Técnicas de Análisis Cuantitativas y Cualitativas
Resolución del ejercicio de evaluación 3

Marcos Rial Docampo

16 de enero de 2016

En este ejercicio de evaluación se nos presentan datos de superficie agrícola abandonada (porcentaje de la superficie agrícola al inicio del periodo), densidad de población (hab/km^2) y altitud media (metros sobre el nivel del mar) de una serie de 50 observaciones tomadas en otros tantos municipios gallegos.

Lo primero que haremos será suprimir la variable categórica X puesto que no ofrece información alguna al tratarse de un código referente a la zona donde se tomaron las observaciones. Una vez hecho esto extraemos gráficamente la relación directa entre las tres variables restantes como se muestra en la figura 1. La estimación numérica de la correlación entre estas variables sería la mostrada en el cuadro 1 mediante el método de Pearson.

	Abandono	Densidad pob.	Elevación
Abandono	1		
Densidad pob.	-0,3620	1	
Elevación	0,8718	-0,5379	1

Cuadro 1: Matriz de correlaciones de las variables.

El siguiente paso previo a la generación de agrupamientos sería el de la estandarización de las variables. Tenemos tres variables cuyos datos no se encuentran entre los mismos intervalos donde, por ejemplo, la variable abandono *abandon.uaa* muestra valores entre 0 y 0,73 mientras que la densidad de población *pop.dens* por ejemplo está entre 9 y 728. Lo que se busca con este estandarizado es el de aproximar los intervalos de los valores de las tres variables y realizar un agrupamiento lo más satisfactorio posible. El procedimiento seguido es el de restar a los valores de cada observación la media y dividir por la desviación típica (1). Una vez realizada esta estandarización se obtienen los datos del cuadro 2 donde a cada valor se le resta la media y divide por la desviación típica.

$$x_i = \frac{X_i - \mu}{\sigma} \tag{1}$$

donde x_i es el valor estandar de la observación i cuyo valor original es X_i, μ es la media de la observación y σ la desviación típica.

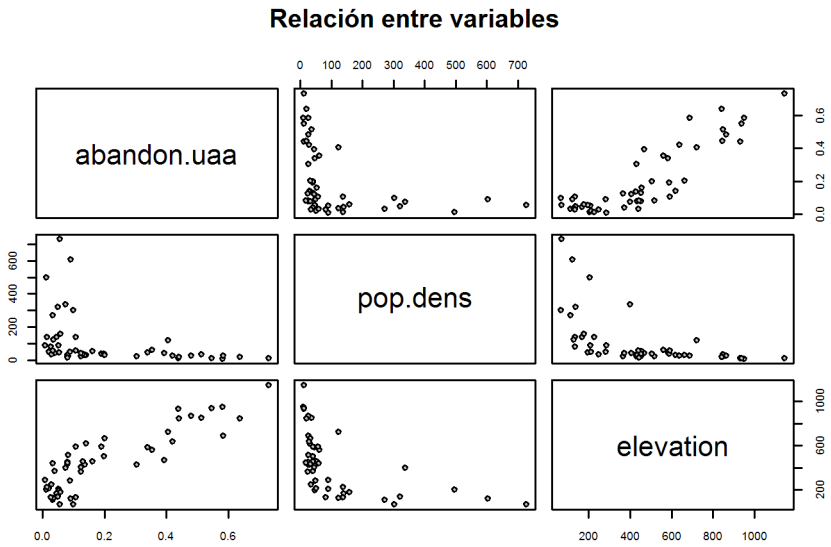

Figura 1: Relación entre variables.

	abandon.uaa		pop.dens		elevation	
	ant.	desp.	ant.	desp.	ant.	desp.
Mínimo	0,01	-0,97	9,38	-0,63	67,00	-1,42
1^{er} cuartil	0,05	-0,74	27,47	-0,52	209,80	-0,90
Mediana	0,11	-0,47	44,50	-0,40	437,00	-0,06
Media	0,20	0,00	106,22	0,00	454,10	0,00
3^{er} cuartil	0,35	0,75	114,55	0,05	612,00	0,58
Máximo	0,73	2,66	727,83	4,07	1145,00	2,54

Cuadro 2: Valores de las variables antes y después del estandarizado.

Pasamos a realizar el agrupamiento jerárquico. Para empezar calculamos la matriz de distancias entre cada observación en tantas dimensiones como variables tenemos, en este caso, tres. Para este ejercicio aplicaremos el cálculo de una distancia euclídea (cálculo habitual para la distancia entre dos puntos).

Para realizar el agrupamiento se aplica el algoritmo de Ward y obtenemos el dendrograma de la figura 2. En él podemos apreciar los agrupamientos en base a la distancia y el peso de la similitud entre grupos (cuanto mayor es este, menor es la similitud entre grupos). Decidimos realizar los grupos cortando el dendrograma por el peso 5 (línea roja a trazos de la figura 2) resultando así un total de 5 grupos.

