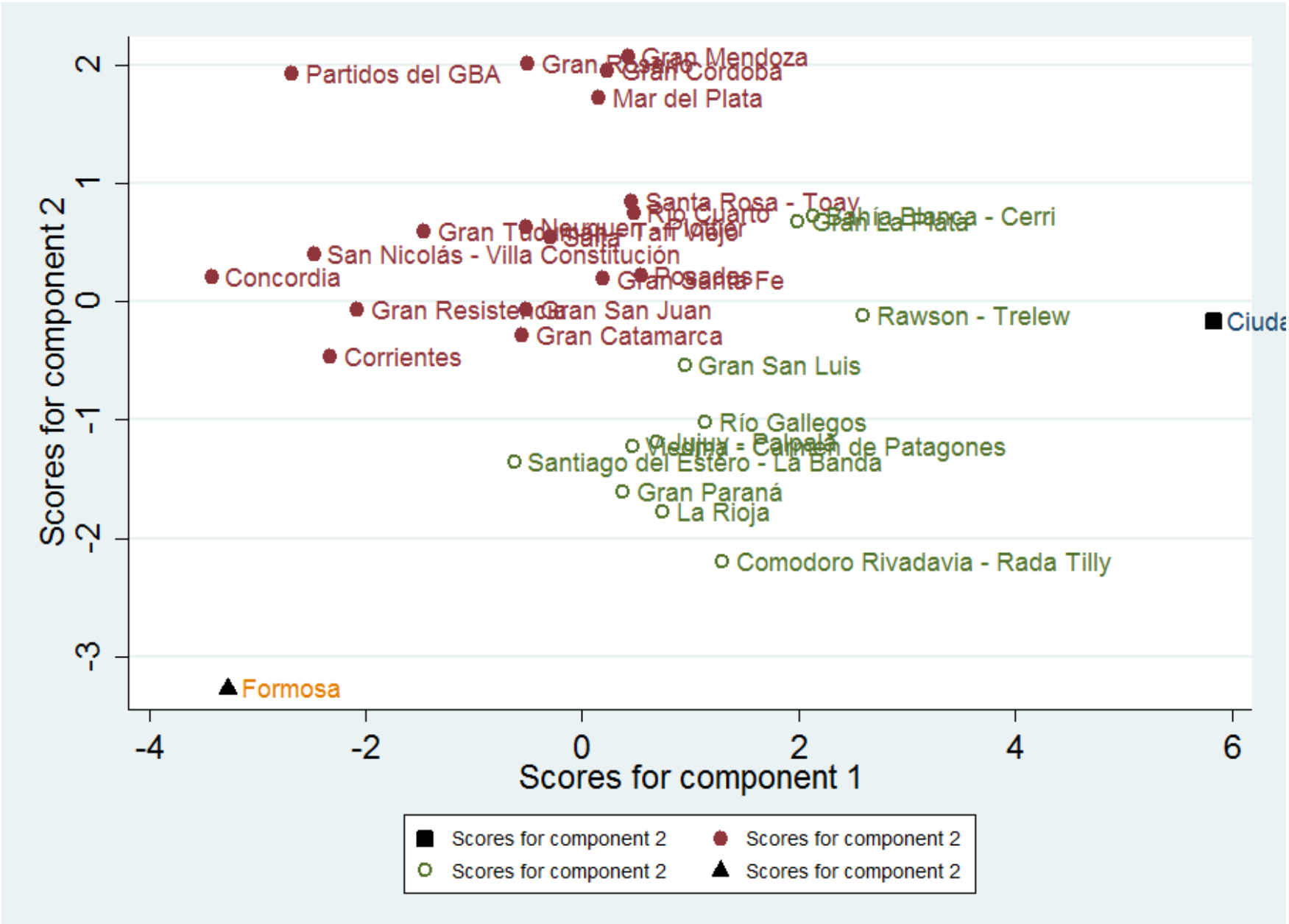
Algunos indicadores socioeconómicos en centros urbanos de Argentina. Año 2020



