Multidimensional Scaling y Modelo Vectorial

OBJETIVOS Y PLANTEAMIENTO DEL PROBLEMA

En colaboración con el Departamento de Pediatría de la Universidad de Salamanca, se ha realizado un estudio en el que se pretende conocer la incidencia de enfermedades alérgicas en los niños de la provincia de Salamanca, considerando la **sintomatología**, el **diagnóstico** emitido por el médico y la **sensibilidad** a los distintos **alérgenos** en los distintos **centros de salud** de la provincia.
Los centros de salud considerados fueron:

SALAMANCA CAPITAL	*PROVINCIA*	
Centro-Universiad (UNI)	*Alba de Tormes (ALB)*	*Pedrosillo(PED)*
Sancti Spiritus-Canalejas (SA)	*Aldeadávila (ALD)*	*Peñaranda (PEN)*
Alamedilla (AL)	*Béjar (BEJ)*	*Periurbana Norte (PERN)*
Garrido Sur (GAS)	*Ciudad Rodrigo (CRO)*	*Periurbana Sur (PERS)*
Garrido Norte (GAN)	*Fuente de S. Estebam (FUE)*	*Santa Marta (SMAR)*
S. Juan (JUA)	*Guijuelo (GUI)*	*Tamames (TAM)*
Pizarrales (PIZ)	*La Alberca (ALBE)*	*Vitigudino (VIT)*
San Bernardo (BER)	*Ledesma (LED)*	*Villoria (VILL)*
Tejares (TEJ)	Lumbrales (LUMB)	
S. José (JOS)		

SÍNTOMAS	ALÉRGENOS
Rinitis	Polvo
Conjuntivitis	Ácaros
Dermatitis	Medicamentos
Gastroenteritis	Veneno de himenópteros
Tos espasmódica	Pólenes
Asma	Alimentos
Inmunodeficiencia	Hongos
Urticaria	Epitelios
Tos	
Reflujo	
Catarro habitual descendente (CHD)	

Los datos considerados originalmente fueron el número de pacientes que presentaba cada una de las categorías de la variable estudiada (sintomatología y sensibilidad), en cada uno de los centros de salud. Para evitar el efecto del tamaño de la población en los distintos centros de salud se consideró finalmente el porcentaje de pacientes que presentaban cada síntoma y la sensibilidad a cada uno de los alérgenos.

El propósito del estudio era la elaboración de un **mapa** de los centros de salud que reflejase el **parecido** de los mismos de acuerdo con el porcentaje de cada uno de los **síntomas** considerados, además se trataba de interpretar dicho mapa a partir de la información que producían los **alérgenos**.

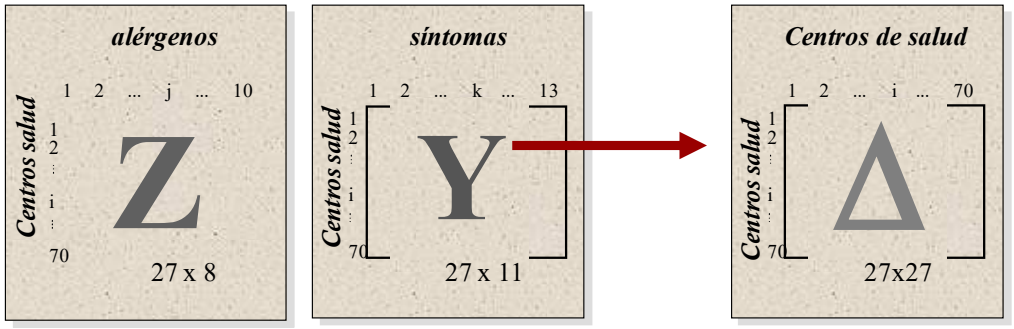
Como medida de la **similaridad** entre los distintos centros se consideró la **distancia euclídea** entre ellos teniendo en cuenta los distintos síntomas.

Para el estudio tiene las siguientes matrices:
En el archivo SINT_ALER.sav aparecen los centros de salud en filas y en columnas los síntomas y los alérgenos y ene l archivo SINTOMAS.sav contiene los porcentajes de cada uno de los síntomas considerados (éstos en filas y los centros de salud en columnas).

Esquema del trabajo práctico que vamos a desarrollar:

Esquema a seguir

Datos

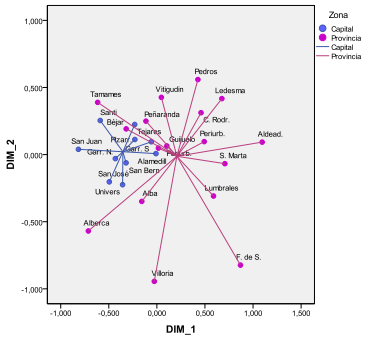


δ_{ij} = grado de parecido (distancia euclid

entre centros de salud

1 MDS

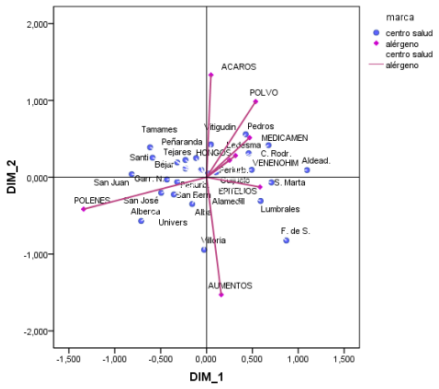
ZPOLVO	ZACAROS	ZHONGOS	ZPOI
-,16273	,40996	,67283	
,18625	,76461	-1,49693	
1,75202	,49196	-1,49693	
-1,17129	-2,78513	,31210	
1,67989	1,82006	2,16725	
-,78275	,17352	,01105	
-1,55866	-1,72010	-,14083	
1,64034	1,20873	-1,49693	
-,09991	,51540	,41246	
1,34953	,94247	1,21527	



