Análisis de parcelas con cabras en Teno

Enfocado en el análisis de componentes principales

Manuel Verdejo García

18/03/2025

Índice

1. **Antecedentes**: Estudio de las comunidades vegetales en zonas con presencia de cabras de la península de Teno. Mete la riqueza y la biomasa, variables que vamos a ignorar, para hacer una regresión.

2. **Objetivos:**
   2.1. Generar variables latentes (no observables) que estén correlacionadas entre sí y maximicen la varianza.
   2.2. Reducir la dimensionalidad de un conjunto de variables, obteniendo un número menor de variables dependientes. Se generan combinaciones lineales de las variables originales, de modo que se retienen aquellas que contienen la mayor cantidad de información:

$$CP_i = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{ki}X_k, \quad 1 \leq i \leq k$$

   2.3. Visualizar de manera gráfica e intuitiva la existencia de estructuras de agrupamiento (clusters) en una población.

3. **Variables:**
   - MO: % materia orgánica
   - Ca, Na, Mg y K (Iones) en meq/100 g y P en ppm
   - C.I.C: complejo de intercambio de cationes
   - pH en meq/100 g
   - C.E: intercambio de cationes en mS/cm
   - Sat: % saturación del suelo

4. **Modelo:**

Dado el tipo de variables consideradas, se emplea un modelo de análisis de componentes principales.

Hipótesis Asociadas: $x_{p\times1} = \mu_{p\times1} + \Lambda_{p\times k} f_{k\times1} + u_{p\times1}$

E(f) = 0, V(f) = I, E(u)=0. Cov (u_i,u_j) = cov(f,u) = 0

$$\boxed{\Sigma = \Lambda\Lambda^t + \Psi}$$

Para que el modelo sea válido, deben cumplirse los siguientes supuestos:

- Las variables deben ser cuantitativas.
- Las variables deben ser normales (aplicando TCL y Kolmogorov). Esta normalidad se va a usar en el test de esfericidad de Barlett.
- Correlación significativa entre variables. Para ello realizamos el test de esfericidad de Barlett, mientras menor sea la significancia mejor. Analizamos también el KMO, que es un estadístico que también estudia la correlación.

5. Estudio Piloto:

No procede en este caso, ya que se supone que se ha realizado un estudio previo.

6. Estudio de la Forma y Tamaño de la Muestra:

De igual manera, no aplica en este contexto.

7. Análisis de datos

7.1. Tipificación de las variables

ZPH	ZC.E	ZSAT	ZP	ZMO	ZCA	ZNa	ZMg	ZK	ZC.I.C
-,28815	,55912	-,56663	,30936	-,24108	-1,09968	-,12346	-,66749	,88637	-,87077
-,28815	-1,08599	,11183	,30936	1,31783	-,92563	-1,64392	-1,18963	,73475	-1,24962
-,02058	-,23297	-,37278	-,41093	,29479	-,33383	-,63028	,27732	,27991	,06878
-,55572	-,47669	-,66355	,30936	-,43594	-1,37817	-1,39051	-,51831	1,49284	-,85562
,51456	,37633	-,17894	-,41093	,78195	-,43827	,38336	,20273	-1,23625	-,29492
,24699	-1,57344	-,17894	-,05079	1,07425	-,43827	-,37687	-,34426	-1,38787	-,62831
1,58484	-1,20786	-1,14817	-,77108	-1,21540	-1,65667	-1,39051	-1,71176	-,62979	-2,31040
1,85241	,86377	,20876	-,77108	-,28980	1,37193	,38336	,65028	1,34122	,81132
1,31727	-1,14693	-1,05125	-,77108	-,87439	1,12825	-,12346	-,12049	,88637	,28093
,24699	1,35121	-,56663	,66950	,29479	,84976	,89019	,02869	-,93302	,37186
-,82330	-,90320	1,66261	,30936	1,22039	-,78638	-,88369	-,29454	,73475	-,40100
-,02058	-,41576	1,66261	,30936	,19736	-,22940	-,88369	-,29454	,73475	-,27977
-,28815	-1,32972	,40260	-,41093	-,04622	-,43827	-1,39051	,97350	-,47817	,41732
-,82330	-,78134	,40260	-,23086	-,04622	-,82119	-,63028	-,04590	-,78141	-,40100
,51456	-,35483	,30568	-,41093	-1,21540	,71051	,12995	,84919	-1,23625	,69009
-1,09087	-,90320	3,01954	-,41093	2,14600	,25796	-,63028	,60055	-,93302	,62947
2,65513	-,35483	,88722	-,05079	-1,06925	,67570	-,12346	-,46858	1,64445	-,32523
,51456	1,35121	,59645	-,05079	-1,11797	1,09344	,38336	1,64482	-,47817	1,47810
,78213	-,47669	,88722	-,23086	-,48466	1,26749	-,63028	,10328	2,40253	,73555

