Licenciatura en Ciencia Ambientales

Matemáticas y Estadística aplicada

Prof. Esperanza Ayuga Téllez

1.- Determinar si en los datos representados por los diagramas siguientes se detecta algún atípico.

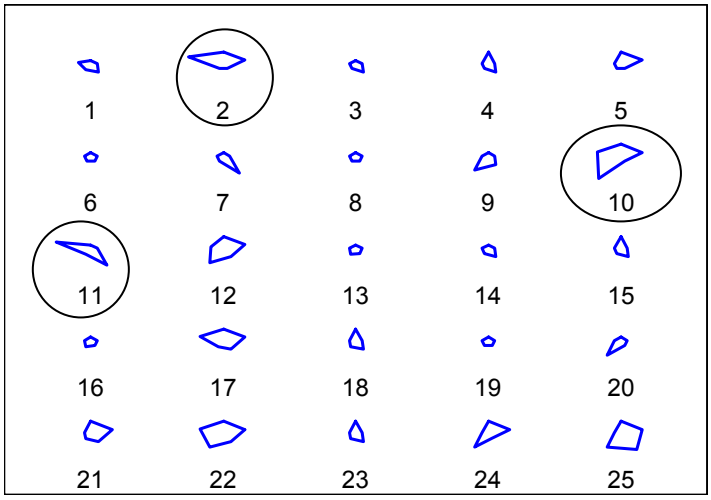


Diagrama de estrella para 25 canteras de Huelva en que se midieron porcentajes de cobertura vegetal en diferentes partes.

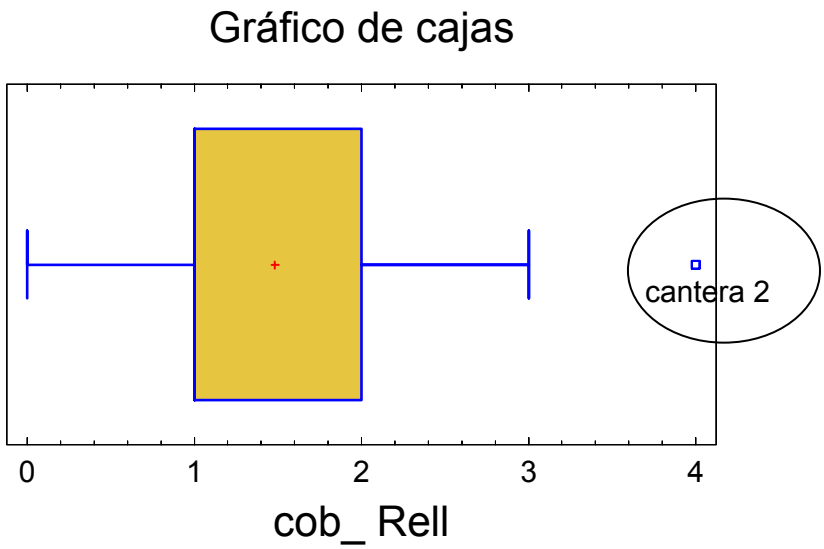


Gráfico de cajas para la variable porcentaje de cobertura vegetal en zonas de rellenado de la cantera

Solución:

Ejercicios

1.- Las canteras que parecen tener el conjunto de variables diferente del resto son la 2, la 10 y la 11. Son posibles datos atípicos.

2.- Calcular para el conjunto de datos mostrado, la matriz de medias, varianzas-covarianzas y la distancia de Mahalanobis del punto 3 al vector de medias.

Cantera	Cobertura ent	Cob-derrub	Cob-rell
1	3	2	2
2	1	1	4
3	3	1	1

Solución:

Cantera	Cob-ent = x	Cob-derrb= y	Cob-rell= z	x^2	y^2	z^2	xy	xz	yz
1	3	2	2	9	4	4	6	6	4
2	1	1	4	1	1	16	1	4	4
3	3	1	1	9	1	1	3	3	1
total	**7**	**4**	**7**	**19**	**6**	**21**	**10**	**13**	**9**

$$\overline{\mathbf{X}} = \frac{1}{3}\begin{bmatrix}7\\4\\7\end{bmatrix} = \begin{bmatrix}2{,}33\\1{,}33\\2{,}33\end{bmatrix}; \qquad \mathbf{S} = \begin{bmatrix}0{,}89 & 0{,}22 & 1{,}22\\0{,}22 & 0{,}22 & -0{,}11\\1{,}22 & -0{,}11 & 1{,}56\end{bmatrix} \qquad \text{Det } \mathbf{S} = -0{,}17$$