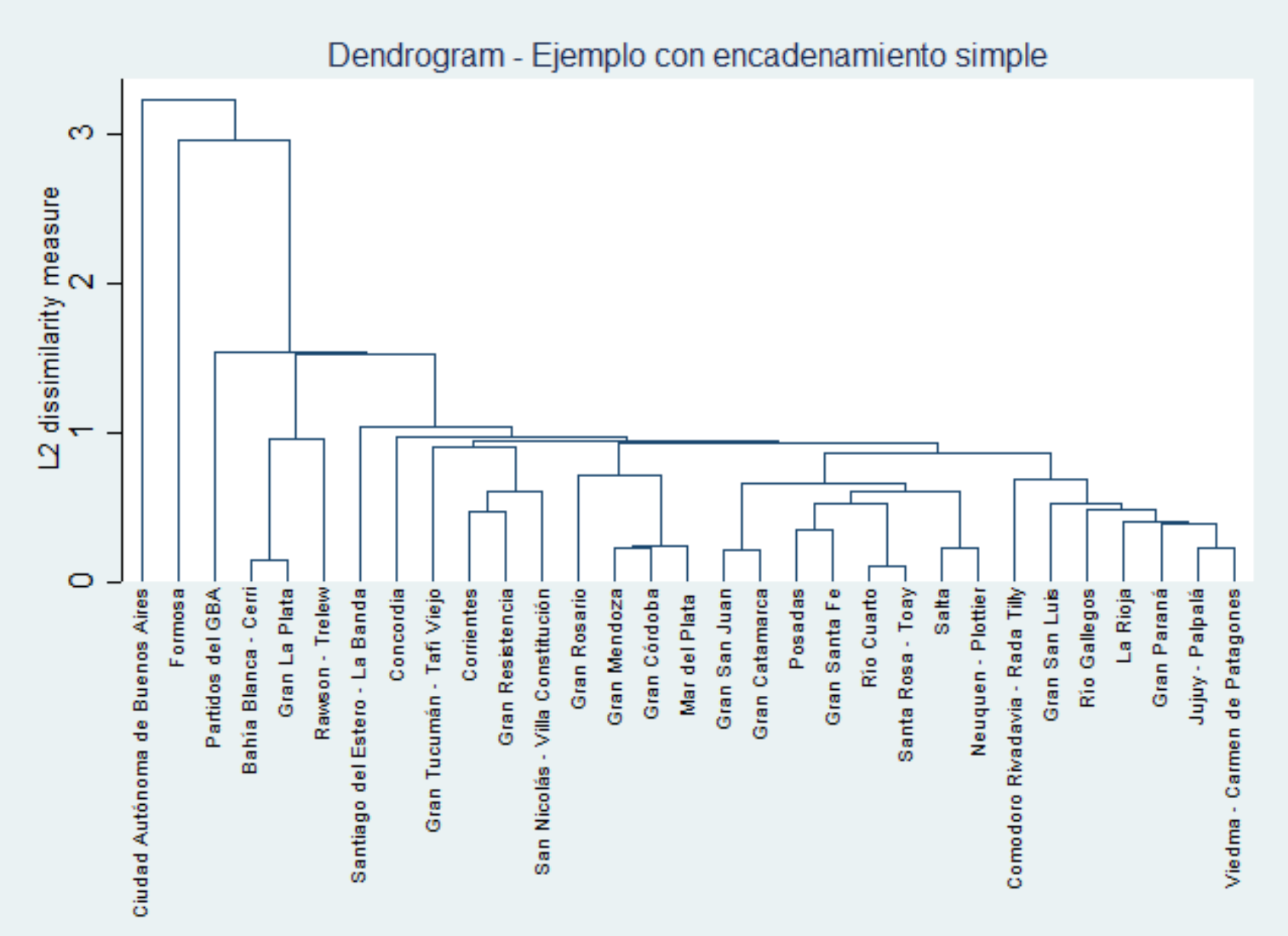
Algunos indicadores socioeconómicos en centros urbanos de Argentina. Año 2020



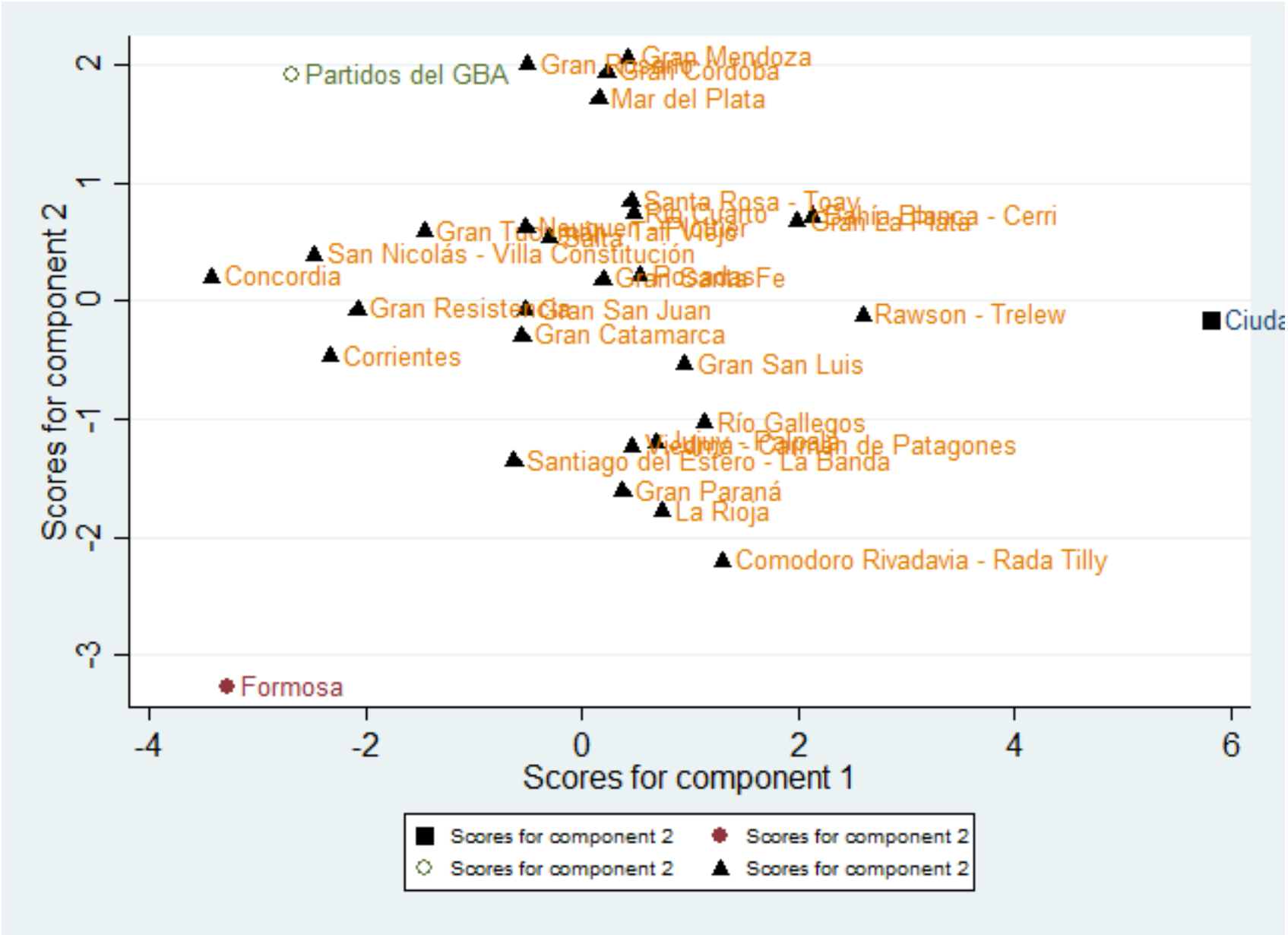
Algunos indicadores socioeconómicos en centros urbanos de Argentina. Año 2020