Las variables han sido tipificadas, es decir, normalizadas a una escala común, con el propósito de evitar que las componentes principales se vean influenciadas por las diferencias de escala entre las variables. Sin esta tipificación, las variables con mayores magnitudes numéricas podrían dominar el análisis, distorsionando la interpretación de las componentes principales.

7.2. Verificación de las hipótesis del Modelo

7.2.1. Cuantitativas:

Ph, P,…, lo son todas salvo la parcela, luego esta última no debe entrar en el análisis, además no entra en el estudio ARCILLA, LIMO, ARENA SPC_RICH.

7.2.2. Normalidad:

Prueba de Kolmogorov-Smirnov para una muestra

		Puntuación Z (PH)	Puntuación Z (C.E)	Puntuación Z (SAT)	Puntuación Z (P)	Puntuación Z (MO)	Puntuación Z (CA)	Puntuación Z (Na)	Puntuación Z (Mg)	Puntuación Z (K)	Puntuación Z (C.I.C)
		39	39	39	39	39	39	39	39	39	39
Media		,0000000	,0000000	,0000000	,0000000	,0000000	,0000000	,0000000	,0000000	,0000000	,0000000
Desv. estándar		1,00000000	1,00000000	1,00000000	1,00000000	1,00000000	1,00000000	1,00000000	1,00000000	1,00000000	1,00000000
Absoluta		,105	,126	,087	,161	,140	,156	,139	,086	,120	,086
Positivo		,101	,126	,087	,161	,140	,156	,139	,086	,120	,086
Negativo		-,105	-,083	-,074	-,129	-,086	-,094	-,085	-,066	-,102	-,066
		,105	,126	,087	,161	,140	,156	,139	,086	,120	,086
		,200[d]	,122	,200[d]	,012	,052	,018	,056	,200[d]	,171	,200[d]
Sig.		,344	,119	,634	,011	,050	,017	,055	,649	,166	,659
Intervalo de confianza al 99%	Límite inferior	,331	,110	,622	,008	,044	,014	,049	,637	,156	,647
	Límite superior	,356	,127	,647	,014	,056	,021	,060	,661	,175	,671

Podemos observar en la tabla que no hay ninguna variable con significación menor que 0.01 por lo que no podemos rechazar la normalidad de ninguna de ellas, sin embargo sí que hay varias con significación menor que 0.05. Si no funcionase la normalidad deberíamos recurrir a estadística no paramétrica.

7.2.3. Correlación:

Prueba de KMO y Bartlett

Medida Kaiser-Meyer-Olkin de adecuación de muestreo		,279
Prueba de esfericidad de Bartlett	Aprox. Chi-cuadrado	350,379
	gl	45
	Sig.	<,001

Como podemos observar, tras hacer el test de esfericidad de Bartlett el nivel de significancia es menor a 0.01 por lo que no podemos rechazar la correlación entre variables. En caso de rechazarla podríamos seguir haciendo el análisis que corresponde solo a los autovalores.

