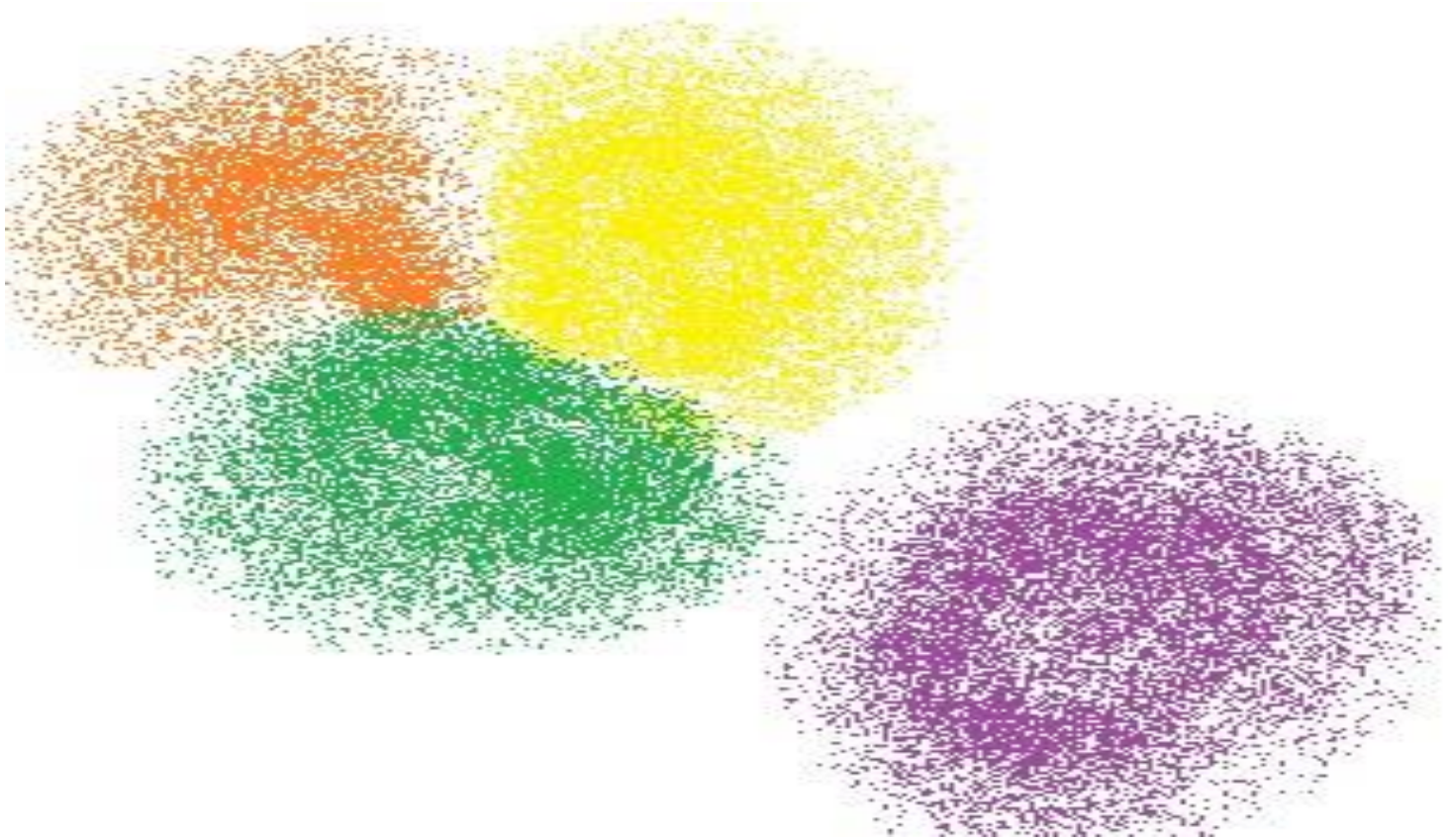
CLUSTERS



Cluster

Clusters es un método de clasificación no supervisado donde se hace un análisis exploratorio de un conjunto de datos no etiquetados.

El objetivo del Clustring es separar un conjunto de datos finite no etiquetados en un conjunto “natural”, escondido en la estructura de los datos.

Medidas de similaridad

Measures	Forms	Comments
Minkowski distance	$D_{ij} = \left(\sum_{l=1}^{d} \left\| x_{il} - x_{jl} \right\|^{1/n} \right)^{n}$	Metric. Invariant to any translation and rotation only for $n = 2$ (Euclidean distance). Features with large values and variances tend to dominate over other features.
Euclidean distance	$D_{ij} = \left(\sum_{l=1}^{d} \left\| x_{il} - x_{jl} \right\|^{1/2} \right)^{2}$	The most commonly used metric. Special case of Minkowski metric at $n = 2$. Tend to form hyperspherical clusters.
City-block distance	$D_{ij} = \sum_{l=1}^{d} \left\| x_{il} - x_{jl} \right\|$	Special case of Minkowski metric at $n = 1$. Tend to form hyperrecutangular clusters.
Sup distance	$D_{ij} = \max_{1 \le l \le d} \left\| x_{il} - x_{jl} \right\|$	Special case of Minkowski metric at $n \to \infty$.

Clúster no Jerárquico

Inicialmente se establece el número de grupos y cada caso se asigna a uno de ellos.

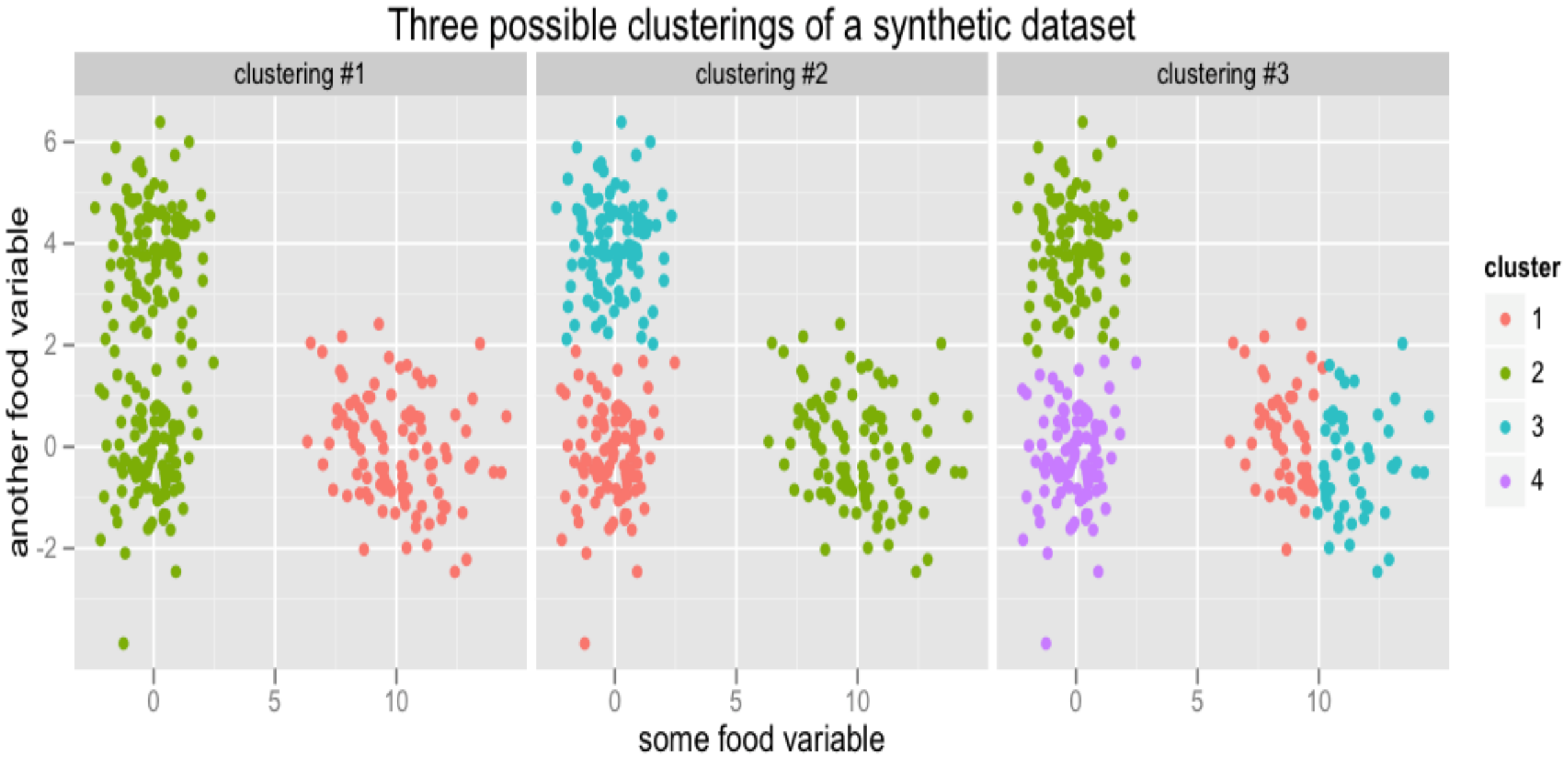
Dado un conjunto $X = \{x_1, \dots, x_2, \dots, x_3\}$, donde $x_j = \left(x_{j1}\ x_{j2}, \dots, x_{jd}\right)$.

Un cluster particional, busca k particones de X, $C = \{C_1, \dots, C_K\}$ $(K \leq N)$ tal que:

1.- $C_i \neq \emptyset, i = 1, \dots, K$;

2.- $\bigcup_{i=1}^{K} C_i = X$;

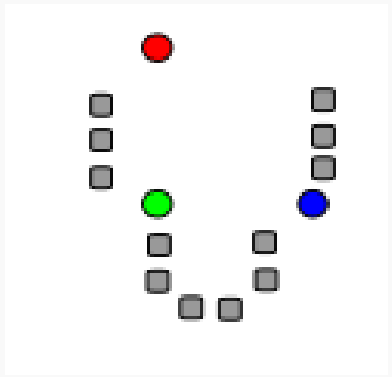
3.- $C_i \cap C_j = \emptyset, i, j = 1, \dots, K\ con\ i \neq j$;


Three possible clusterings of a synthetic dataset
clustering #1
clustering #2
clustering #3
another food variable
some food variable
6
4
2
0
-2
0
5
10
cluster
1
2
3
4

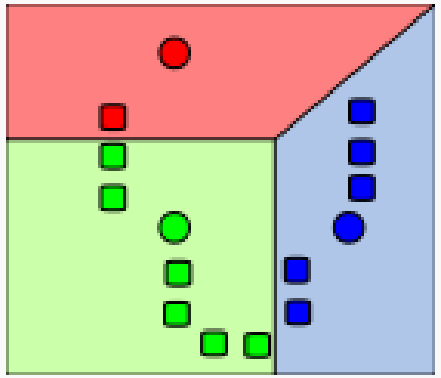
K-Means

K-means es un método de agrupamiento, que tiene como objetivo la partición de un conjunto n observaciones en k grupos en el que cada observación pertenece al grupo más cercano a la media.

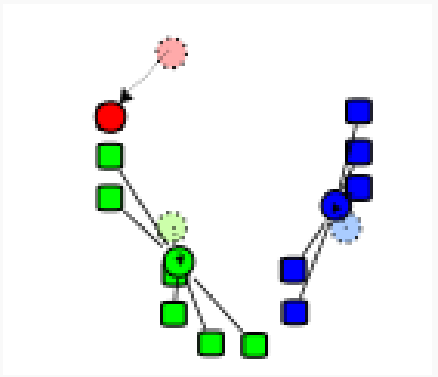
Demonstration of the standard algorithm



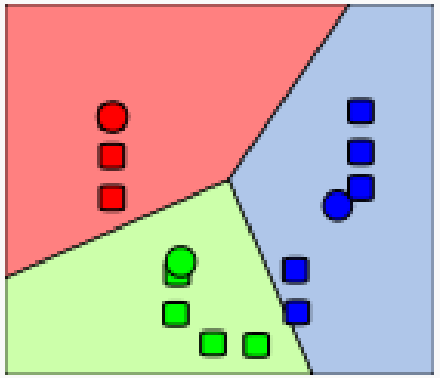
1) *k* initial "means" (in this case *k*=3) are randomly selected from the data set (shown in color).



2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



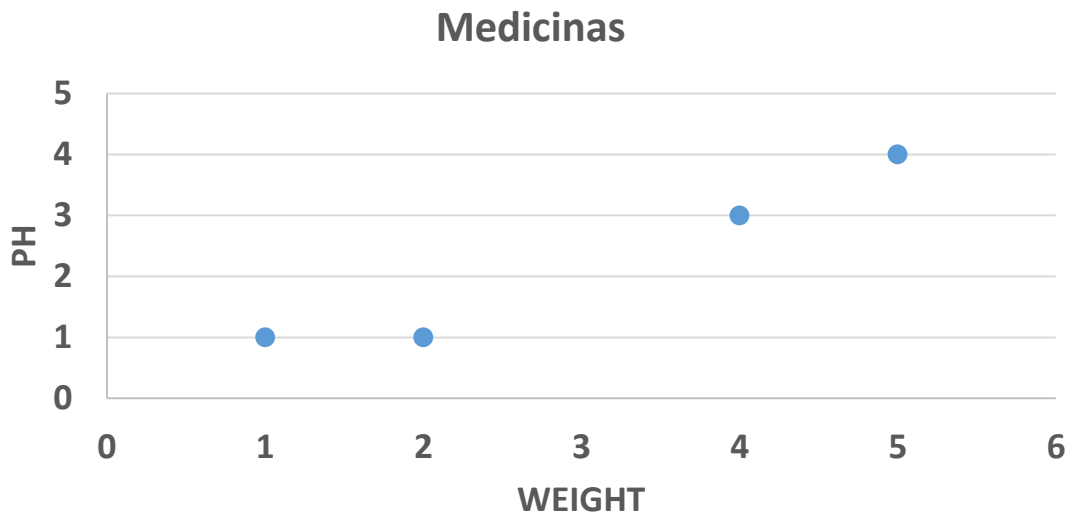
3) The centroid of each of the *k* clusters becomes the new means.