2 Modelo vectorial

$$Polvo = \hat{\beta}_{1Polvo} Dim1 + \hat{\beta}_{2\,Polvo} Dim2 + \hat{\beta}_{3\,Polvo} Dim3$$

$$T = 0{,}535 Dim1 + 0{,}985 Dim2 + -0{,}613 Dim3$$



1.- MDS DE LA MATRIZ DE SINTOMAS

1.1.-Datos

Abra el archivo **SÍNTOMAS.sav**. En dicha matriz aparecen en filas los síntomas y en columnas los lugares (centros de salud). Es por tanto, una matriz 11x27.

1.2.- PROXCAL:

Vamos a realizar el análisis mediante la opción **PROXCAL** (porque nos permite guardar las coordenadas obtenidas en el análisis. (En el ALSCAL, las coordenadas se obtienen como resultado del análisis, pero hay que editarlas) (Figura 1).

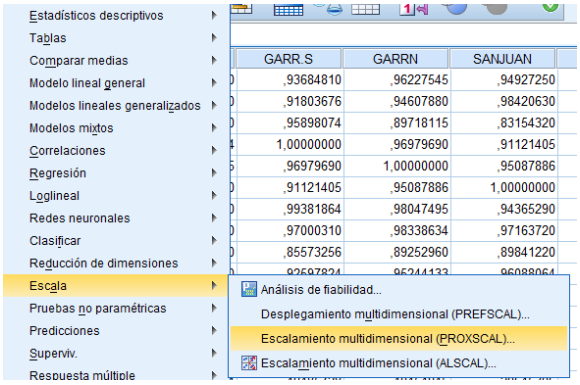


Figura 1

Tenemos que calcular las Similaridades por tanto, tenemos que **Crear proximidades a partir de los datos**, y **Una fuente matricial**. (Figura 2).

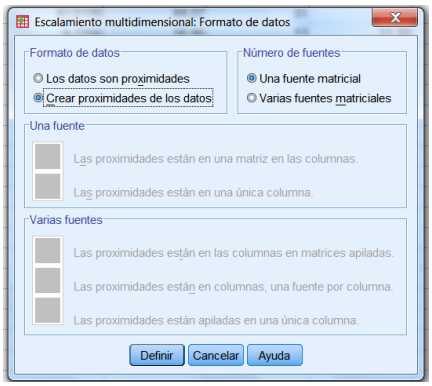


Figura 2

En primer lugar, le decimos al programa con qué columnas de datos vamos a trabajar. Para ello añadimos las "variables", en este caso son los diferentes centros de salud que se encuentran en las columnas de nuestra matriz de datos. (Figura 3).

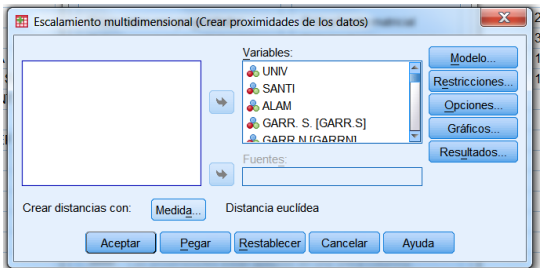


Figura 3

A continuación, en la opción **Medida**, seleccionaremos la medida de disimilaridad a calcular. En este caso, seleccionaremos **Distancia Euclídea**. Lo hacemos entre variables, y sin estandarizar. (Figura 4).

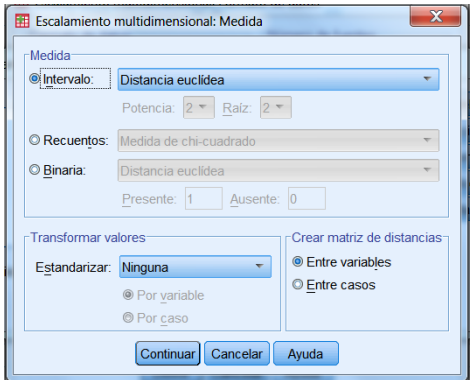


Figura 4

La opción que elegiremos en **Modelo** es un modelo de **Razón**, pues las distancias de partida son distancias euclídeas. Queremos la solución para **1-7** dimensiones, (para poder luego decidir con cuántas realizamos el análisis). (Figura 5).

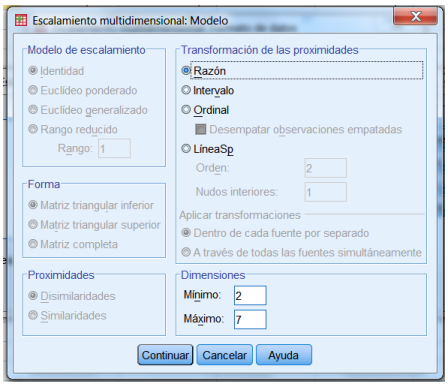


Figura 5

Elija en **Opciones** como configuración inicial la obtenida mediante el método de **TORGERSON**. (Figura 6).