7.3. Reducir el número de componentes principales:

7.3.1. Proporción de varianza explicada:

Varianza total explicada

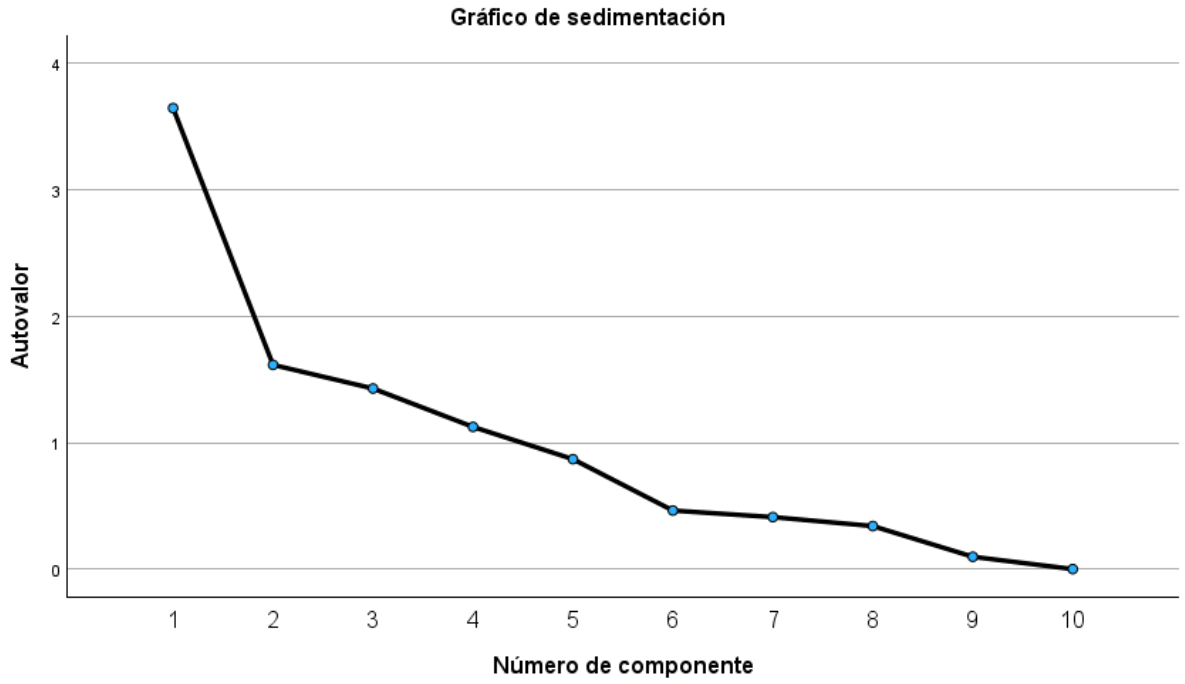
	Autovalores iniciales			Sumas de cargas al cuadrado de la extracción		
Componente	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	3,646	36,465	36,465	3,646	36,465	36,465
2	1,616	16,157	52,621	1,616	16,157	52,621
3	1,429	14,285	66,907	1,429	14,285	66,907
4	1,125	11,252	78,158	1,125	11,252	78,158
5	,870	8,699	86,857			
6	,463	4,635	91,492			
7	,412	4,116	95,608			
8	,341	3,408	99,016			
9	,098	,978	99,994			
10	,001	,006	100,000			

Método de extracción: análisis de componentes principales.

Obtenemos una tabla con tantas componentes principales como variables originales, la columna del total son los autovalores, es decir, la varianza de cada componente principal. El

porcentaje acumulado nos muestra cuánto explica cada componente del conjunto total y nos ayuda a determinar el punto en el que debemos realizar el corte. En este caso, el corte ocurre alrededor del 95%, lo que corresponde a la séptima componente. Al reducir las variables de 10 a 7, hemos logrado disminuir la dimensionalidad, lo cual es positivo. Sin embargo, esta reducción también implica una desventaja: no podremos representar los datos de manera óptima, ya que las variables restantes pueden no ser suficientes para separar de forma clara los clústeres, aún así clústeres diferentes pueden quedar cerca entre sí.

7.3.2. Test ladera:



Esta gráfica nos indica con cuántas componentes debemos quedarnos. Como está ordenado de mayor a menor nos quedamos cuando se empieza a suavizar la pendiente, es decir, la 6 o 7.

7.3.3. Criterio de Kaiser:

Consiste en quedarse con las CP con autovalores mayores que uno, que en este caso serían las 4 primeras. Este criterio tiene mucha menos validez que los anteriores y es generalmente utilizado como una herramienta complementaria, más que como un criterio principal, y se recomienda su uso solo para confirmar o respaldar sospechas previas acerca de la cantidad de componentes relevantes, pero no como la única base para tomar decisiones.