La distancia de Mahalanobis del la cantera 3 al vector de medias se calcula con:

$$d_3 = \left(\left|\begin{bmatrix}0{,}67 & -0{,}33 & -1{,}33\end{bmatrix}\mathbf{S}^{-1}\begin{bmatrix}0{,}67\\-0{,}33\\-1{,}33\end{bmatrix}\right|\right)^{1/2} = \left(\left|\begin{bmatrix}0{,}67 & -0{,}33 & -1{,}33\end{bmatrix}\begin{bmatrix}-1{,}9 & 2{,}8 & 1{,}7\\2{,}8 & 0{,}59 & -2{,}1\\1{,}7 & -2{,}1 & -0{,}88\end{bmatrix}\begin{bmatrix}0{,}67\\-0{,}33\\-1{,}33\end{bmatrix}\right|\right)^{1/2}$$

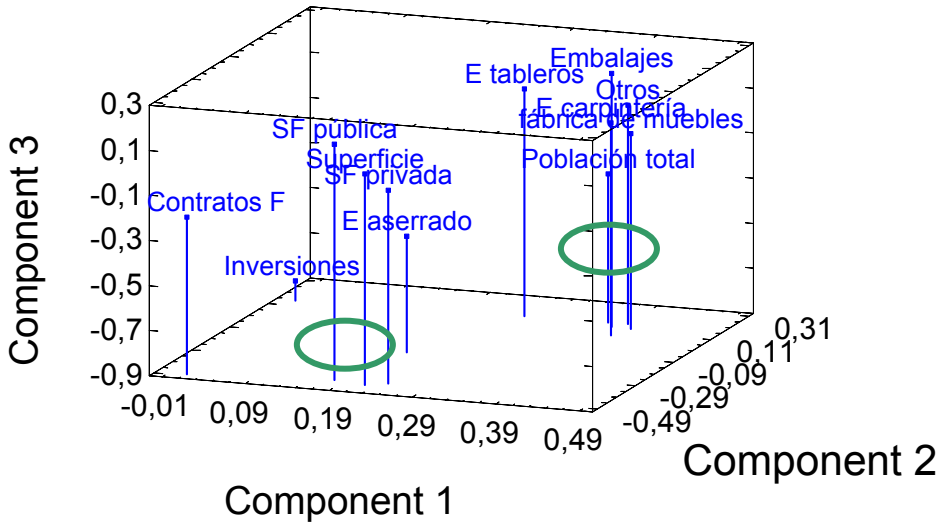
$$d_3 = \left(\left|\begin{bmatrix}-4{,}5 & 4{,}6 & 3\end{bmatrix}\begin{bmatrix}0{,}67\\-0{,}33\\-1{,}33\end{bmatrix}\right|\right)^{1/2} = (8{,}6)^{1/2} = 2{,}93$$

Trabajo cooperativo

1.- En el estudio de las características de empleo en el sector forestal se han observado, para las 17 comunidades autónomas, un total de 12 variables: Población de la comunidad autónoma, Superficie de la CA, superficie forestal de propiedad privada y la de propiedad pública (las tres en hectáreas), el porcentaje de contratos e inversiones por hectárea del sector sobre el total de la comunidad y además el número de empleos en la industria de aserrado, carpintería, embalajes, fábrica de muebles y tableros.

Se consideran en el análisis, tres componentes principales que explican el 82,62% de la variabilidad total.

Discutir y explicar las relaciones de las variables entre sí y con cada una de las componentes.



SOLUCIÓN:

La Superficie de cada Comunidad Autónoma, junto con la superficie pública y privada tienen comportamientos similares en cuanto a las tres componentes. El empleo en la industria de aserrado es muy similar a las anteriores en relación a las dos primeras componentes, pero inferior en cuanto a la componente 3 (C3).