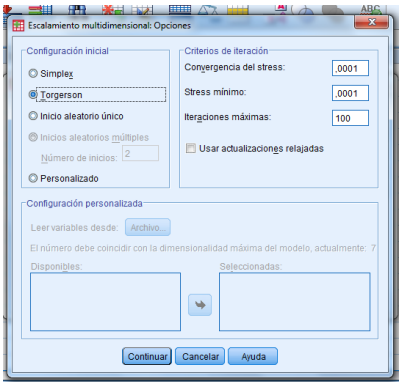


Figura 6

En cuanto a los gráficos que queremos nos muestre en los resultados, podemos marcar las siguientes opciones, para que nos muestre el gráfico para los centros de salud (**espacio común**), así como los gráficos de proximidades originales frente a transformadas (disparidades) y el de proximidades transformadas frente a distancias, para ver cómo es el ajuste. También queremos que nos dibuje el ***Scree Plot*** o diagrama del codo, para el **Stress**, y así poder decidir en relación a en cuántas dimensiones deberemos presentar la solución. Representa el **Stress Bruto Normalizado** (Figura 7 y 8).

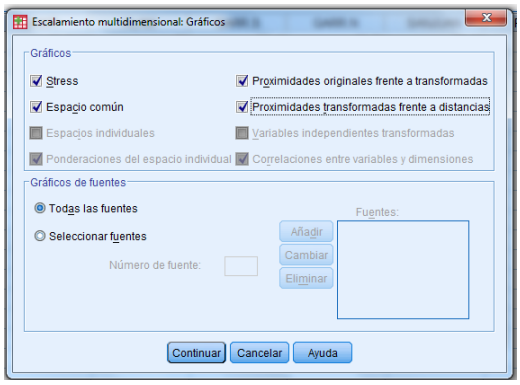
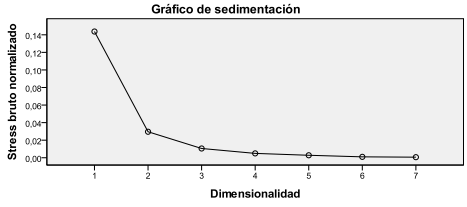


Figura 7

Realizamos el análisis y observamos el diagrama de sedimentación. A continuación se presenta esta figura.

Figura 8

La otra opción, si no se quiere representar esta medida, sino otra medida de bondad / falta de ajuste, es editar la tabla siguiente que aparecerá en la ventana de resultados y pegarla en Excel, y realizar allí la representación deseada (Figuras 9, 10, 11 y 12).



Bondad de ajuste

Medidas de ajuste y stress

Dimensionalidad:2

Stress bruto normalizado	,02960
Stress-I	,17206[a]
Stress-II	,36982[a]
S-Stress	,06172[b]
Dispersión explicada (D. A.F.)	,97040
Coeficiente de congruencia de Tucker	,98509

PROXSCAL minimiza el stress bruto normalizado.

a. Factor para escalamiento óptimo = 1,031.
b. Factor para escalamiento óptimo = ,947.

Figura 9

STRESS BRUTO NORMALIZADO

STRESS-I $S_1 = \left[\frac{\sum_{i,j}(\hat{\delta}_{ij} - d_{ij})^2}{\sum_{i,j} d_{ij}^2}\right]^{1/2}$

STRESS-II $S_2 = \left[\frac{\sum_{i,j}\left(\hat{\delta}_{ij} - d_{ij}\right)^2}{\sum_{i,j}\left(d_{ij} - d_{\cdot\cdot}\right)^2}\right]^{1/2}$

S-STRESS $SS_1 = \left[\frac{\sum_{i,j}(\hat{\delta}_{ij}^2 - d_{ij}^2)^2}{\sum_{i,j}(d_{ij}^2)^2}\right]^{1/2}$

DISPERSIÓN EXPLICADA (D.A.F.)

COEFICIENTE DE CONGRUENCIA DE

Figura 10

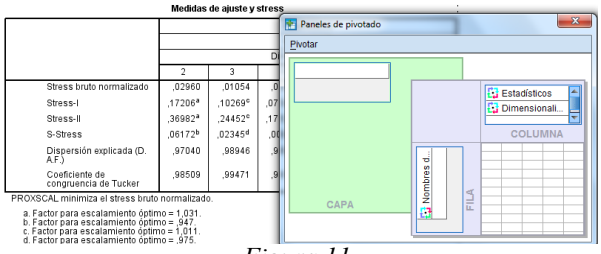


Figura 11

Medidas de ajuste y stress

	Dimensionalidad						
	1	2	3	4	5	6	7
Stress bruto normalizado	,14374	,02960	,01054	,00496	,00281	,00102	,00064
Stress-I	,37914[m]	,17206[k]	,10269[i]	,07042[g]	,05305[e]	,03200[c]	,02535[a]
Stress-II	,65524[m]	,36982[k]	,24452[i]	,17627[g]	,13594[e]	,08355[c]	,06668[a]
S-Stress	,23513[n]	,06172[l]	,02345[j]	,00992[h]	,00491[f]	,00166[d]	,00070[b]
Dispersión explicada (D. A.F.)	,85626	,97040	,98946	,99504	,99719	,99898	,99936
Coeficiente de congruencia de Tucker	,92534	,98509	,99471	,99752	,99859	,99949	,99968

PROXSCAL minimiza el stress bruto normalizado

Figura 12

Esta tabla la copiamos y la llevamos a Excel, donde podremos realizar otras representaciones en relación a la bondad de ajuste o falta de ajuste (Por ejemplo, Figura 13).