4) Steps 2 and 3 are repeated until convergence has been reached.

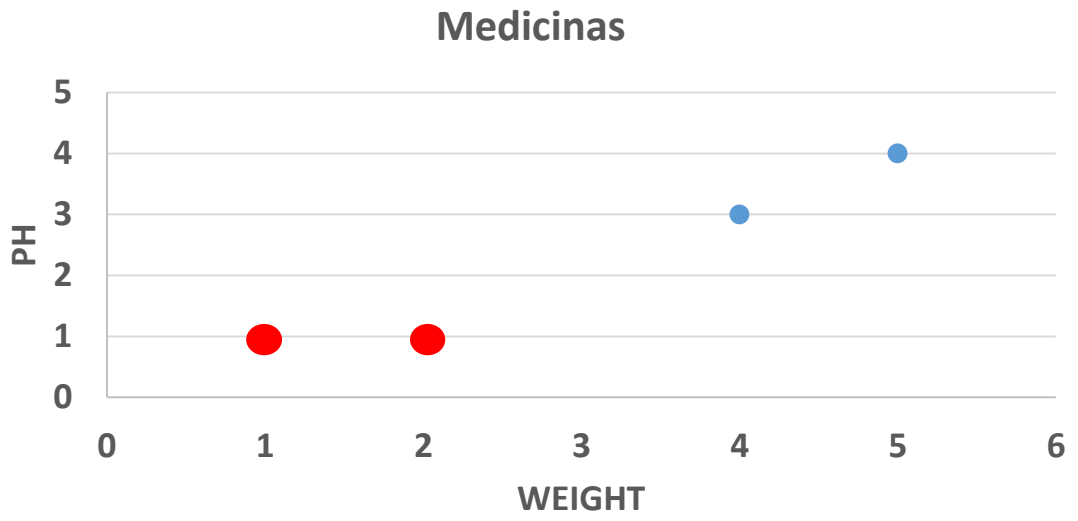
Ejemplo *K-Means: Inicio*

Elemento	*Peso*	*PH*
Medicina A	1	1
Medicina B	2	1
Medicina C	4	3
Medicina D	5	4



Ejemplo *K-Means: Inicio*

Elemento	*Peso*	*PH*
Medicina A	1	1
Medicina B	2	1
Medicina C	4	3
Medicina D	5	4

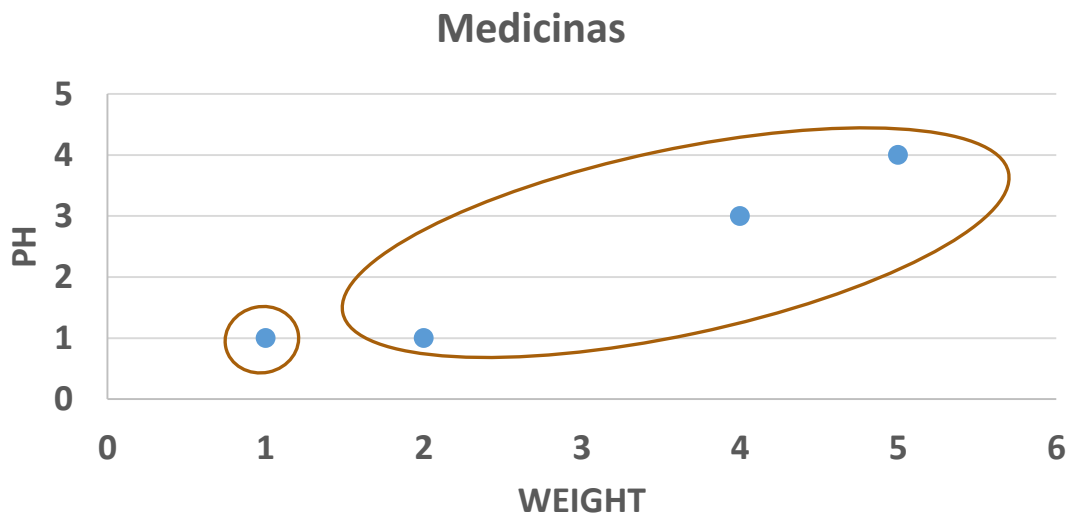


	MedA	*MedB*	*MedC*	*MedD*	*Centroide 1*	*Centroide 2*
X	1	2	4	5	1	2
Y	1	1	3	4	1	1

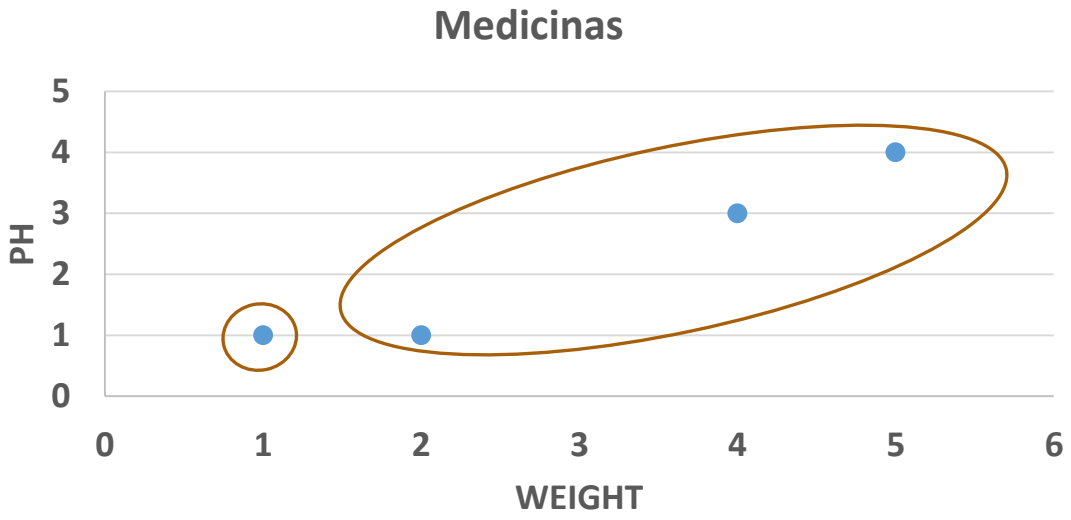
Ejemplo *K-Means: Iteración 1*

	MedA	*MedB*	*MedC*	*MedD*	*Centroide 1*	*Centroide 2*
X	1	2	4	5	1	2
Y	1	1	3	4	1	1

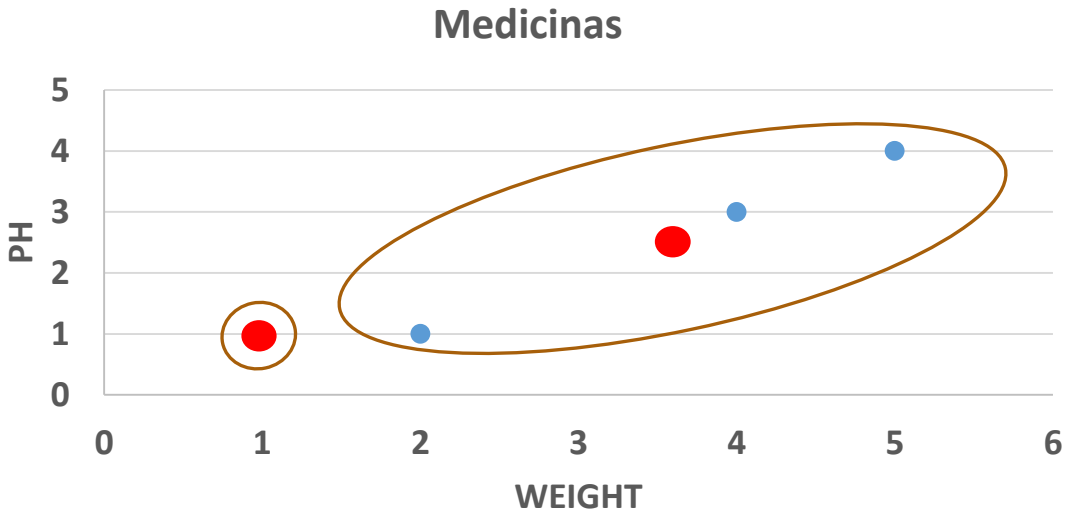
Distancia Euclidiana	*MedA*	*MedB*	*MedC*	*MedD*
Centroide 1	0	1	3,6	5,0
Centroide 2	1	0	2,8	4,2



Ejemplo *K-Means: Iteración 2*



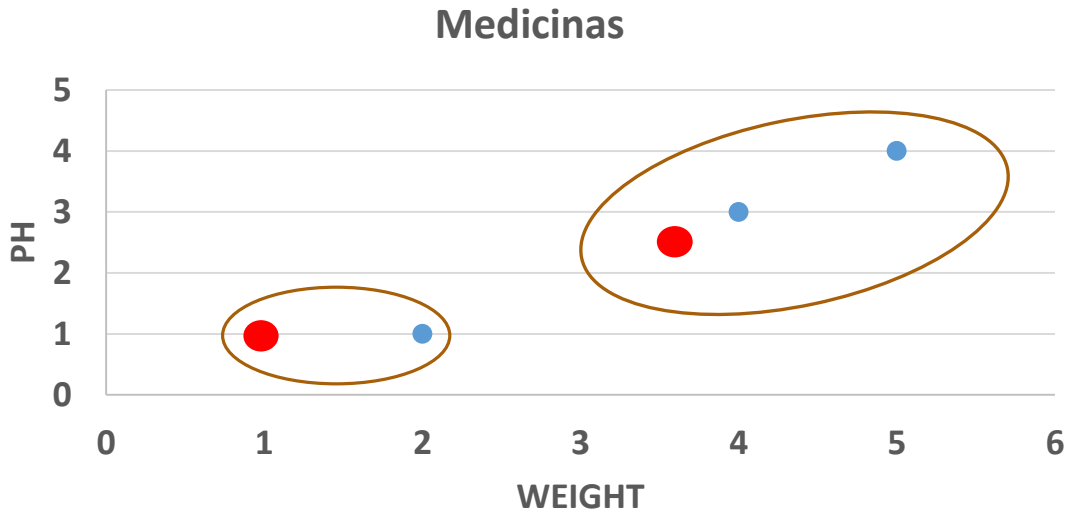
	Centroide 1	*Centroide 2*	*Centroide 1*	*Centroide 2*
Nuevos Centroides	1	(2+4+5)/3	1	3,7
	1	(1+3+4)/3	1	2,7



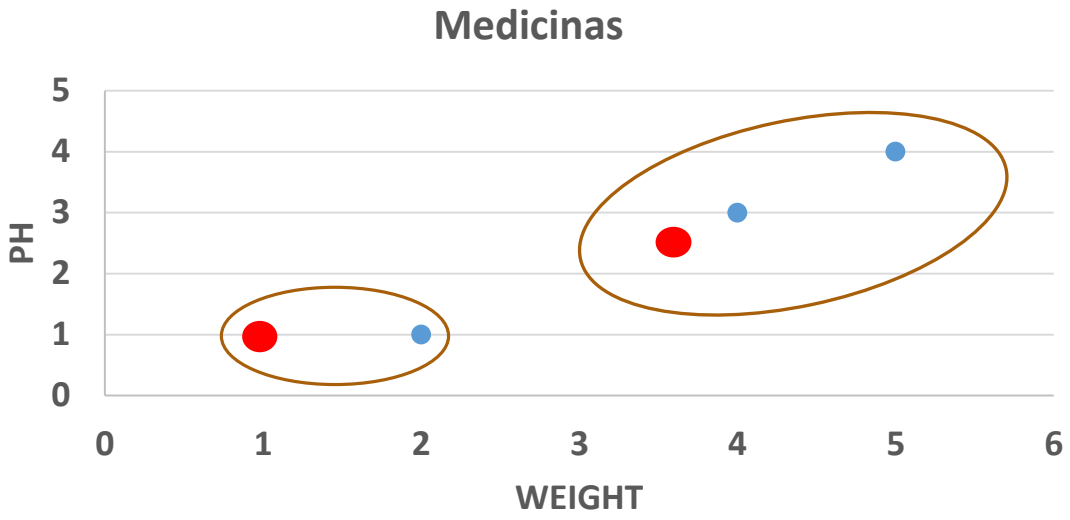
Ejemplo *K-Means: Iteración 2*

	MedA	*MedB*	*MedC*	*MedD*	*Centroide 1*	*Centroide 2*
X	1	2	4	5	1	3,7
Y	1	1	3	4	1	2,7