7.4. Interpretación de las CP:

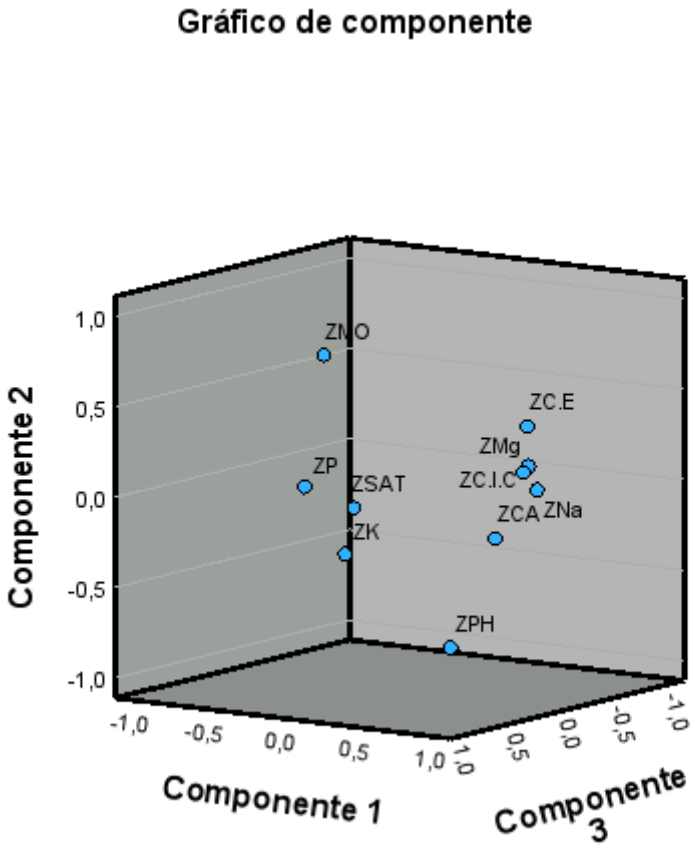
7.4.1. Coeficientes de carga o loadings:

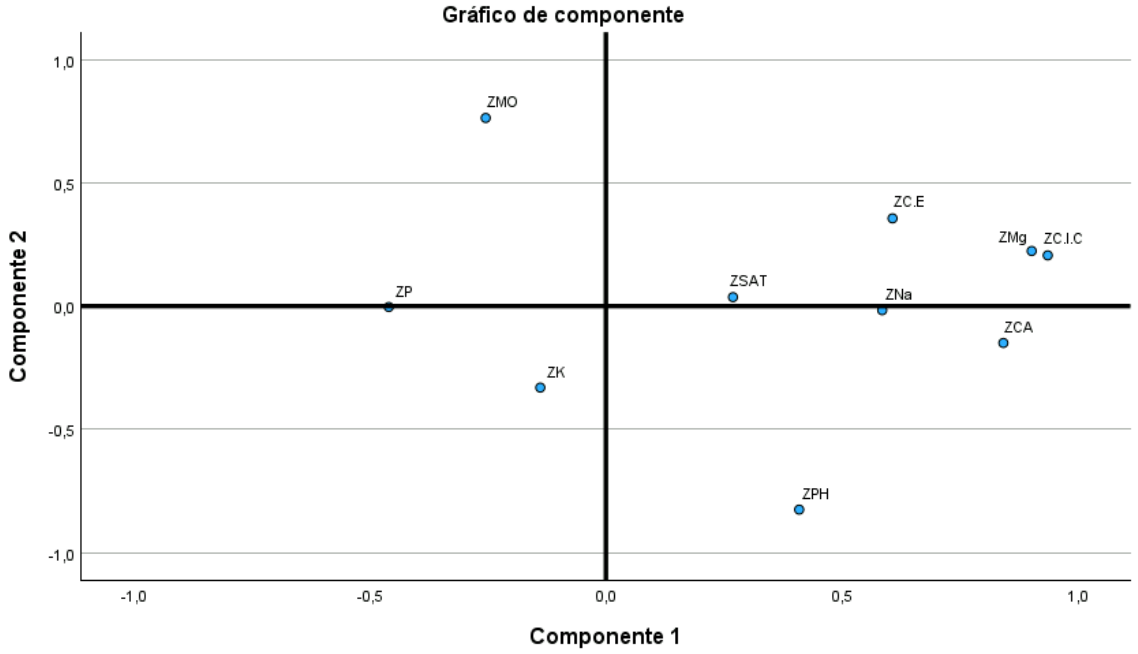
Matriz de componente[a]