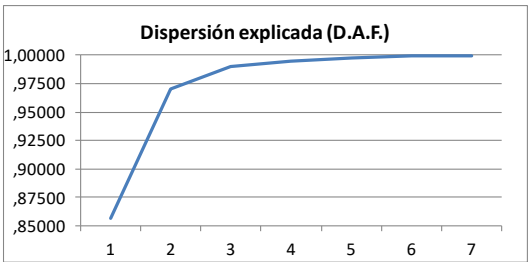


Figura 13

Como podemos observar el codo se produce en la segunda/tercera dimensión, si bien cuando observamos los valores de stress, vemos que con 2 dimensiones obtenemos un *stress bruto normalizado* del 3% que en los términos establecidos por Kruskal sería nos proporciona un buen valor, del mismo modo que si observamos la *Dispersión explicada* en dos dimensiones es del 97%.

Sin embargo, a efectos de la práctica, vamos a realizar la solución en tres dimensiones.

1.3.- Solución con tres dimensiones:

Una vez que hemos visto que con 3 dimensiones es suficiente, vamos a realizar el análisis con 3 dimensiones.

Volvemos al menú **Analizar → Escalas → Proxcal → Definir → Modelo**. Ahí seleccionamos de 2-3 dimensiones. Y luego sólo nos queda entrar en el botón **Resultados**: ahí podemos elegir información adicional para mostrar en la ventana de resultados (como por ejemplo que nos muestre la matriz de distancias, etc.), pero lo que es más importante, **nos permite guardar en un fichero las coordenadas del espacio común** (es decir, la de los lugares) para poder utilizarlas en ulteriores análisis (esta opción deberemos marcarla porque luego las vamos a utilizar). Al fichero lo llamamos: **COORDCENTROSALUD**. (Figura 14).

1.4.- Resultados e interpretación:

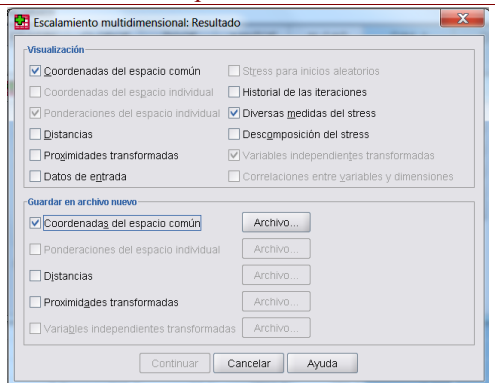


*Figura 14 (ventana en la versión PASW)**

Las medidas de ajuste están en la tabla siguiente (Figura 15).

Medidas de ajuste y stress

Stress bruto normalizado	,01054
Stress-I	,10267[a]
Stress-II	,24436[a]
S-Stress	,02339[b]
Dispersión explicada (D. A.F.)	,98946
Coeficiente de congruencia de Tucker	,99472

PROXSCAL minimiza el stress bruto normalizado.

a. Factor para escalamiento óptimo = 1,011.
b. Factor para escalamiento óptimo = ,975.

Figura 15

Luego tenemos la tabla donde nos muestra las coordenadas de los centros de salud en las tres dimensiones en que hemos pedido la solución.(Figura 16):

Espacio común

Coordenadas finales

Dimensionalidad:3

	Dimensión		
	1	2	3
UNIV	-,356	-,224	,199
SANTI	-,589	,254	-,124
ALAM	-,009	,007	,351
GARR. S.	-,058	,095	,030
GARR N	-,432	-,030	,019
SAN JUAN	-,816	,039	-,111
PIZAR	-,229	,113	,036
SAN BERN	-,320	-,061	-,014
TEJARES	-,230	,224	-,477
SAN JOSE	-,495	-,203	,174
BEJAR	-,319	,192	-,037

Figura 16

A continuación aparece el diagrama de Shepard. (Figuras 17 y 18) (distancia transformada=disparidad).

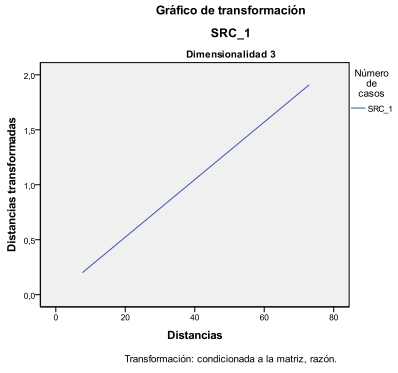


Figura 17

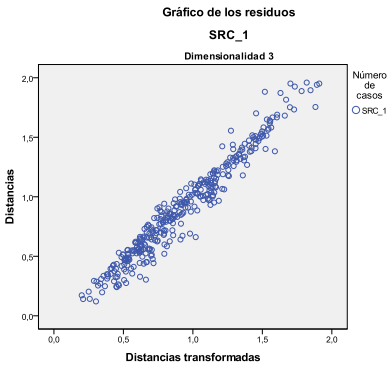


Figura 18

1.3.- Representación gráfica de los centros de salud para ponerles las marcas:

El programa nos muestra el mapa perceptual combinando las tres dimensiones (en forma de matriz de gráficos). (Figuras 19).

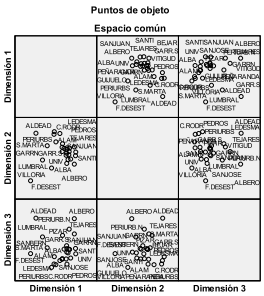


Figura 19

(*)Abrimos el fichero **COORDCENTROSALUD** que contiene tres columnas denominadas DIM_1; DIM_2 y DIM_3, que son las coordenadas de cada centro en cada una de las dimensiones de la solución. Copiamos los nombres de los centros de Salud (que aparecen en la ***vista de variables*** del archivo síntomas) y los pegamos en una nueva columna a continuación de las dimensiones.

Construimos una nueva variable, que llamaremos **ZONA**, con valores ***1=Capital*** y ***2=Provincia***, y le ponemos el valor correspondiente a cada Centro de Salud dependiendo de la zona a la que pertenezcan (Figura 20)

.