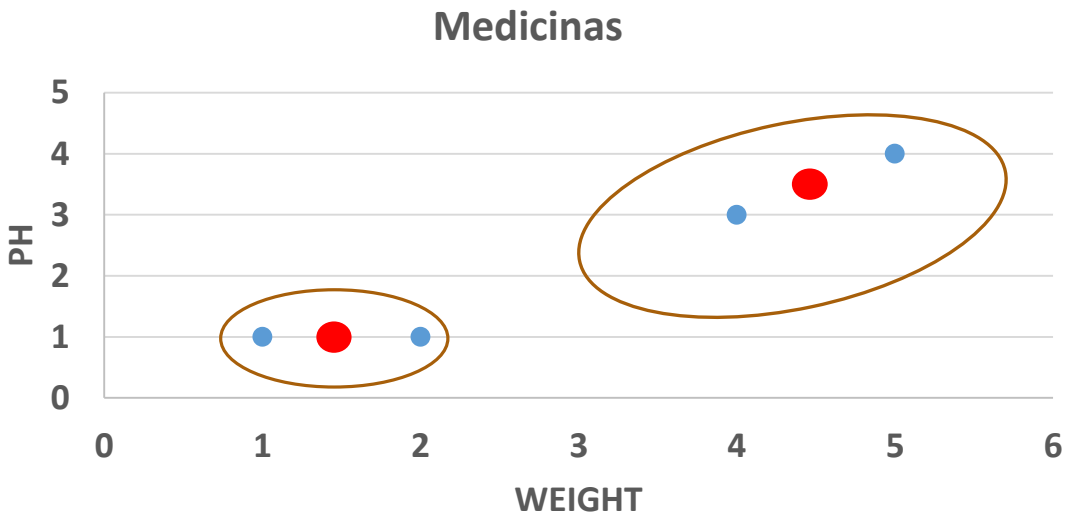
Distancia Euclidiana	*MedA*	*MedB*	*MedC*	*MedD*
Centroide 1	0	1	3,6	5,0
Centroide 2	3,1	2,4	0,5	1,9



Ejemplo *K-Means: Iteración 3*



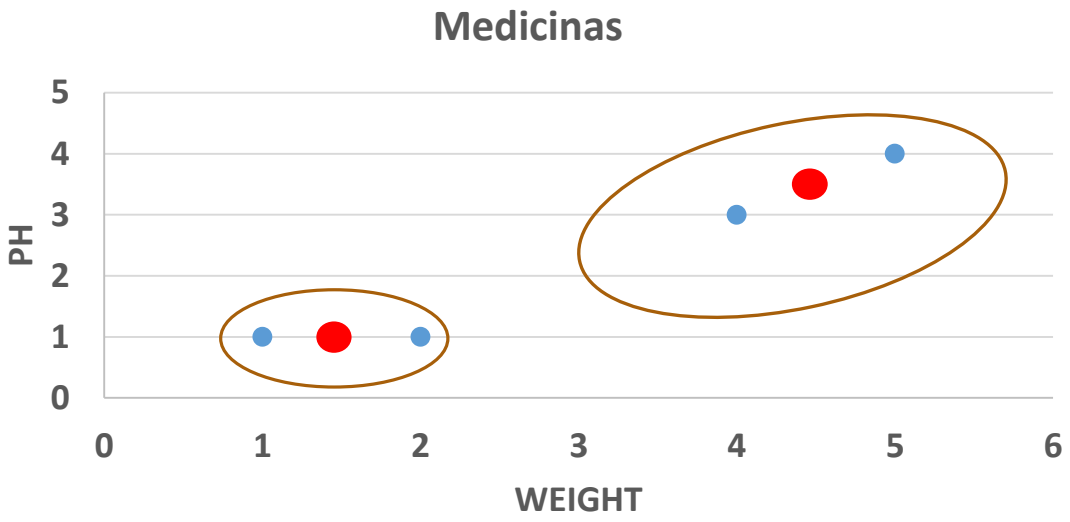
Nuevos Centroides	*Centroide 1*	*Centroide 2*	*Centroide 1*	*Centroide 2*
	(1+2)/2	(4+5)/2	1,5	4,5
	(1+1)/2	(3+4)/2	1	3,5



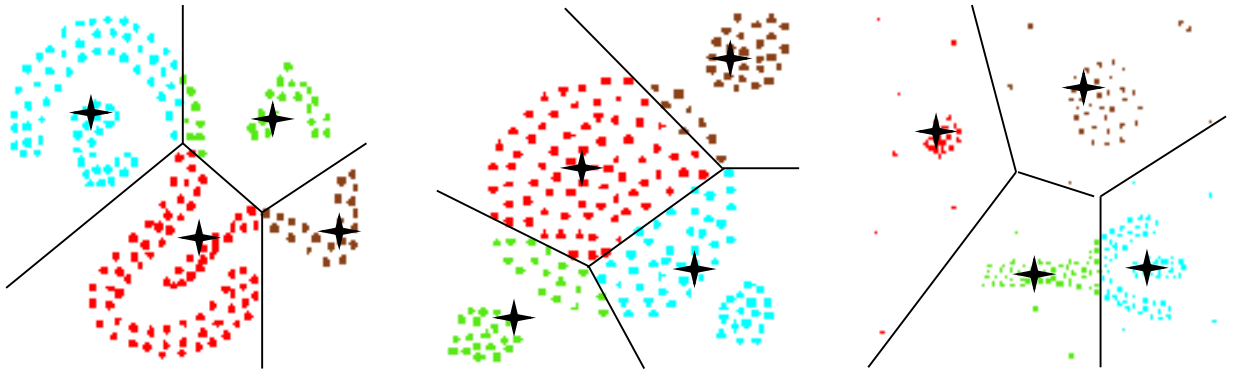
Ejemplo *K-Means: Iteración 3*

	MedA	*MedB*	*MedC*	*MedD*	*Centroide 1*	*Centroide 2*
X	1	2	4	5	1,5	4,5
Y	1	1	3	4	1	3,5

DISTANCIA EUCLIDIANA	*MedA*	*MedB*	*MedC*	*MedD*
Centroide 1	1	1	3,2	4,6
Centroide 2	4,3	3,5	0,7	0,7



DBSCAN: Density Based Spatial Clustering of Applications with Noise (Ester et al., KDD'1996)



Results of a *k*-medias algorithm for *k*=4

Idea Básica: En datos espaciales, los clusters son regions densas, separadas por regions con objetos de baja densidad.

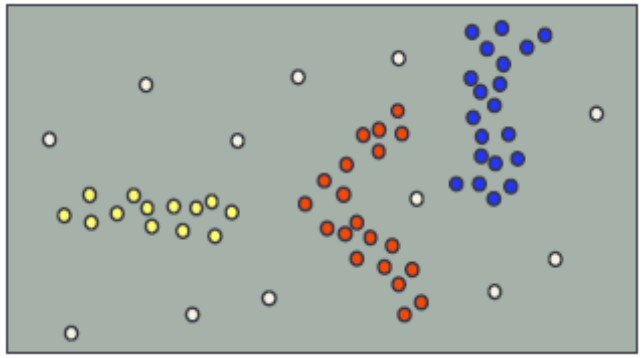
DBSCAN

Se basa en una noción de cluster basada en densidad.

Descubre clusters de forma arbitraria en bases de datos espaciales con ruido.

Idea básica

- Agrupar puntos en alta densidad.
- Marca como valores atípicos los puntos que se encuentran solos en regiones de baja densidad.



DBSCAN

Densidad del punto local en un punto p definido por dos parámetros:

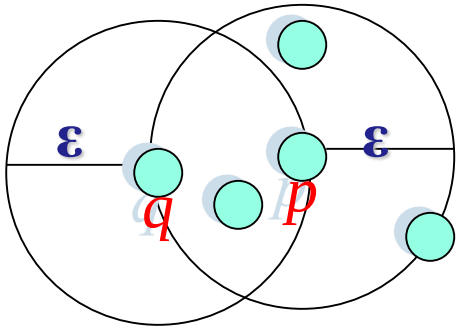
1.- **ε** : Radio para la vecindad (ε-Neighborhood)del punto p.

- ε-Neighborhood: todos los puntos dentro de un radio de ε desde el punto p
- N_ε (p): = {q en el conjunto de datos D | dist (p, q) ≤ ε}

2.- MinPts: Número mínimo de puntos en el vecindario dado N (p)

DBSCAN

ε-Neighborhood de un punto contiene al menos MinPts



ε-Neighborhood de p
ε-Neighborhood de q

Si el MinPts es 4

La densidad de p es Alta

La densidad de q es Baja

DBSCAN

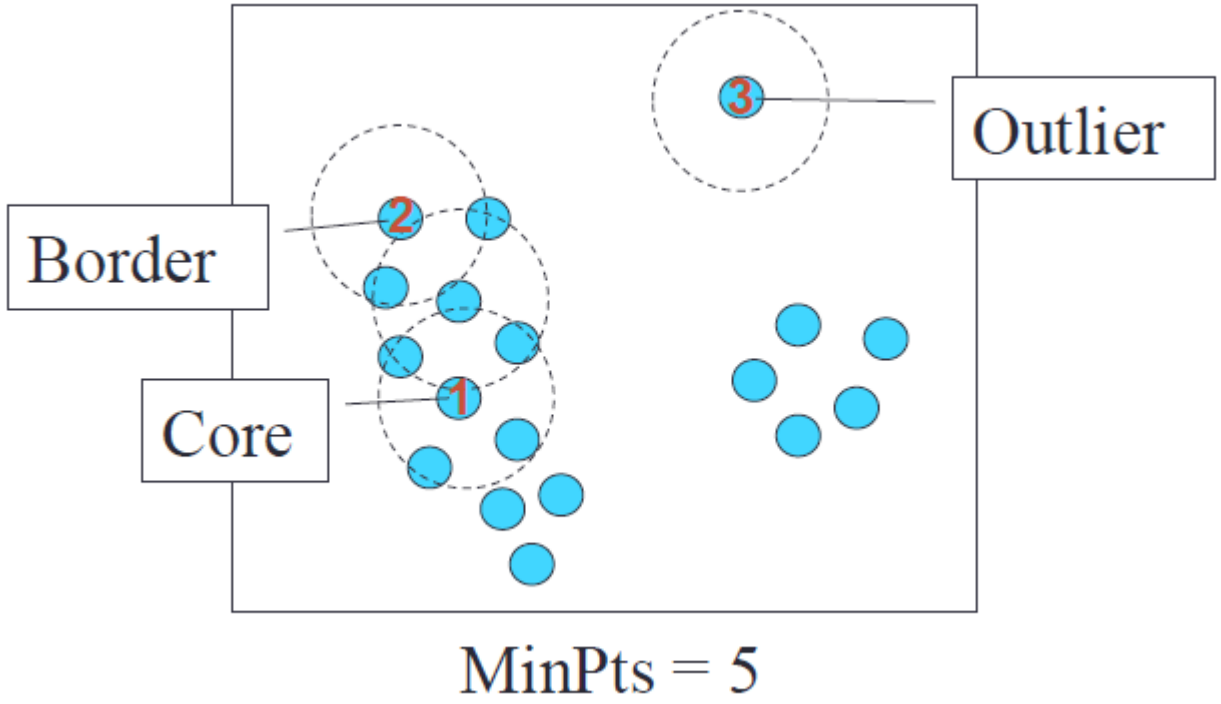
Core, Border & Outlier

Hay tres categorías para cada punto:

Punto Centro: Un punto es Centro si la densidad es Alta

Punto de Borde: Un punto es de Borde si la densidad es baja pero pertenece al vecindario de un Punto Centro.

Punto de Ruido o Fuera de Rango: Cualquier punto que no cea un Punto Cntro ni un Punto Borde.



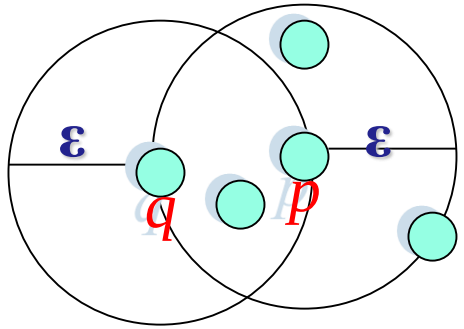
DBSCAN

Un objeto q es **directamente alcanzable por densidad (directly density-reachable)** desde el objeto p si p es un Punto Centro y q está en el ε-Neighborhood de p.

q es directamente alcanzable por densidad desde p.

p no es directamente alcanzable por densidad desde q.

La densidad-accesibilidad es asimétrica



MinPts = 4

DBSCAN

Un punto p es directamente alcanzable por densidad desde p2.