Las Inversiones tienen un valor muy diferente al resto en la C3, semejante a los Contratos en la componente 1 (C1) y a la población de la comunidad autónoma, número de empleos en la industria de carpintería, embalajes, fábrica de muebles, tableros y otros en la componente 2 (C2).

La población de la comunidad autónoma, número de empleos en la industria de carpintería, embalajes, fábrica de muebles, y otros son muy semejantes en las dos primeras componentes, siendo algo más diferentes respecto a la C3, siendo las variables menos semejantes la población y el empleo en embalajes.

La Superficie forestal de cada Comunidad Autónoma, junto con la superficie pública y privada, el empleo en la industria de aserrado y de tableros presentan valores muy similares de la C1, siendo éste último distinto del resto en la C2.

2.- Si la tabla de pesos de las componentes es la siguiente:

	Componente 1	Componente 2	Componente 3
Superficie	0,219324	-0,453028	0,0382444
SF pública	0,178156	-0,433496	0,146876
SF privada	0,241187	-0,427964	-0,0417792
Población total	0,367764	0,173313	-0,241123
Inversiones	-0,00184632	0,150816	-0,811412
Contratos F	0,012363	-0,481973	-0,201413
E aserrado	0,206315	-0,175499	-0,384792
E tableros	0,268341	0,171292	0,109518
E carpintería	0,391927	0,0814292	0,040674
Embalajes	0,378394	0,146876	0,223347
fábrica de muebles	0,402208	0,141752	-0,034498
Otros	0,391723	0,173709	0,073039

Obtener las ecuaciones de cada componente e interpretarlas.

SOLUCIÓN:

Componente 1 = 0,219324 Superficie +0,178156 SF pública +0,241187 SF privada+ +0,367764 Población total -0,00184632 Inversiones +0,012363 Contratos F +0,206315 E aserrado +0,268341 E tableros +0,391927 E carpintería +0,378394 Embalajes + +0,402208 fábrica de muebles +0,391723 Otros

Componente 2 = -0,453028 Superficie -0,433496 SF pública -0,427964 SF privada +

+0,173313 Población total +0,150816 Inversiones -0,481973 Contratos F -0,175499 E aserrado +0,171292 E tableros +0,0814292 E carpintería +0,146876 Embalajes +

+0,141752 fábrica de muebles+ 0,173709 Otros

Componente 3 = 0,0382444 Superficie +0,146876 SF pública -0,0417792 SF privada –

-0,241123 Población total -0,811412 Inversiones -0,201413 Contratos F -0,384792 E aserrado +0,109518 E tableros +0,040674 E carpintería +0,223347 Embalajes –

-0,034498 fábrica de muebles +0,073039 Otros

En cada componente se resalta en rojo las variables de mayor peso.

La C1 es una suma de la mayoría de variables relacionadas con el empleo, el tamaño de la superficie forestal y las dimensiones de la Comunidad con pesos semejantes. Apenas tienen peso en ella ni las inversiones ni el número de contratos.

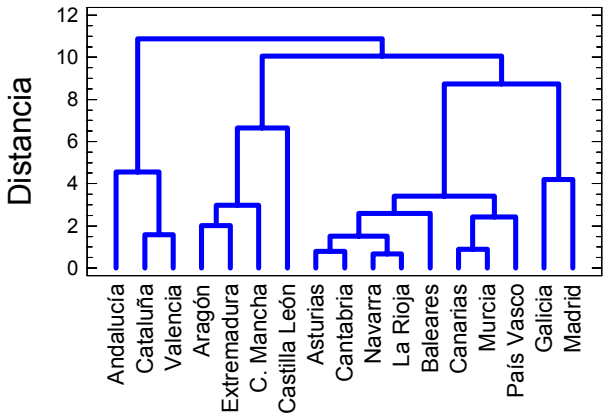
En la C2 influyen fundamentalmente las superficies y los contratos con signo negativo.

En la C3 el peso principal corresponde a las inversiones con signo negativo.