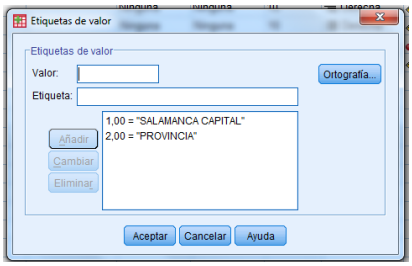


Figura 20

Ahora nos vamos a la opción ***Cuadros de diálogo antiguos---dispersión/puntos*** (para hacer el diagrama). (Figura 21)

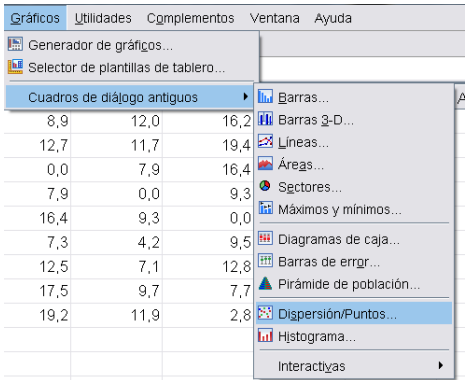


Figura 21

Y elegimos el gráfico *"**Dispersión simple**"*. (Figura 22).

Figura 22

Queremos que las dimensión 1 se represente en el eje X y la dimensión 2 en el eje Y, por tanto añadimos las variables en los lugares correspondientes. (Figura 23 y 24)

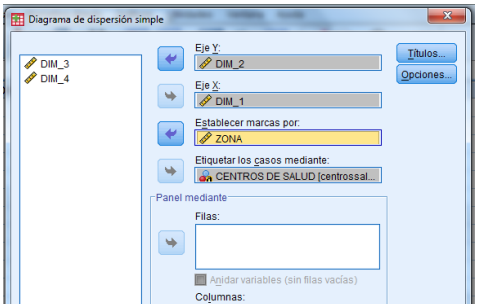


Figura 23

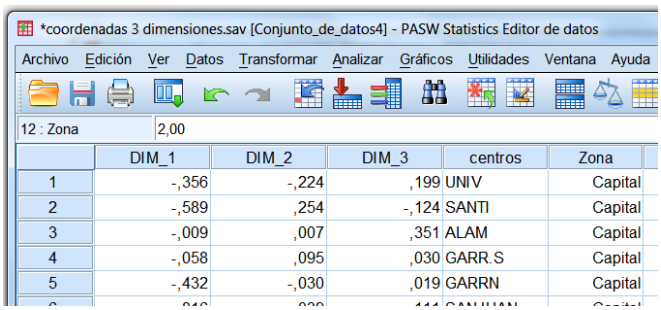


Figura 24

Queremos que cada punto venga identificado, por eso añadimos la variable CENTROS DE SALUD en la **casilla *"etiquetar los casos mediante"***, y luego queremos que marque con un símbolo y/o color diferente los lugares dependiendo de a qué ZONA pertenecen. Por tanto, añadimos la variable ZONA, en la casilla de ***"establecer marcas por".*** Por último hacemos clic en el botón ***Opciones*** y marcamos la opción ***"Mostrar el gráfico con las etiquetas de caso",*** para que efectivamente le ponga las etiquetas a los lugares. (Figura 25)

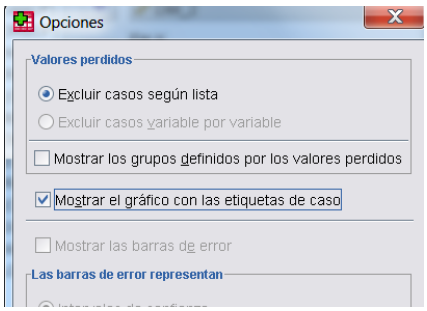


Figura 25

1.4.- Resultados e interpretación:

El resultado será el que aparece a continuación (Figuras 26, 27 y 28). Se han representado todos los planos hasta la dimensión 3. *Hemos unido cada centro de salud con el centro de gravedad de su zona para visualizar mejor los resultados*.