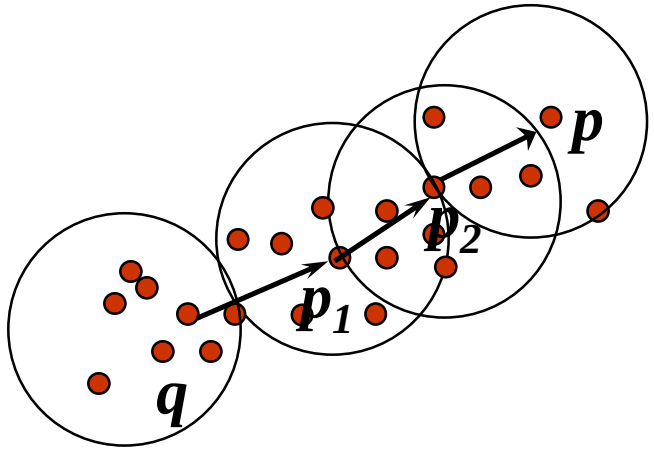
p2 es directamente alcanzable por densidad desde p1.

p1 es directamente alcanzable por densidad desde q.

p, p2, p1, q forman una cadena.

p es (**indirectamente) alcanzable por densidad** desde q

q **no es (indirectamente) alcanzable por densidad** densidad-accesible desde p

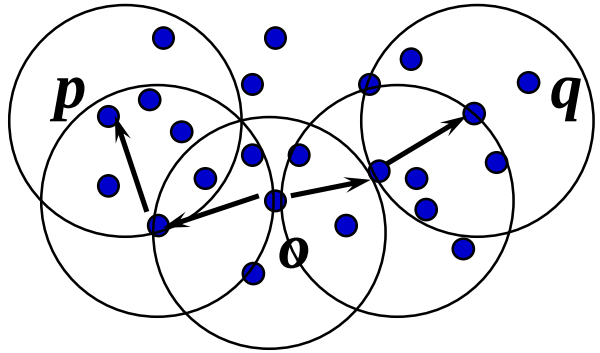


MinPts = 7

DBSCAN

Un par de puntos p y q están **conectados por densidad** si son comúnmente alcanzable por densidad desde punto o.

La **conexión por densidad** es simétrica.



DBSCAN: Descripción formal

Dado un conjunto de datos D, parámetro ε y umbral MinPts.

Un Cluster C es un subconjunto de objetos que satisfacen dos criterios:

Conectado: $\forall$ p, q $\in$ C: p y q están conectados por densidad.

Máximo: $\forall$ p, q $\in$ C y q es alcanzable por densidad desde el Punto Centro p, entonces q $\in$ C.

DBSCAN: Algoritmo

Seleccionar en forma arbitraria un punto p

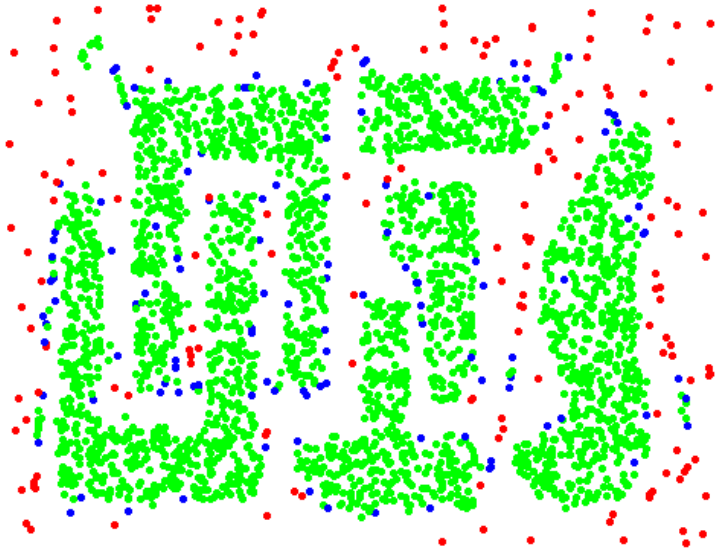
Recuperar todos los puntos densidad alcanzables de p definido ε y MinPts.

Si p es un punto central, se forma un grupo.

Si p es un punto de borde, no se puede alcanzar la densidad desde p y DBSCAN visita el siguiente punto de la base de datos.

Continuar el proceso hasta que todos los puntos hayan sido procesados.

Original Points



Point types: core,
border and outliers

$\varepsilon = 10$, MinPts = 4

Ejercicio

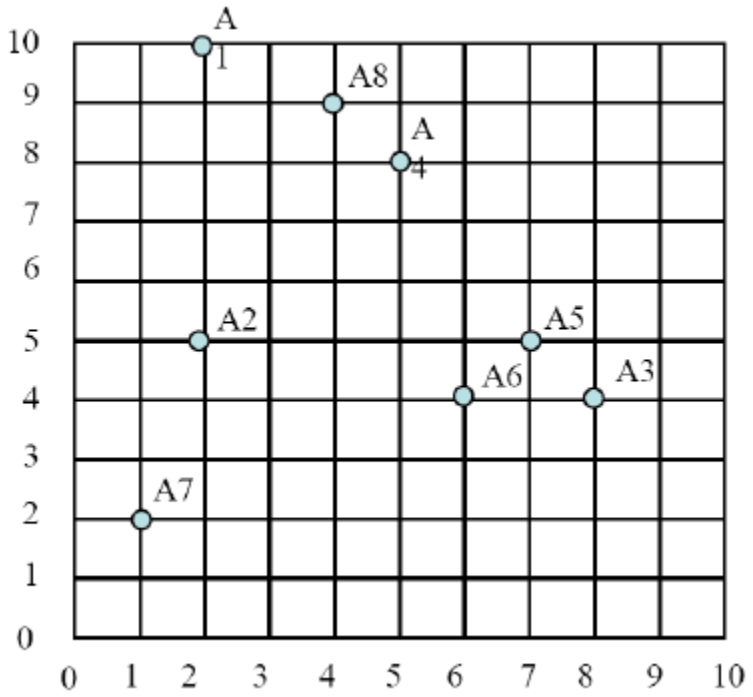
Agrupar los 8 puntos de la figura utilizando el algoritmo DBSCAN.

Número mínimo de puntos en el "vecindario":

MinPts = 2

Radio del "vecindario":

Epsilon $\sqrt{2} \leq \varepsilon < \sqrt{10}$



Ejercicio resuelto

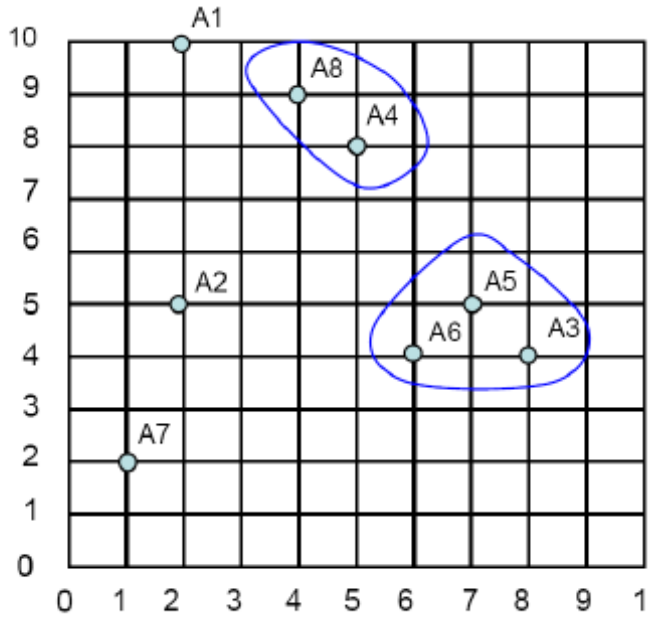
Distancia euclídea

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Ejercicio resuelto

Epsilon = $\sqrt{2}$

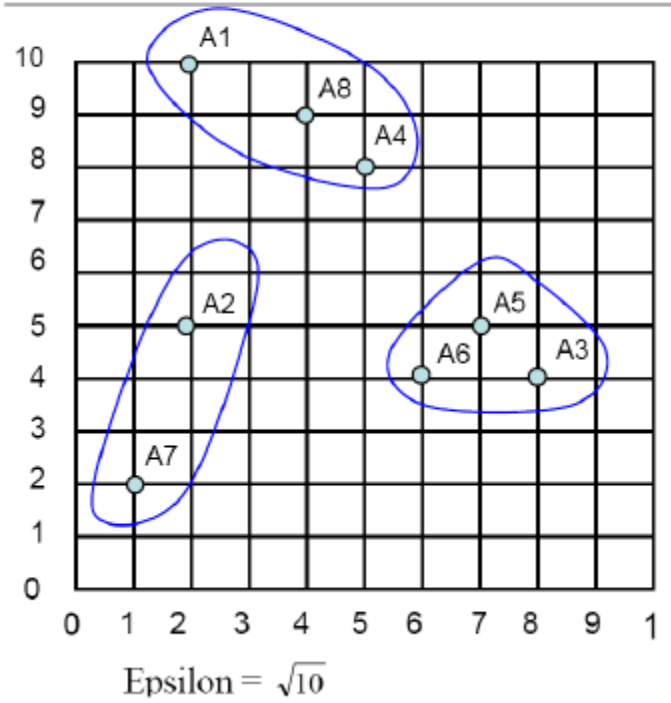
A1, A2 y A7 no tienen vecinos en su vecindario,
por lo que se consideran "outliers" (no están en zonas densas):



Ejercicio resuelto

Epsilon = $\sqrt{10}$

Al aumentar el valor del parámetro Epsilon,
el vecindario de los puntos aumenta y todos quedan agrupados:



Validación de Clusters

Es de importancia evaluar el resultado de los algoritmos de clustering, sin embargo, es difícil definir cuando el resultado de un agrupamiento es aceptable. Por esta razón existen técnicas e índices para la validación de un agrupamiento realizado.

Validación de Clusters

Metricas de Validación Internas están basadas usualmente en los criterios de cohesión y separación:

Cohesión: El miembro de cada clúster debe ser lo más cercano posible a los otros miembros del mismo clúster.

Separación: Los clúster deben estar ampliamente separados entre ellos. Existen varios enfoques para medir esta distancia entre clúster: distancia entre el miembro más cercano, distancia entre los miembros más distantes o la distancia entre los centroides.

Validación de Clusters

Sum of Squared Within (SSW)

Medida interna especialmente usada para evaluar la Cohesión de los clústeres que el algoritmo de agrupamiento generó.

$$SSE = \sum_{i=1}^{k} \sum_{x \epsilon C_i} dist^2(m_i, x)$$

Siendo k el número de clústeres, x un punto del clúster Ci y mi el centroide del clúster Ci.

Validación de Clusters

Es una medida de separación utilizada para evaluar la distancia inter-clúster (Separación).

Siendo k el número de clústeres, nj el número de elementos en el clúster j, cj el centroide del clúster j y x es la media del data set.

$$SSB = \sum_{j=1}^{k} n_j \, dist^2 \, (c_j - \overline{x})$$

Validación de Clusters

Los índices o medidas basadas en las «sumas de cuadrados» presentadas anteriormente se caracterizan por medir o cuantificar la dispersión de los puntos a nivel inter-cluster e intra-cluster. Los índices son:

Ball y Hall (1965)	$\frac{SSW}{k}$
Calinski y Harabasz (1974)	$\frac{SSB/(k-1)}{SSW/(n-k)}$
Hartigan (1975)	$log(\frac{SSB}{SSW})$
Xu (1997)	$\mathrm{d} * \log\left(\sqrt{\frac{SSW}{dN^2}}\right) + \log(k)$

Siendo k el número de clústeres, N el número de datos y d la dimensión de los datos.

Davies-Bouldin index (DB)

$$DB = \frac{1}{k} \sum_{i=1, i \neq j}^{k} \max \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

Donde k es el número de clústeres, σi es la distancia promedio entre cada punto en el clúster i y el centroide del clúster, σj es la distancia promedio entre cada punto del clúster j y el centroide del clúster, y $d(ci,cj)$ es la distancia entre los centroides de los 2 clústeres .

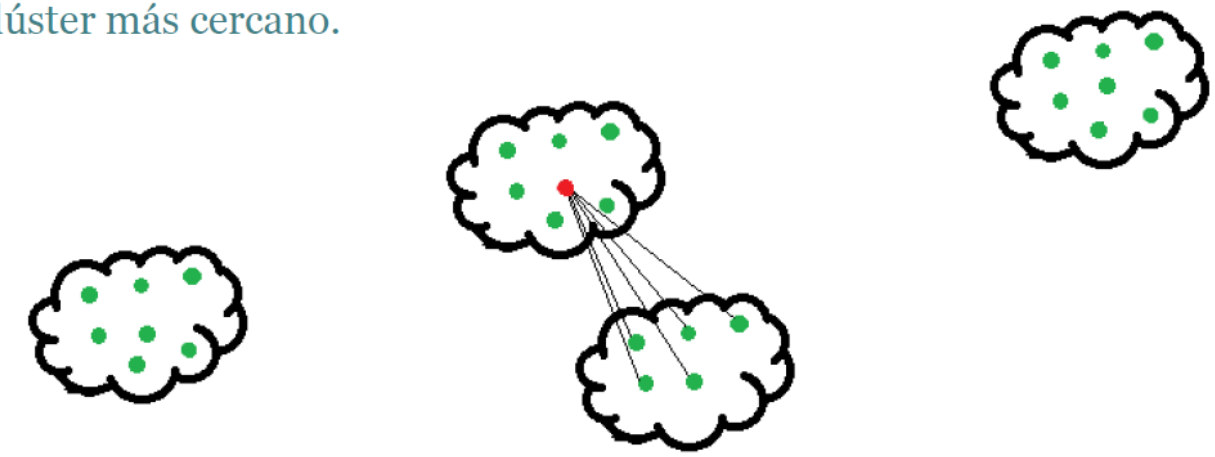
Valores pequeños para el índice DB indica clústeres compactos, y cuyos centros estas bien separados los unos de los otros. Consecuentemente el número de clústeres que minimiza el índice DB se toma como el óptimo.