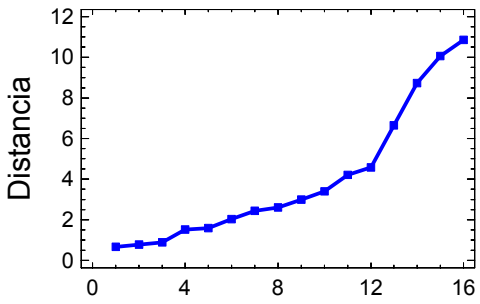
Trabajo cooperativo

1.- En un estudio sobre el sector forestal se han obtenido datos de tres variables que representan el tamaño del sector productivo (P), la población empleada en el sector (E) y las inversiones realizadas en el sector (I) y se trata de agrupar las comunidades autónomas españolas en grupos homogéneos para las variables consideradas.

Se obtienen el dendrograma y el gráfico de aglomeración de los conglomerados con los dos métodos siguientes:

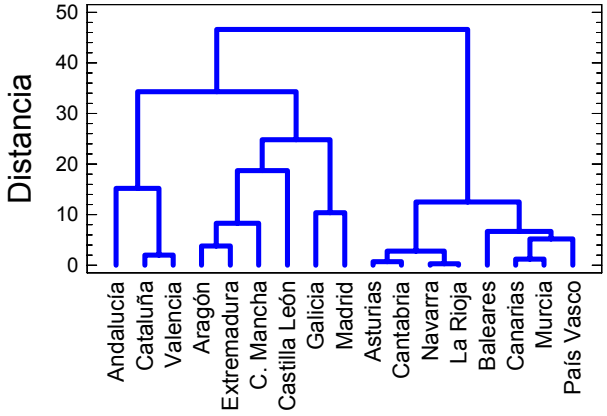


a) Dendrograma

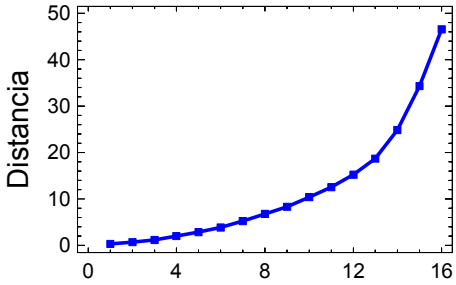


b) Gráfico de aglomeración

Método de el vecino más lejano y distancia ciudad-bloque.



a) Dendrograma



b) Gráfico de aglomeración

Método de Ward y distancia ciudad-bloque.

Discutir y explicar qué método escogerías de los dos, cuántos grupos se escogerían y qué comunidades se incluyen en cada grupo.

SOLUCIÓN:

Con el método del vecino más lejano o distancia máxima, agrupamos en el paso 12 con una distancia ciudad-bloque de 4,5. Se obtienen seis grupos:

Grupo1: Andalucía
Grupo 2: Cataluña y Valencia
Grupo 3: Aragón, Extremadura y Castilla-La Mancha.
Grupo 4: Castilla-León.
Grupo 5: Asturias, Cantabria, Navarra, La Rioja, Baleares, Canarias, Murcia y País Vasco.
Grupo 6: Galicia y Madrid.

Con el método de Ward, agrupamos en el paso 12 con una distancia ciudad-bloque de 20. Se obtienen cuatro grupos:

Grupo1: Andalucía, Cataluña y Valencia
Grupo 2: Aragón, Extremadura, Castilla-La Mancha y Castilla-León.
Grupo 3: Asturias, Cantabria, Navarra, La Rioja, Baleares, Canarias, Murcia y País Vasco.
Grupo 4: Galicia y Madrid.

Los dos métodos consiguen agrupamientos similares. En el caso del vecino más lejano se separan del grupo 1 Andalucía y del 2 Castilla-León.

Sin información adicional, cualquiera de las dos agrupaciones es válida e informativa.

Parece que, desde el punto de vista del empleo, el peso del sector productivo forestal y las inversiones las comunidades se pueden considerar bien agrupadas con el método de Ward. Cuando no hay más información, este método es más discriminativo en la determinación de los grupos y conduce a grupos más homogéneos.