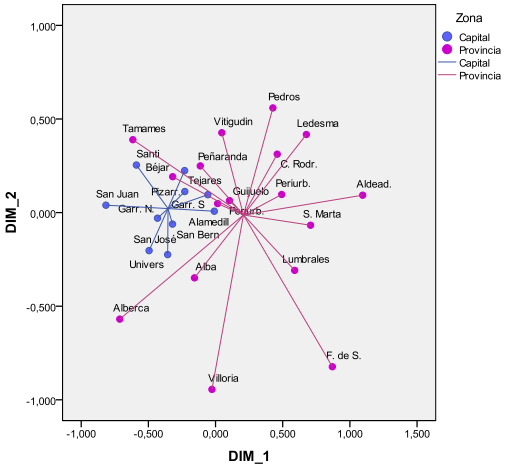


Figura 26

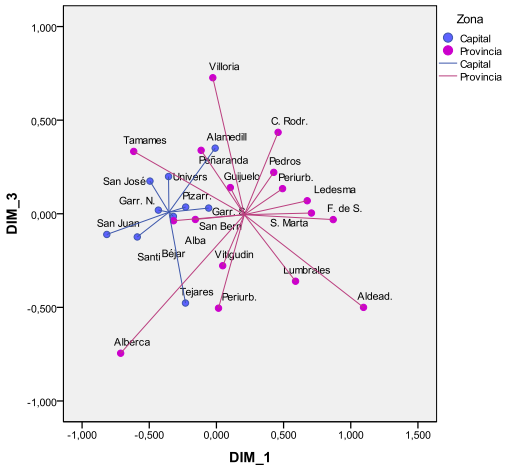


Figura 27

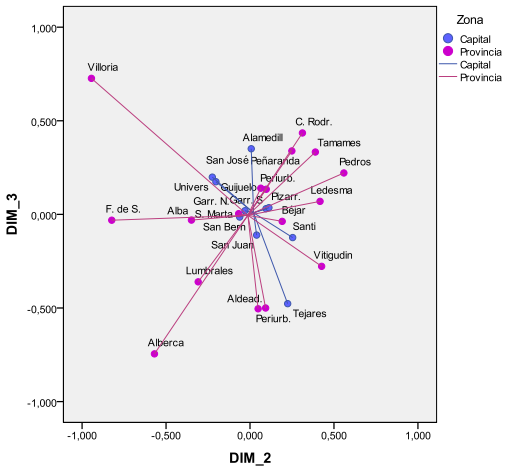


Figura 28

2.- MODELO VECTORIAL.

Vamos a encontrar las direcciones del espacio (del mapa perceptual) máximamente relacionadas con los alérgenos), es decir, la dirección de máximo incremento para cada uno de los alérgenos. Eso lo vamos a conseguir mediante regresión múltiple, donde cada alérgeno será la variable dependiente y cada dimensión la independiente. El coeficiente de regresión para cada dimensión nos dará las coordenadas de la punta del vector que represente a cada variable (alérgeno).

2.1.- Abrimos el archivo

SINT_ALERG.sav. Lo primero que vamos a hacer es a estandarizar los datos. Para ello, en el menú **Analizar → Estadísticos descriptivos → Descriptivos**, señalamos la opción: ***Guardar valores tipificados como variables***. De este modo, a continuación, en el archivo, nos

aparecen las correspondientes variables estandarizadas (Figuras 29, 30 y 31).

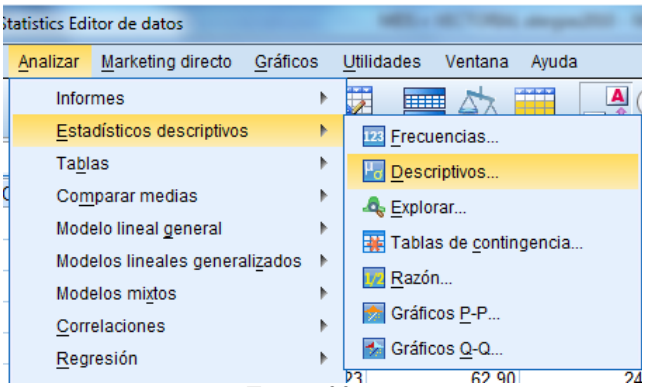


Figura 29

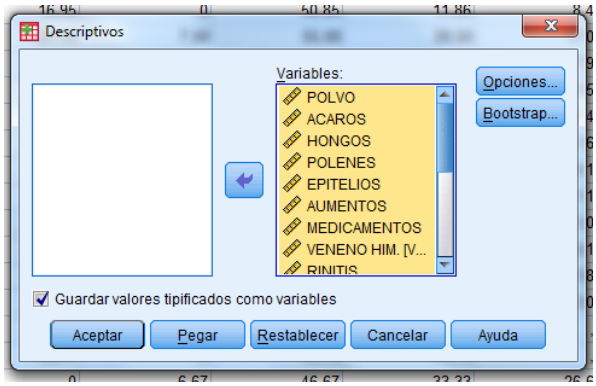


Figura 30

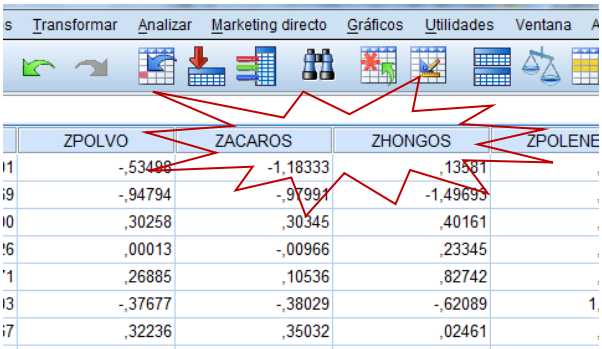


Figura 31

2.3.- Copiamos las tres columnas de las coordenadas en las tres dimensiones en este fichero de datos de los alérgenos/síntomas.

2.4.- Vamos a encontrar las coordenadas de las puntas del vector que se corresponde a cada variable (alérgeno). Eso lo hacemos ajustando un modelo de regresión múltiple para cada variable (alérgeno) (que juega el papel de dependiente) sobre las coordenadas de las tres dimensiones (consideradas como variables independientes).

El ajuste lo realizaremos para todas las variables a la vez. Las coordenadas de la punta de la flecha de cada vector para cada variable serán los coeficientes de regresión obtenidos para cada variable sobre cada dimensión.

Lo vamos a hacer en la opción ***Modelo lineal general*** y allí elegimos ***Multivariante*** (Figura 3).

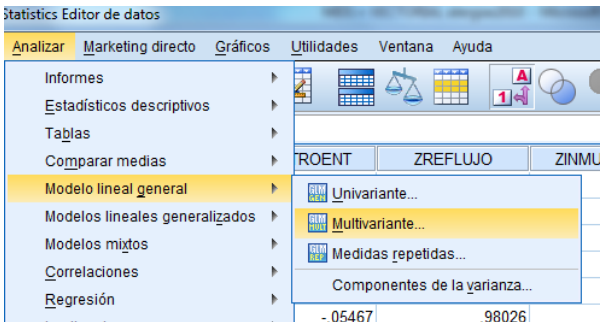


Figura 32

A las variables (alérgenos) les asignamos el papel de variables dependientes, y como covariables elegiremos las dimensiones. (DIM_1, DI_2 y DIM_3) (Figura 3).