Coeficiente de Silhouette

Dado un punto x del conjunto de datos :

- Cohesión a(x) : distancia promedio de x a todos los demás puntos en el mismo clúster.

- Separación b(x) : distancia promedio de x a todos los demás puntos en el clúster más cercano.



Coeficiente de Silhouette

El coeficiente de silhouette para el punto x está definido como:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

- Donde el valor de s(x) puede variar entre -1 y 1. –
  - -1 = mal agrupamiento
  - 0 = indiferente
  - 1 = bueno

El coeficiente de silhouette para el punto x está definido como:

$$SC = \frac{1}{N}\sum_{i=1}^{N} s(x)$$

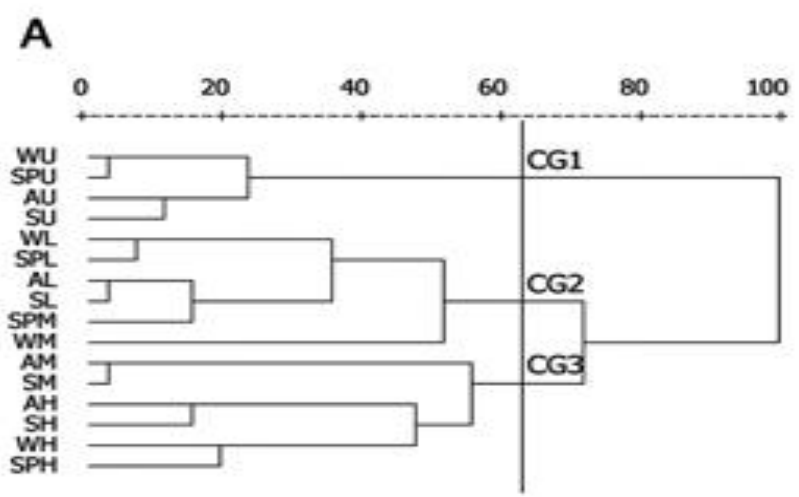
Clúster Jerárquico

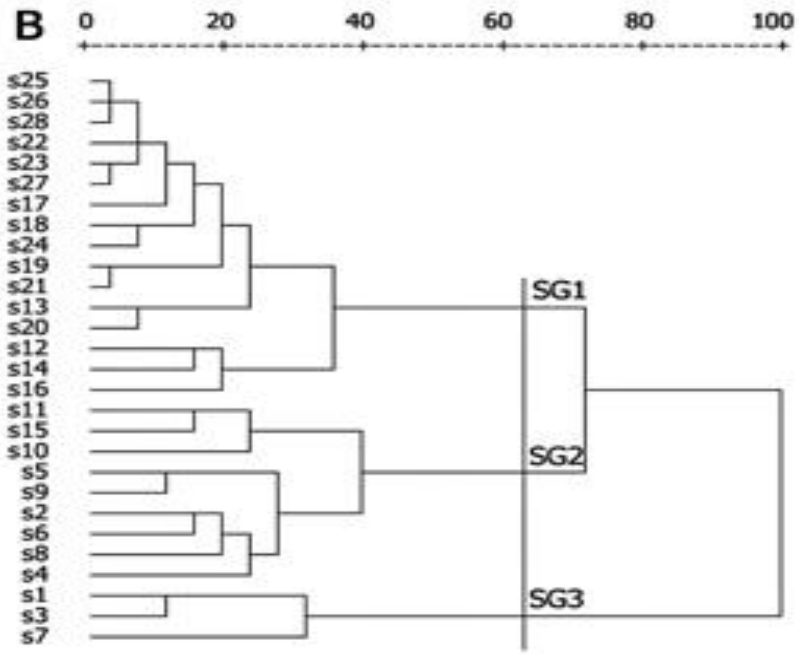
Inicialmente cada caso es un grupo en sí mismo y sucesivamente se van fusionando grupos cercanos hasta que todos los individuos confluyen en un solo grupo.

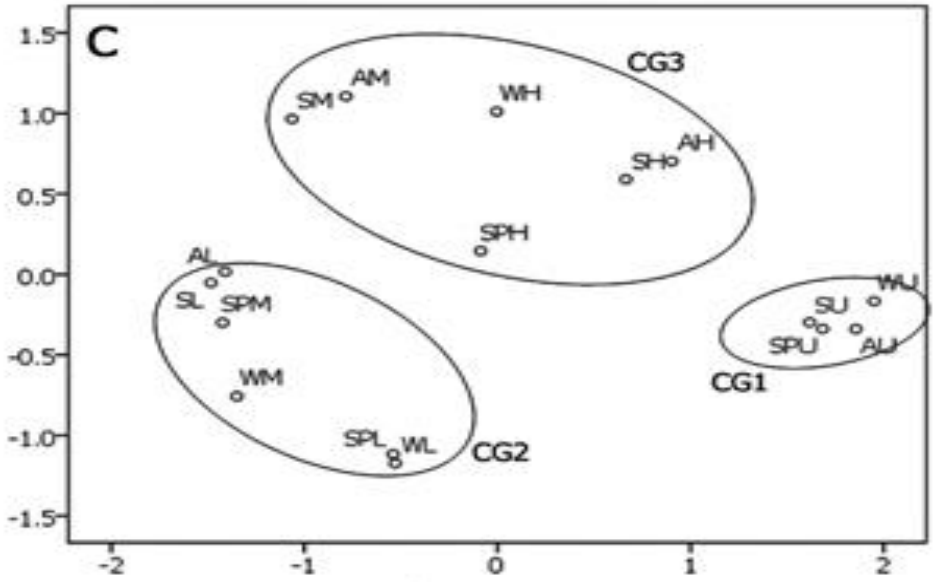
Un cluster jerarquico busca constrir una estructura de árbol anidado de X,

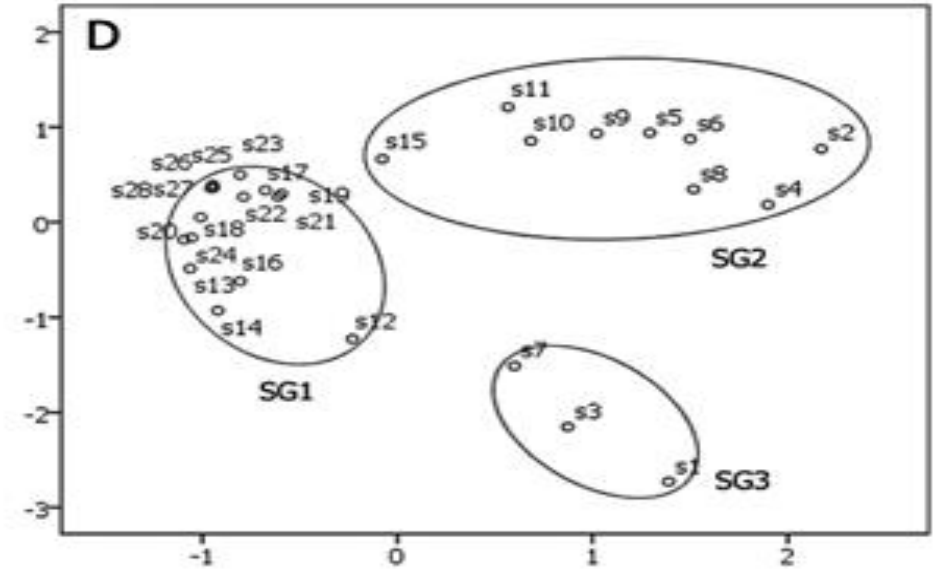
$\mathrm{H} = \{H_1, \ldots, H_Q\}$ $(Q \leq N)$ tal que $C_i \in H_m$, $C_i \in H_l$, con $m > l$, implicando que:

$C_i \in C_j \; o \; C_i \cap C_j = \emptyset \; \forall \; i, j \neq i, m, l = 1, \ldots, Q.$


A
0
20
40
60
80
100
WU
SPU
AU
SU
WL
SPL
AL
SL
SPM
WM
AM
SM
AH
SH
WH
SPH
CG1
CG2
CG3


B
0
20
40
60
80
100
s25
s26
s28
s22
s23
s27
s17
s18
s24
s19
s21
s13
s20
s12
s14
s16
s11
s15
s10
s5
s9
s2
s6
s8
s4
s1
s3
s7
SG1
SG2
SG3


C
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2
-1
0
1
2
AM
SM
WH
AH
SH
SPH
AL
SL
SPM
WM
SPL
WL
WU
SU
SPU
AU
CG1
CG2
CG3


D
2
1
0
-1
-2
-3
-1
0
1
2
s11
s10
s9
s5
s6
s2
s15
s8
s4
s23
s26
s25
s17
s28
s27
s19
s22
s21
s20
s18
s24
s16
s13
s14
s12
s7
s3
s1
SG1
SG2
SG3

Clúster Jerárquico: Utilizando el Método del vecino más cercano

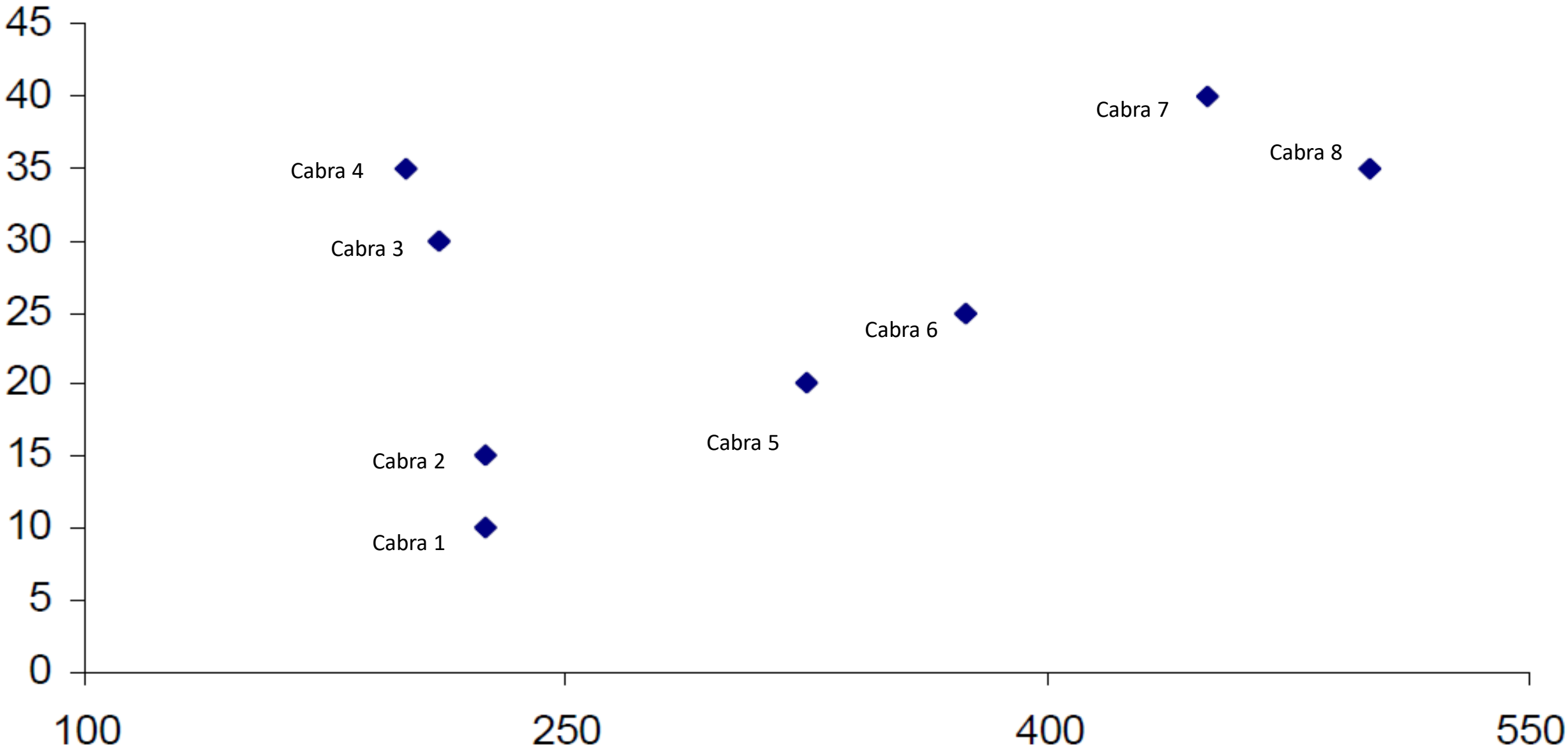
Comienza uniendo las **dos observaciones más cercanas**.