Algunos indicadores socioeconómicos en centros urbanos de Argentina. Año 2020



Algunos indicadores socioeconómicos en centros urbanos de Argentina. Año 2020



ANEXO - Medidas de similaridad

- Cuando en la muestra coexisten variables contínuas y atributos el problema se complica, ya que puede ocurrir que las variables contínuas tengan mayor importancia en el procedimiento de clasificación. Cuando esto no sea deseable la solución es trabajar con similaridades.
- El coeficiente de similaridad entre dos elementos i y h, en base a la variable j se define como una función s_{jih} no negativa y simétrica que satisface:

$$s_{jii} = 1, \qquad 0 \leq s_{jih} \leq 1, \qquad s_{jih} = s_{jhi}$$

Medidas de similaridad

- Si obtenemos las similaridades entre dos elementos para cada variable podemos combinarlas en un coeficiente de similaridad global entre los dos elementos.
- El coeficiente propuesto de Gower es:

$$s_{ih} = \frac{\sum_{j=1}^{p} w_{jih} s_{jih}}{\sum_{j=1}^{p} w_{jih}}$$

donde w_{jih} es una variable dummy que es igual a 1 si la comparación de estos dos elementos mediante la variable j tiene sentido y será igual a cero si no queremos incluir esa variable en la comparación.

Medidas de similaridad

- En el caso de los atributos, se pueden agrupar las variables binarias en grupos homogéneos para tratarlas conjuntamente.
- Si suponemos que todos los atributos tienen el mismo peso, podemos construir una medida de similaridad entre dos elementos A y B respecto a todos estos atributos contando el número de atributos que estan presentes:
  - En ambos elementos: (a)
  - En A pero no en B: (b)
  - En B pero no en A: (c)
  - En ninguno de los dos elementos: (d)
- Estas cuatro cantidades forman una tabla de asociacion entre elementos y serviran para construir medidas de similitud entre los dos elementos comparados.

Medidas de similaridad

Para calcular un coeficiente de similitud entre dos individuos a partir de su tabla de asociación se utilizan habitualmente:

- Proporción de coincidencias: se calcula como el número total de coincidencias sobre el número de atributos totales.

$$s_{ih} = \frac{a+d}{a+b+c+d}$$

- Proporción de apariciones: cuando la ausencia de un atributo no es relevante podemos excluir las ausencias y calcular la proporción de veces en que el atributo aparece en ambos elementos.

$$s_{ih} = \frac{a}{a+b+c}$$

Para una variable continua, se puede calcular:

$$s_{ih} = 1 - \frac{|x_{ij} - x_{hj}|}{max(x_j) - min(x_j)}$$

Este coeficiente se encuentra entre 0 y 1.

Medidas de similaridad

- Una vez obtenida la similaridad global entre elementos podemos transformar los coeficientes en distancias.
- La manera más simple es definir $d_{ih} = 1 - s_{ih}$ pero en algunos casos puede no satisfacer la desigualdad triangular.
- Si la matriz de similaridades es definida positiva, la distancia:

$$d_{ih} = \sqrt{2(1 - s_{ih})}$$

satisface la desigualdad triangular y por lo tanto satisface todas las características de una medida de distancia.