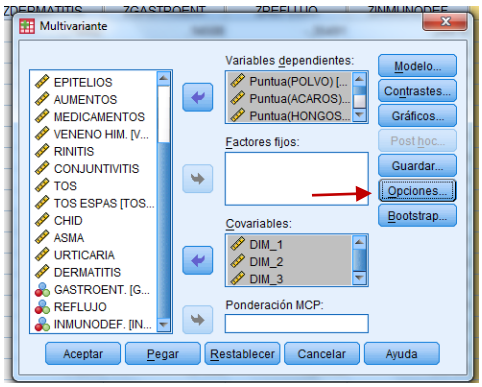


Figura 33

En el botón ***opciones*** tenemos que seleccionar que nos muestre las ***estimaciones de los parámetros*** (Figura).

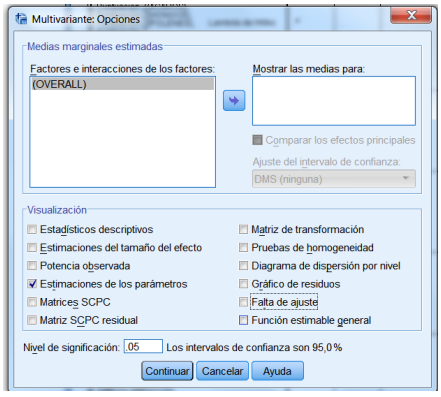


Figura 34

Continuamos.

2.5.- En los resultamos editamos (doble clic sobre) la tabla de estimaciones de parámetros, y seleccionamos, en el menú ***Pivotar*** la opción ***Paneles de pivotado***. Esto es simplemente para poder colocar en la tabla los resultados de forma que podamos copiar de forma sencilla los coeficientes de regresión parciales de cada variable dependiente (variables ambientales) sobre cada una de las dimensiones (v. independientes). (Figura 3).

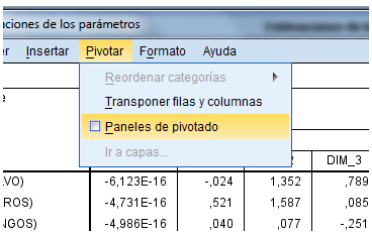


Figura 35

En la siguiente figura, que es la correspondiente a la ventana ***Paneles de Pivotado***, dejamos las opciones como se muestra en la figura. (Figura).

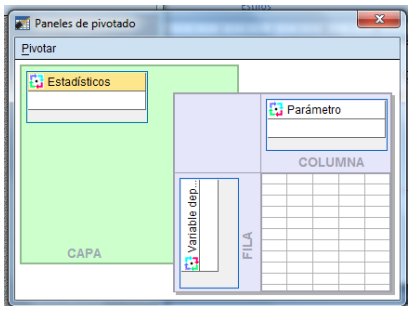


Figura 36

Le debe quedar como aparece en la Figura .

De esta forma, observe cómo le quedan en tres columnas los coeficientes de regresión parcial de cada variable para cada dimensión (señalados con un círculo en la figura anterior).

Estimaciones de los parámetros

Estadísticos: B

Variable dependiente	Parámetro			
	Intersección	DIM_1	DIM_2	DIM_3
Puntua(POLVO)	-6,399E-16	0,535	0,985	-0,613
Puntua(ACAROS)	-4,024E-16	0,048	1,331	-0,036
Puntua(HONGOS)	-4,952E-16	0,251	0,221	0,079
Puntua(POLENES)	5,411E-16	-1,340	-0,415	-0,698
Puntua(EPITELIOS)	6,406E-18	0,582	-0,127	-0,522
Puntua(AUMENTOS)	1,120E-16	0,160	-1,528	1,924
Puntua(MEDICAMENTOS)	1,683E-16	0,469	0,515	-0,735
Puntua: VENENO HIM.	1,559E-16	0,315	0,282	0,283

Figura 37

2.6.- Copiamos los resultados para DIM_1, DIM_2 y DIM_3, (es decir, los coeficientes de regresión parciales de cada variable ambiental para cada una de las dimensiones, que ahora funcionan como variables independientes). *(Como las DIM_i son incorreladas, los coeficientes de regresión no cambian aunque ajustemos el modelo para cada plano por separado, o bien para las tres dimensiones de forma conjunta).*

2.7 *Los pegamos en el archivo COORDCENTROSALUD, debajo de las columnas DIM_1, DIM_2 y DIM_3 (colocándolas, por tanto, ***debajo de las coordenadas de los centros de salud para cada dimensión***).*

2.8.- Por último ponemos los nombres de los alérgenos para que aparezcan como etiquetas en el gráfico correspondiente (los podemos copiar del archivo alérgenos, si vamos a vista de variables).

2.9.- En otra columna que llamaremos MARCA, diferenciamos las filas correspondientes a las coordenadas de los centros de salud y las correspondientes a las coordenadas de los alérgenos (por ejemplo CENTROS y ALERGENOS) y vamos a construir el gráfico (análogamente al caso que vimos anteriormente). Al tener colocadas las coordenadas para situar las puntas de las flechas de las variables, debajo de las coordenadas para los lugares, quedan automáticamente representadas. Ahora, establecemos marcas por la variable MARCA que hemos construido.