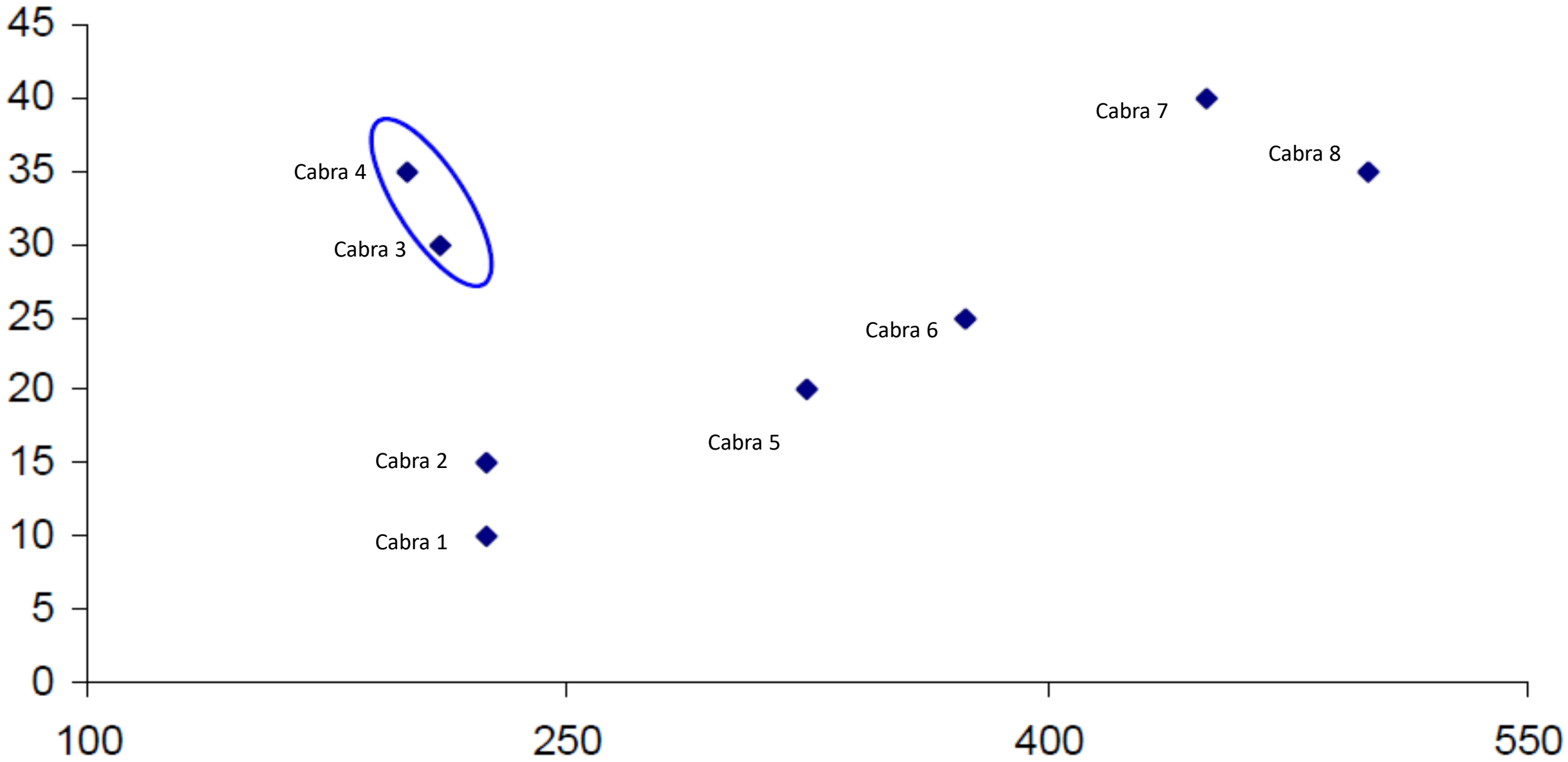
A continuación, el grupo se sustituye por una observación que lo representa.

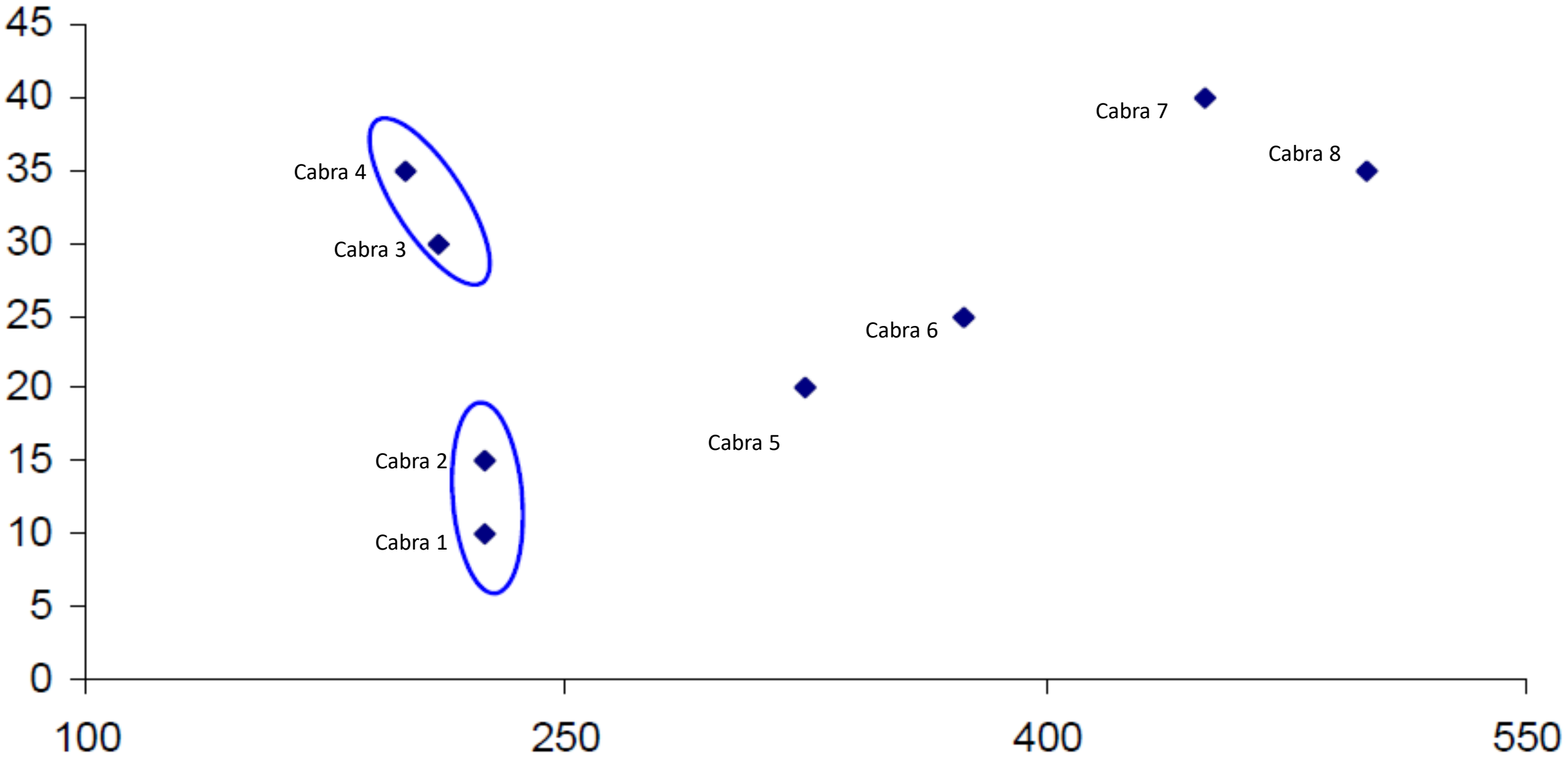
Las distancias entre los grupos a fusionar se calculan tomando las **observaciones más cercanas de cada grupo.**

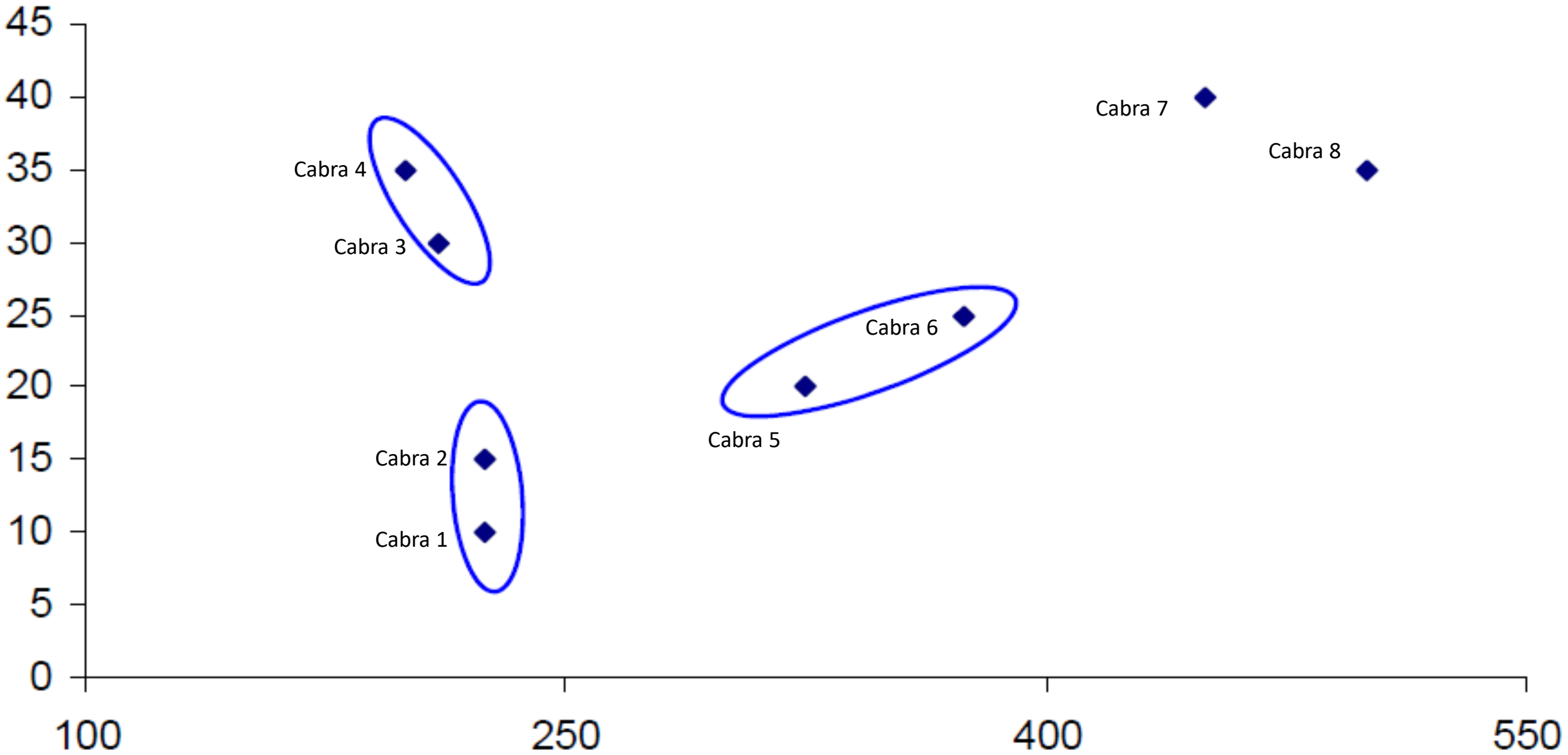
Ejemplo

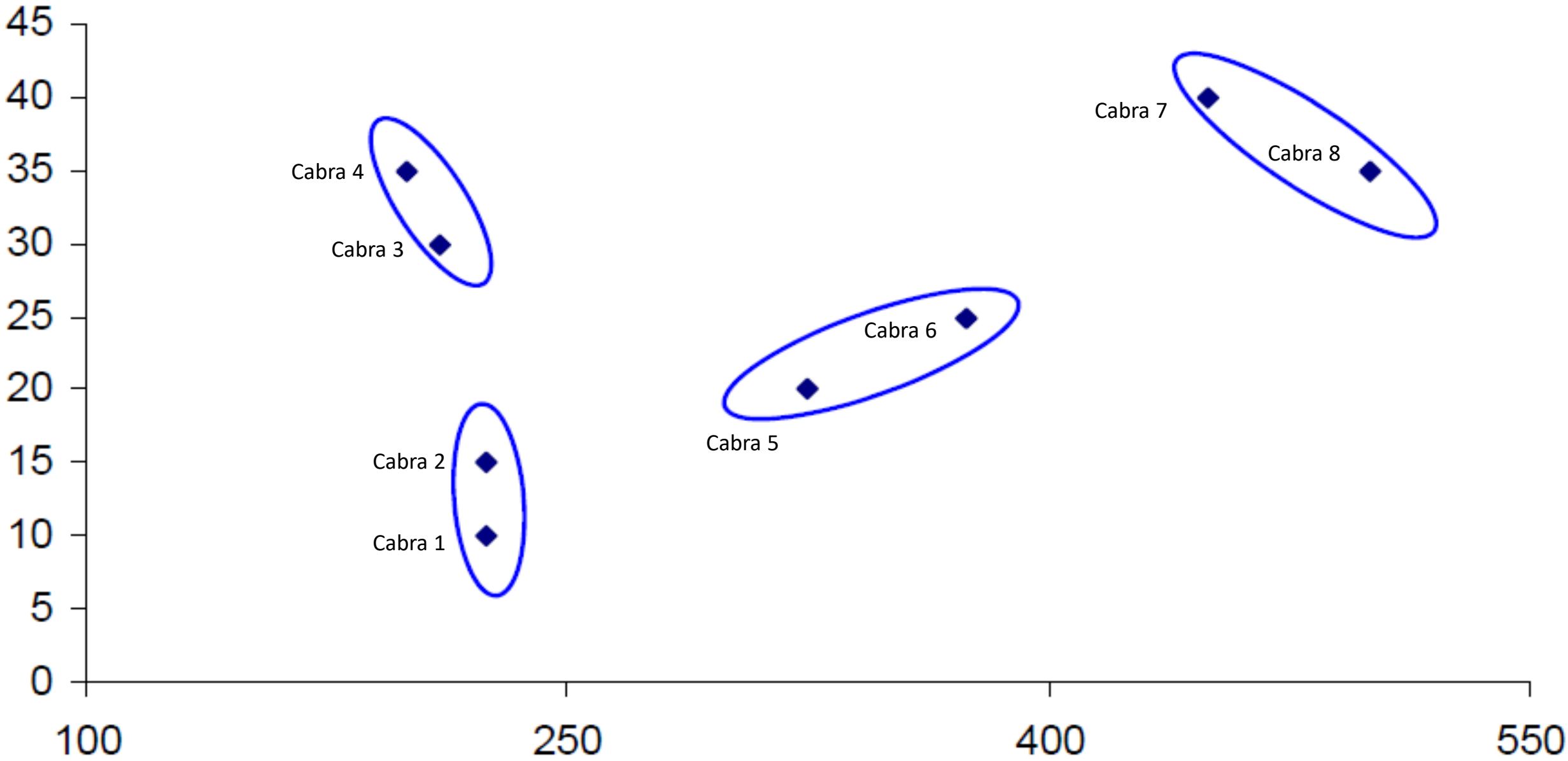
Cabra	Producción leche	Rendimiento quesero
1	225	10
2	225	15
3	210	30
4	200	35
5	325	20
6	375	25
7	450	40
8	500	35