	Componente						
	1	2	3	4	5	6	7
Puntuación Z(PH)	,409	-,825	,103	,026	-,125	,067	-,022
Puntuación Z(SAT)	,269	,036	,821	-,256	-,127	,031	,410
Puntuación Z(C.E)	,607	,356	-,342	,290	,256	-,311	,269
Puntuación Z(P)	-,460	-,004	,246	-,549	,601	-,218	-,086
Puntuación Z(MO)	-,255	,763	,356	,224	,040	,305	-,120
Puntuación Z(CA)	,842	-,149	,296	-,014	,253	,057	-,270
Puntuación Z(Na)	,585	-,017	-,465	-,281	,348	,441	,196
Puntuación Z(Mg)	,902	,224	,066	-,051	-,216	-,142	-,063
Puntuación Z(K)	-,139	-,330	,322	,736	,418	,036	,078
Puntuación Z(C.I.C)	,936	,206	,167	,012	,056	-,046	-,166

Para ayudar a la interpretación es útil calcular los coeficientes de correlación entre CP y las variables, llamados coeficientes de carga o loadings. Aquellas variables que presenten las mayores correlaciones con las respectivas componentes principales son las que más explican a éstas. Para ayudar a interpretar cada variable latente ignoramos algunas de las últimas columnas para eliminar componentes y señalamos los valores más altos y esos son los que mayormente explica la CP. En caso de que todos sean parecidos rotamos.

7.4.2. Plots of loadings o Gráfico de Saturaciones:





Es la representación de las variables originales sobre las componentes principales. Podemos observar que las variables MG, CA y C.I.C son las que más explican a la CP1 y PH y MO son las que más explican a la CP2.

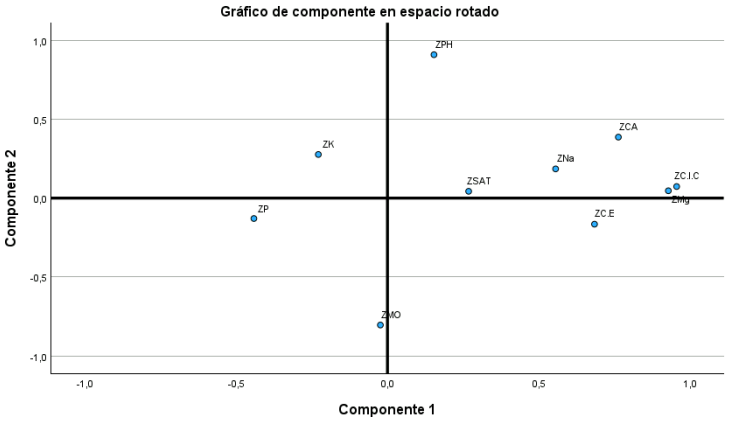
7.4.3. Rotación de las CP:

Hay varios tipos de rotación, varimax, quartimax... Probamos con todas y vemos cual da mejores resultados.

Varimax:

Matriz de componente rotado[a]

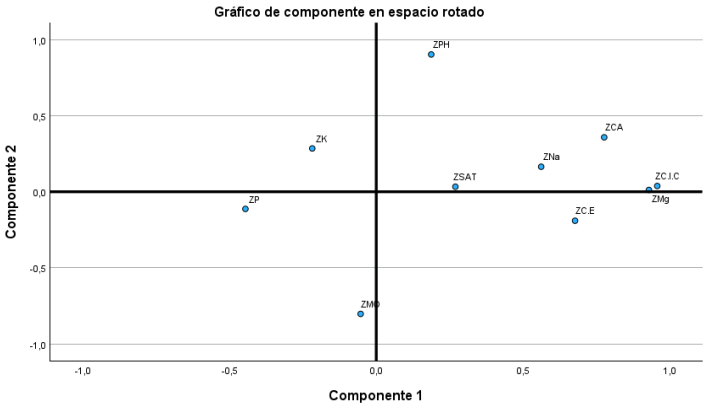
	Componente 1	Componente 2
Puntuación Z(PH)	,154	,908
Puntuación Z(SAT)	,268	,043
Puntuación Z(C.E)	,684	-,165
Puntuación Z(P)	-,441	-,129
Puntuación Z(MO)	-,023	-,804
Puntuación Z(CA)	,763	,386
Puntuación Z(Na)	,556	,185
Puntuación Z(Mg)	,928	,046
Puntuación Z(K)	-,228	,276
Puntuación Z(C.I.C)	,956	,073



Quartimax:

Matriz de componente rotado[a]