El grupo 1 se caracteriza por una baja densidad de población en una zona de elevación media respecto del resto de grupos, al igual que el grupo 5 pero presentando menos porcentaje de abandono que este, en torno a un 50 % menos. El grupo 2 presenta niveles

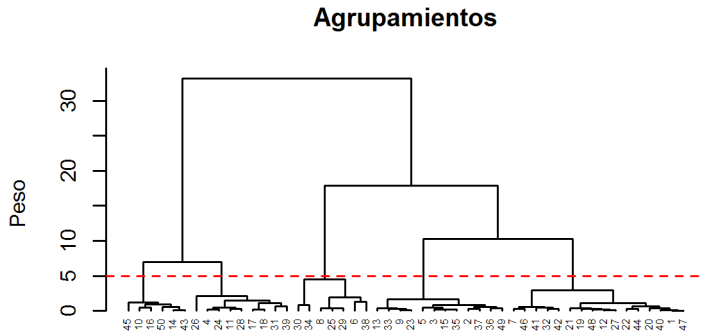

Figura 2: Dendrograma resultante del agrupamiento de observaciones.

bajos de las tres variables. En el grupo 3 claramente se agrupan observaciones hechas en municipios rurales de interior con mucha elevación, mucho abandono y una densidad de población baja. Al contrario que con el grupo 4 en el que su densidad alta y su elevación baja nos indica que se trata de observaciones de municipios costeros. Estos últimos son los grupos más claramente identificables de entre los cinco propuestos, dejando en duda a los otros tres.

Nos ayudamos de los gráficos de cajas de la figura 3 para analizar los datos de las observaciones por grupos. Extraemos la siguiente información:

- En la variable abandono los grupos 2 y 4 presentan mucha similaridad y podríamos decir que el grupo 1 se les parece levemente con una media próxima a la de estos otros grupos. Destaca el grupo 3 por ser el más diferente.
- En cuanto a la variable densidad de población los grupos 1, 3 y 5 se parecen notablemente. Se destaca la gran amplitud de los datos de las observaciones del grupo 4.
- En la variable elevación hay diferencias en todos los grupos. Los datos están notablemente bien agrupados.

Para realizar un agrupamiento no jerárquico aplicamos el algoritmo *k-means*. Utilizaremos, al aplicar dicho algoritmo, los datos escalados empleados en el agrupamiento jerárquico así como un total de 5 grupos. Adicionalmente, al comparar la coincidencia entre ambos métodos de clasificación resulta el cuadro 3 En él vemos que las dos agrupaciones son bastante similares (obviando el hecho de que los nombres de los grupos no coinciden) sobre todo en el caso del grupo 4 jerárquico y 3 no jerárquico donde solo se distinguen por una observación agrupada en otro grupo.

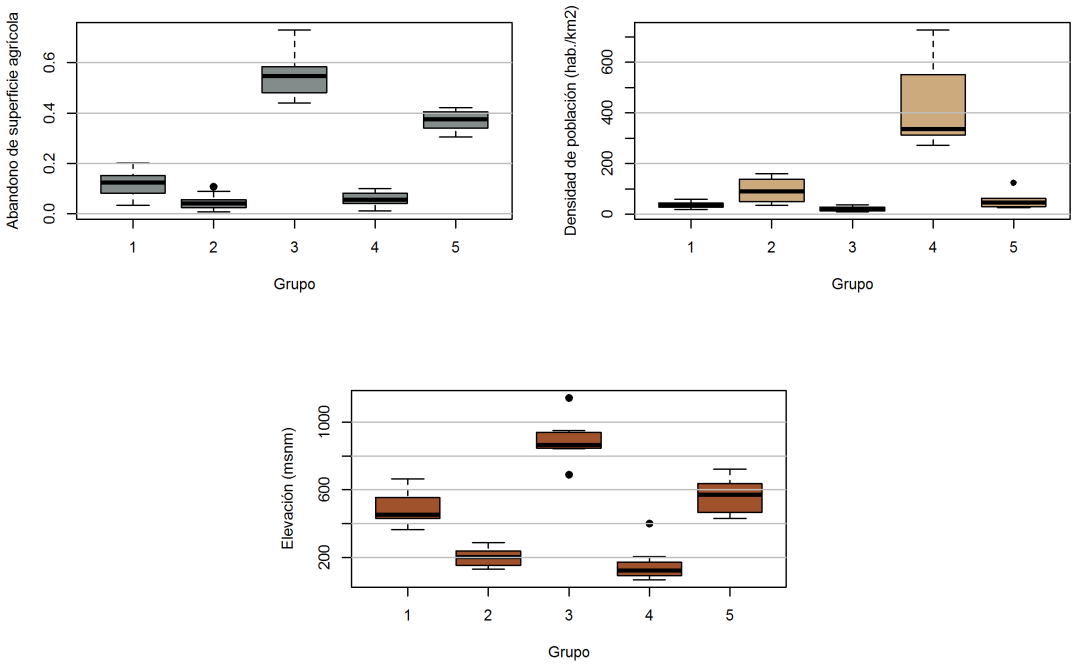

Figura 3: Diagramas de cajas de las observaciones agrupadas.

Grupos	1	2	3	4	5
1	0	1 0	0	15	
2	0	12	0	0	0
3	8	0	0	1	0
4	0	1	6	0	0
5	2	0	0	0	4

Cuadro 3: Tabla de comparación entre los dos agrupamientos

El algoritmo *k-means* crea un número k de centroides iniciales, tantos como grupos, originalmente repartidos de forma aleatoria entre las observaciones. Los grupos se generan asignando las observaciones al centroide cuyo valor se aproxime más al valor del dato. El valor del centroide se actualiza al de la media del grupo al que representa y vuelve a realizarse el paso anterior. El proceso es iterativo hasta conseguir convergencia o un número de iteraciones predeterminado.