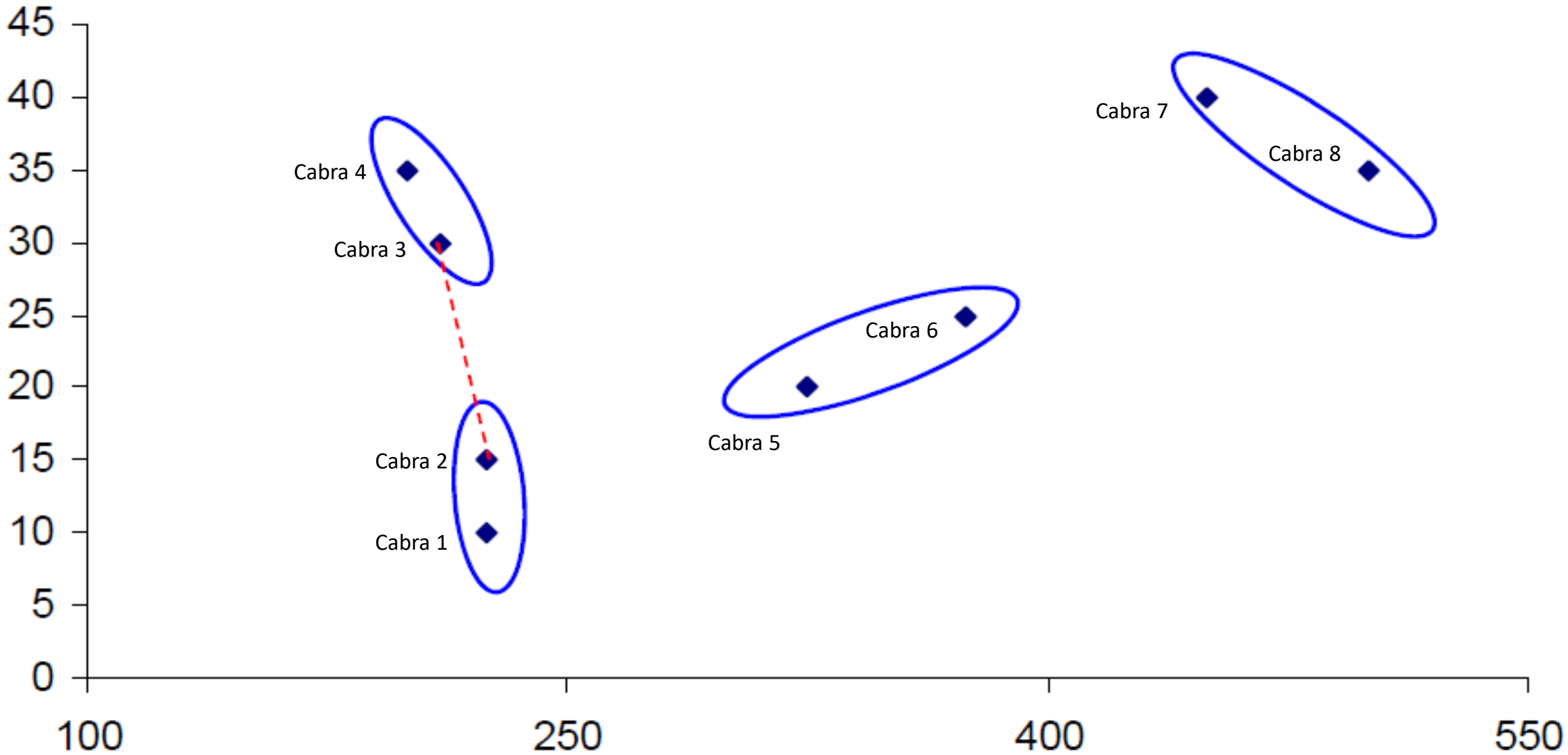

45
40
35
30
25
20
15
10
5
0
100
250
400
550
Cabra 4
Cabra 3
Cabra 2
Cabra 1
Cabra 5
Cabra 6
Cabra 7
Cabra 8

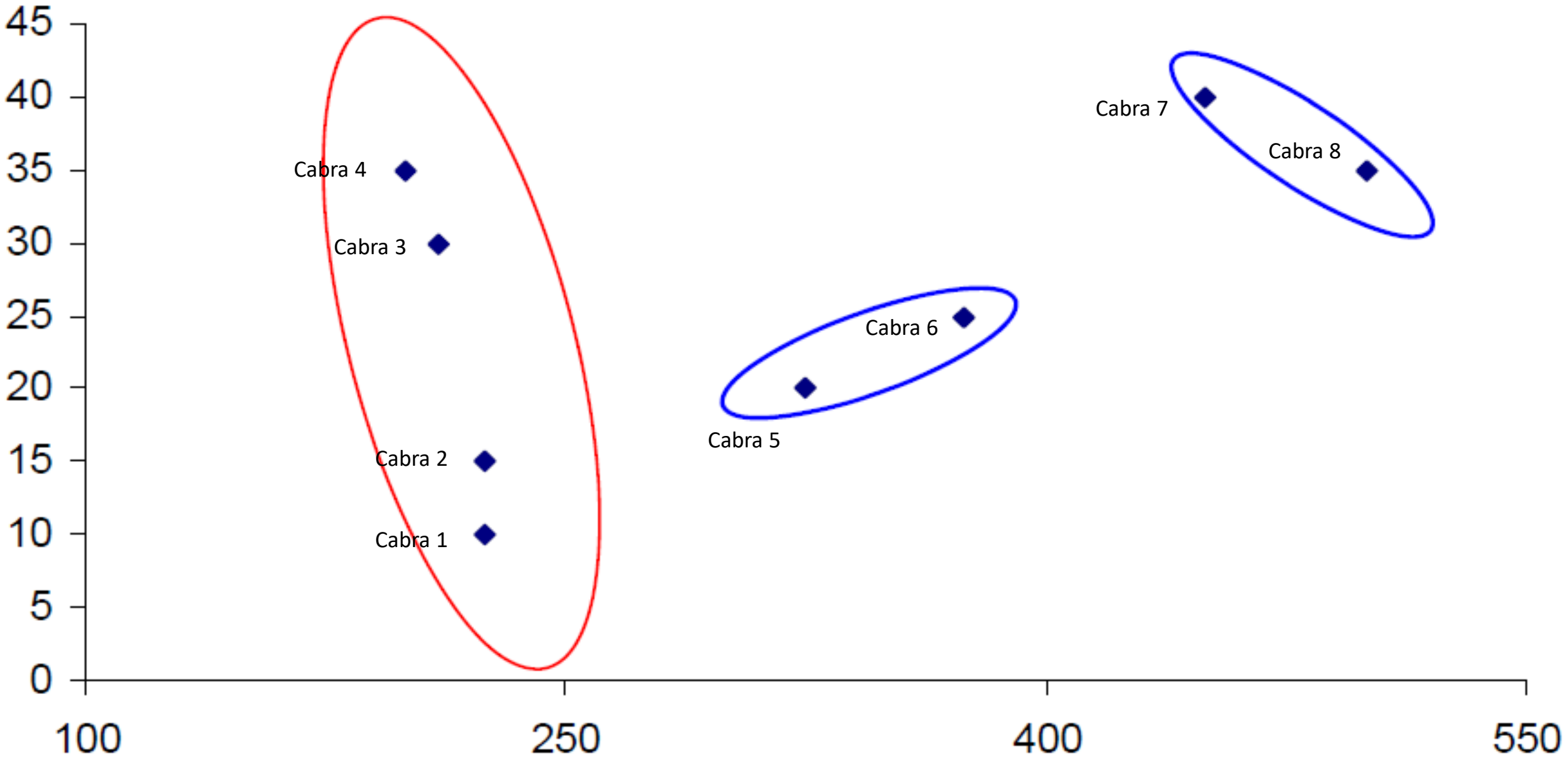

45
40
35
30
25
20
15
10
5
0
100
250
400
550
Cabra 4
Cabra 3
Cabra 2
Cabra 1
Cabra 5
Cabra 6
Cabra 7
Cabra 8

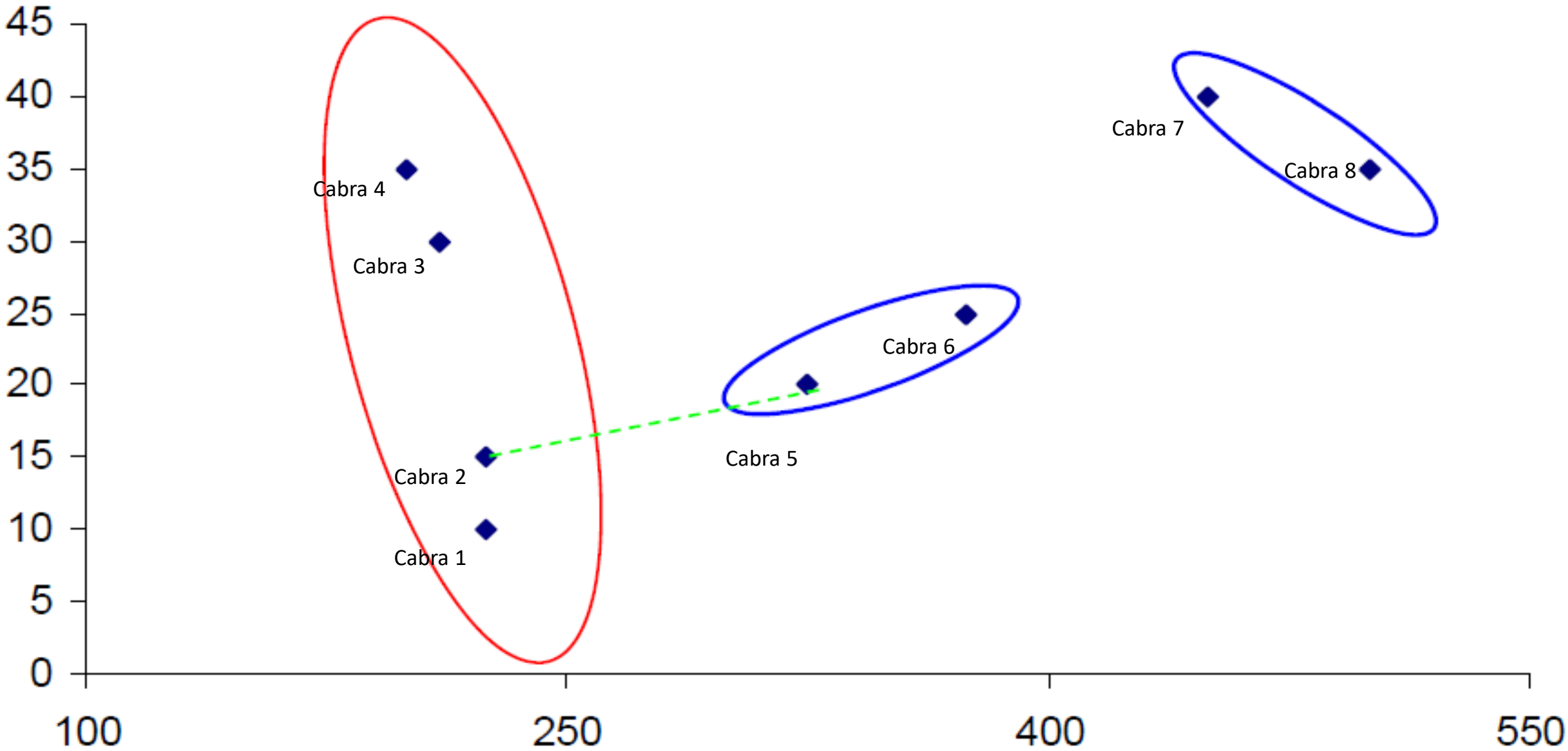

45
40
35
30
25
20
15
10
5
0
100
250
400
550
Cabra 4
Cabra 3
Cabra 2
Cabra 1
Cabra 5
Cabra 6
Cabra 7
Cabra 8