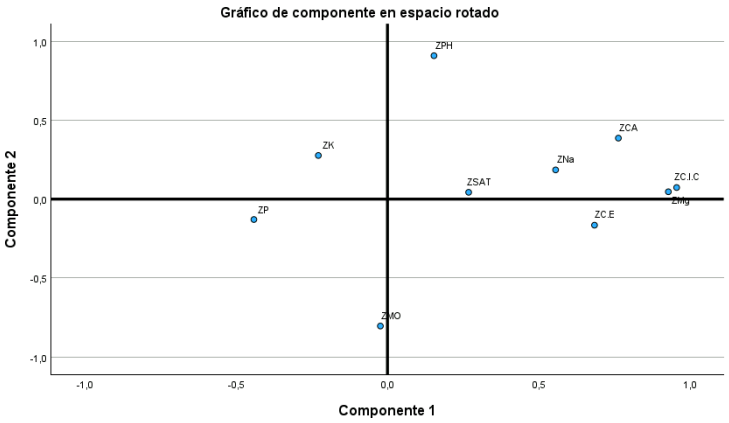
	Componente 1	2
Puntuación Z(PH)	,187	,902
Puntuación Z(SAT)	,270	,033
Puntuación Z(C.E)	,678	-,191
Puntuación Z(P)	-,446	-,113
Puntuación Z(MO)	-,053	-,803
Puntuación Z(CA)	,777	,358
Puntuación Z(Na)	,562	,164
Puntuación Z(Mg)	,929	,012
Puntuación Z(K)	-,218	,285
Puntuación Z(C.I.C)	,958	,038



Equamax:

Matriz de componente rotado[a]

	Componente 1	2
Puntuación Z(PH)	,154	,908
Puntuación Z(SAT)	,268	,043
Puntuación Z(C.E)	,684	-,165
Puntuación Z(P)	-,441	-,129
Puntuación Z(MO)	-,023	-,804
Puntuación Z(CA)	,763	,386
Puntuación Z(Na)	,556	,185
Puntuación Z(Mg)	,928	,046
Puntuación Z(K)	-,228	,276
Puntuación Z(C.I.C)	,956	,073

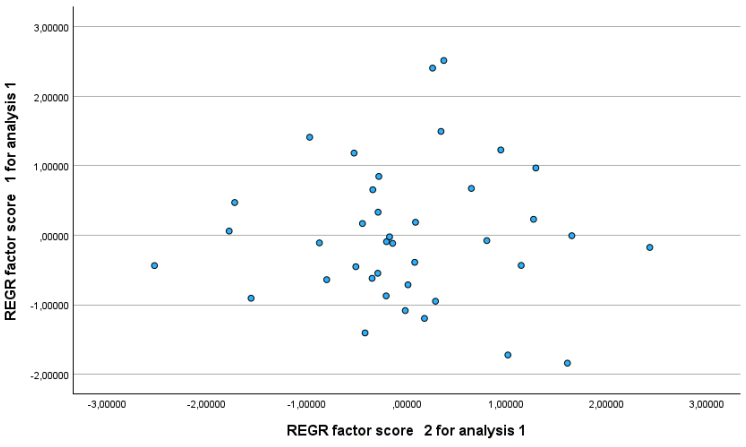


Como podemos observar todas las rotaciones dan resultados muy similares (e iguales en el caso de varimax y equamax). Aún así, podemos observar que las variables MG y C.I.C son las que más explican a la CP1 y PH y MO son las que más explican a la CP2.

7.4.4. Análisis cluster de las CP:

Matriz de coeficiente de puntuación de componente

	Componente 1	2
Puntuación Z(PH)	,020	,523
Puntuación Z(C.E)	,203	-,187
Puntuación Z(SAT)	,077	-,009
Puntuación Z(P)	-,125	-,020
Puntuación Z(MO)	,015	-,477
Puntuación Z(CA)	,211	,132
Puntuación Z(Na)	,156	,039
Puntuación Z(Mg)	,268	-,092
Puntuación Z(K)	-,074	,194
Puntuación Z(C.I.C)	,275	-,080



En la tabla obtenemos los coeficientes de cada variable tipificada en la combinación lineal que genera las dos primeras variables principales. Podemos observar que en la primera no hay ninguno que destaque especialmente y en la segunda destacan el PH y el MO. En el gráfico de dispersión se representan las muestras tomadas respecto de ambas PC y podemos observar una falta de varianza que nos impide visualizar claramente clústeres.