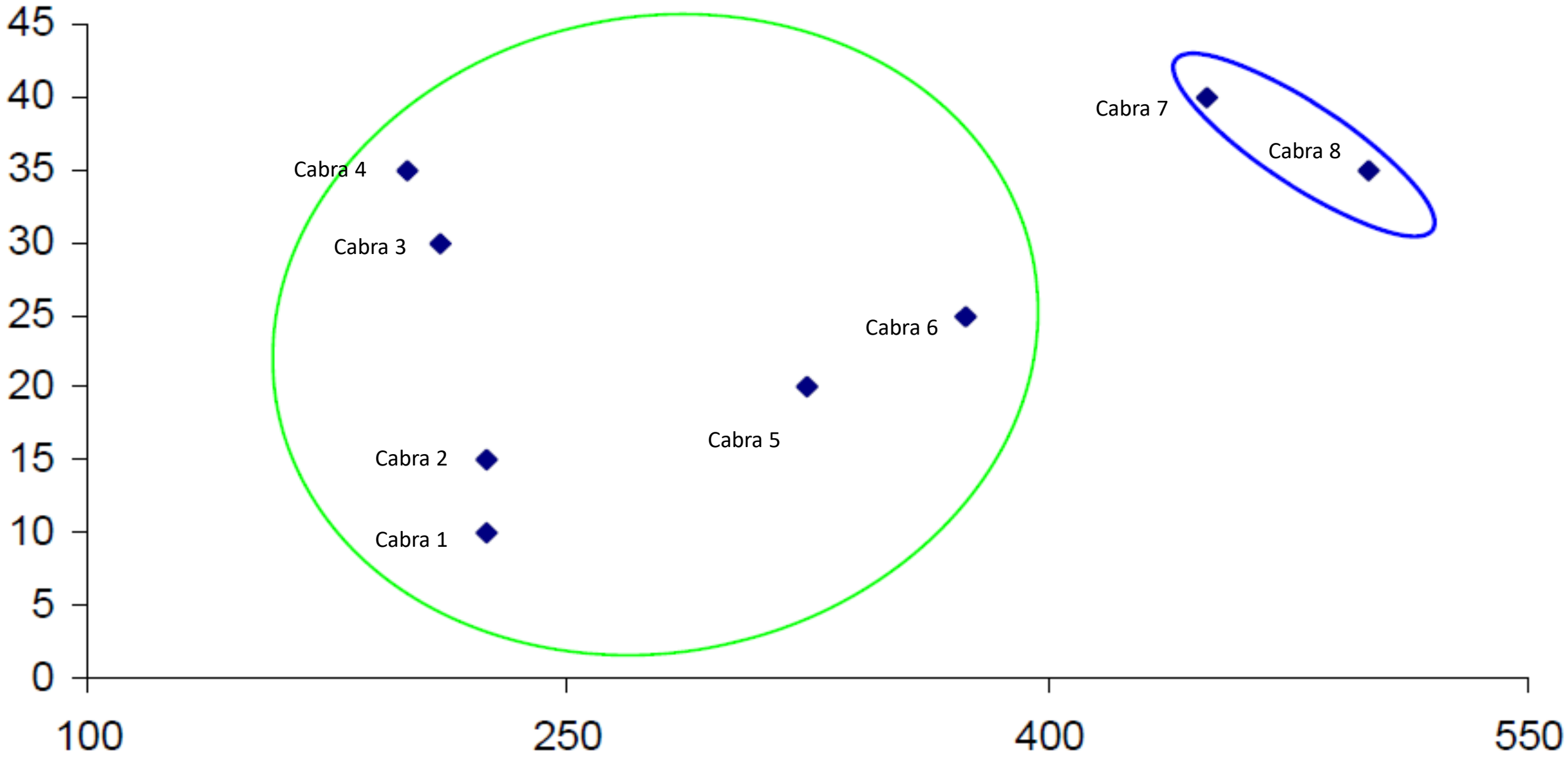

45
40
35
30
25
20
15
10
5
0
100
250
400
550
Cabra 4
Cabra 3
Cabra 7
Cabra 8
Cabra 6
Cabra 5
Cabra 2
Cabra 1

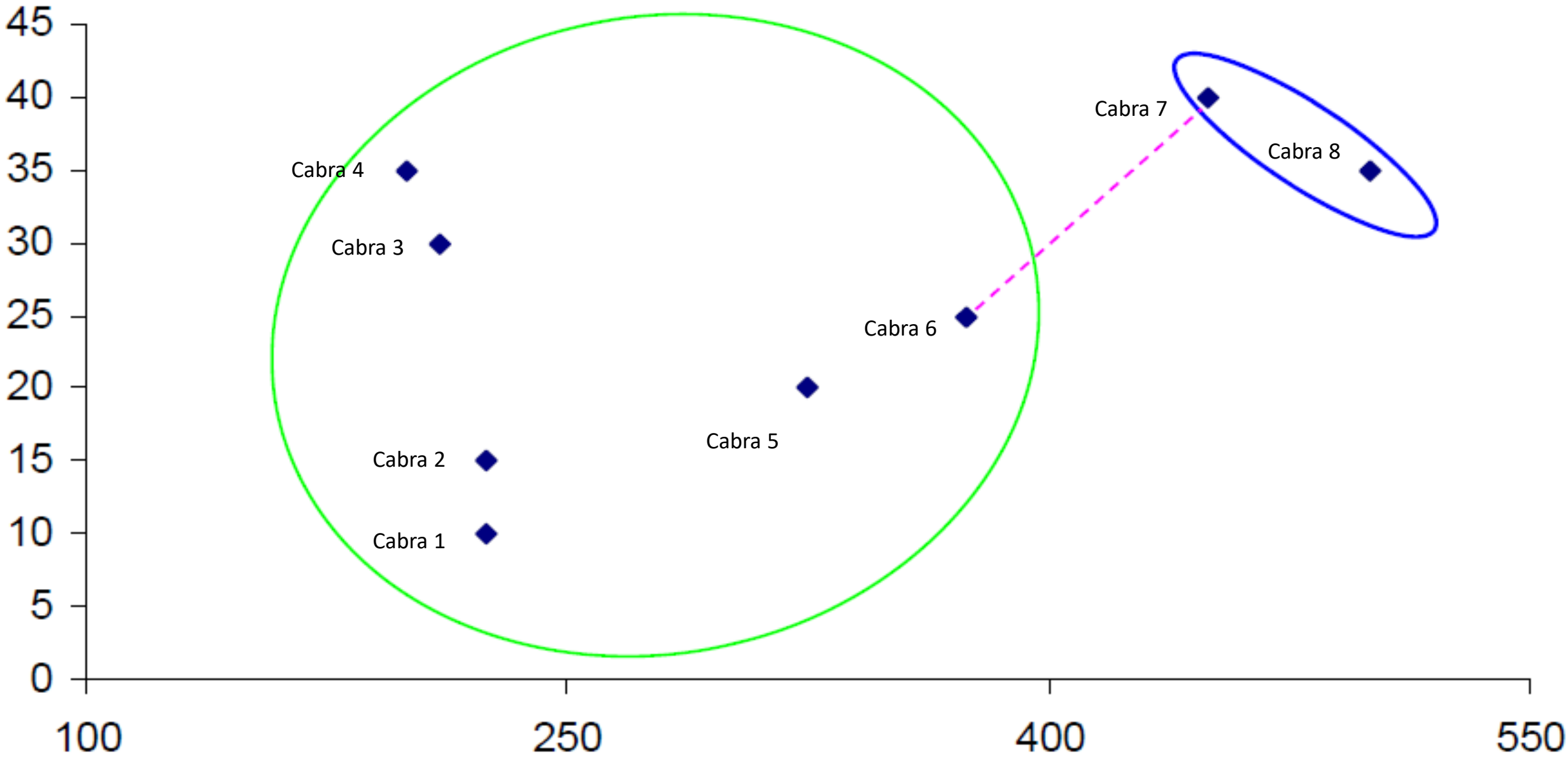

45
40
35
30
25
20
15
10
5
0
100
250
400
550
Cabra 4
Cabra 3
Cabra 2
Cabra 1
Cabra 6
Cabra 5
Cabra 7
Cabra 8

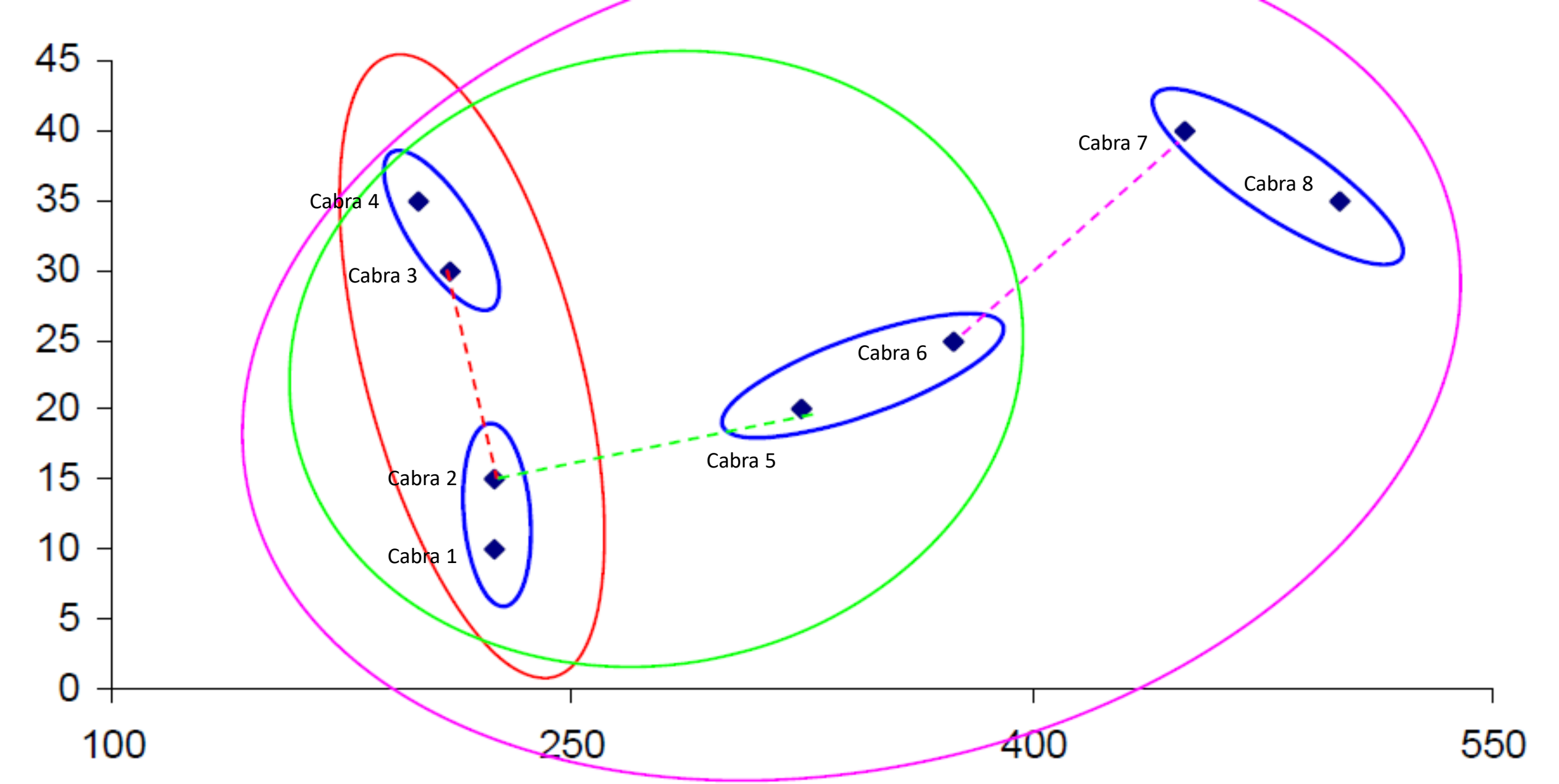

45
40
35
30
25
20
15
10
5
0
100
250
400
550
Cabra 4
Cabra 3
Cabra 2
Cabra 1
Cabra 5
Cabra 6
Cabra 7
Cabra 8


45
40
35
30
25
20
15
10
5
0
100
250
400
550
Cabra 4
Cabra 3
Cabra 2
Cabra 1
Cabra 5
Cabra 6
Cabra 7
Cabra 8


45
40
35
30
25
20
15
10
5
0
100
250
400
550
Cabra 4
Cabra 3
Cabra 2
Cabra 1
Cabra 5
Cabra 6
Cabra 7
Cabra 8


45
40
35
30
25
20
15
10
5
0
100
250
400
550
Cabra 4
Cabra 3
Cabra 6
Cabra 5
Cabra 2
Cabra 1
Cabra 7
Cabra 8


45
40
35
30
25
20
15
10
5
0
100
250
400
550
Cabra 4
Cabra 3
Cabra 2
Cabra 1
Cabra 5
Cabra 6
Cabra 7
Cabra 8


45
40
35
30
25
20
15
10
5
0
100
250
400
550
Cabra 4
Cabra 3
Cabra 2
Cabra 1
Cabra 5
Cabra 6
Cabra 7
Cabra 8

Dendograma