2.10.- Dibujamos el gráfico (en la misma opción en que dibujamos el gráfico en el primer apartado de MDS). Elegimos qué queremos colocar en el Eje Y (DIM_2) en el Eje X (DIM_1) y establecemos marcas (la variable MARCA) y etiquetamos los casos mediante Centros Salud. (No olvidarse: en el botón inferior **Opciones** debemos decirle: ***Mostrar el gráfico con las etiquetas de caso)***

Se obtiene el gráfico que se puede editar para "arreglarlo visualmente", y poder exportarlo, por ejemplo al PowerPoint para poner los vectores que representan a las variables.

Repetimos para las demás combinaciones de dimensiones. Hay que tener en cuenta la significación de los coeficientes de regresión de cada variable (alérgeno) sobre cada dimensión para interpretación, o bien, interpretar la bondad de ajuste (Figuras 38.1 y 38.2)

Estimaciones de los parámetros

Estadísticos= Sig.

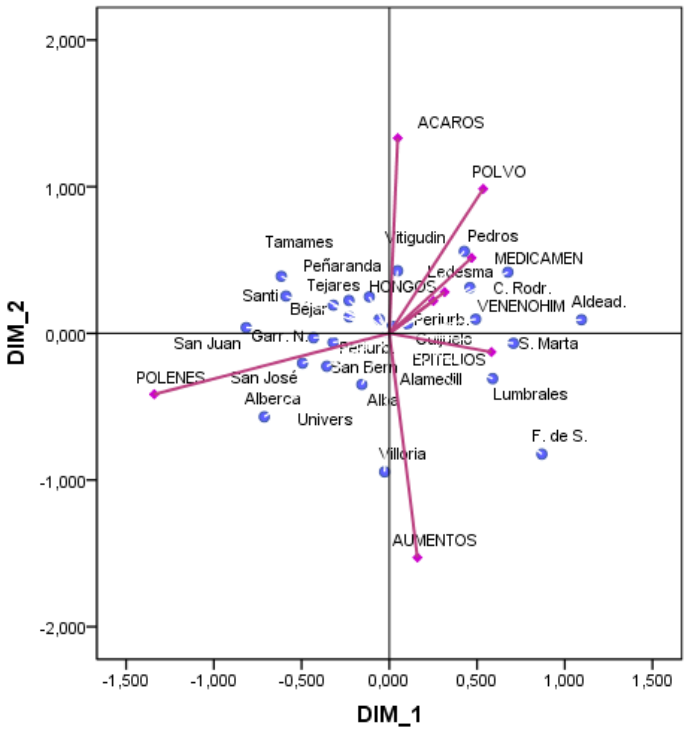
Variable dependiente	Parámetro			
	Intersección	DIM_1	DIM_2	DIM_3
Puntua(POLVO)	1,000	,148	,063	,283
Puntua(ACAROS)	1,000	,896	,015	,949
Puntua(HONGOS)	1,000	,542	,702	,902
Puntua(POLENES)	1,000	,000	,301	,122
Puntua(EPITELIOS)	1,000	,144	,817	,395
Puntua(AUMENTOS)	1,000	,481	,000	,000
Puntua(MEDICAMENTOS)	1,000	,228	,344	,226
Puntua: VENENO HIM.	1,000	,440	,622	,655

Figura 38.1

Estimaciones de los parámetros				
Estadísticos= Sig.				
Variable dependiente		Parámetro		
		DIM_1	DIM_2	DIM_3
	POLVO	,148	,063	,283
	ACAROS	,896	,015	,949
	HONGOS	,542	,702	,902
	POLENES	,000	,301	,122
	EPITELIOS	,144	,817	,395
	AUMENTOS	,481	,000	,000
	MEDICAMENTOS	,228	,344	,226
	VENENO HIM.	,440	,622	,655

Figura 38.2

2.11.- Resultados e Interpretación (Figura 40 y 41)

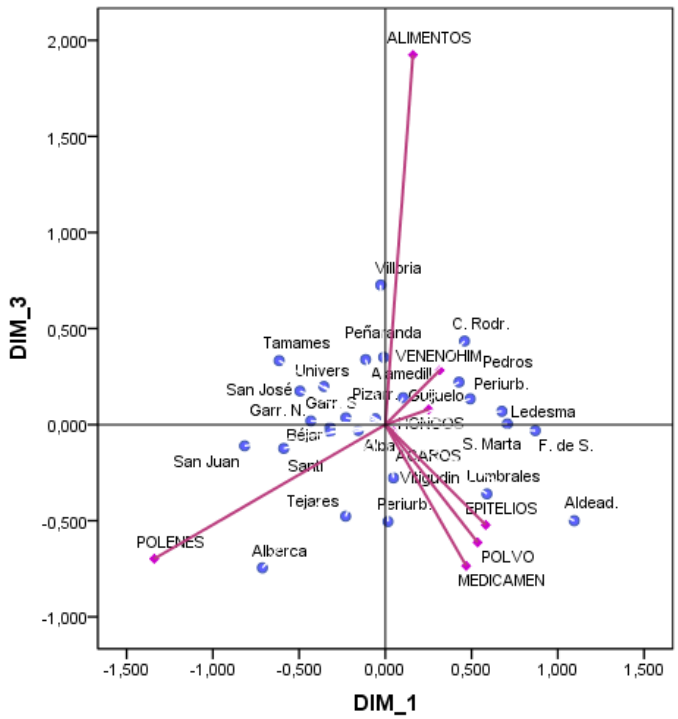


marca
- centro salud
- alérgeno

centro salud
alérgeno

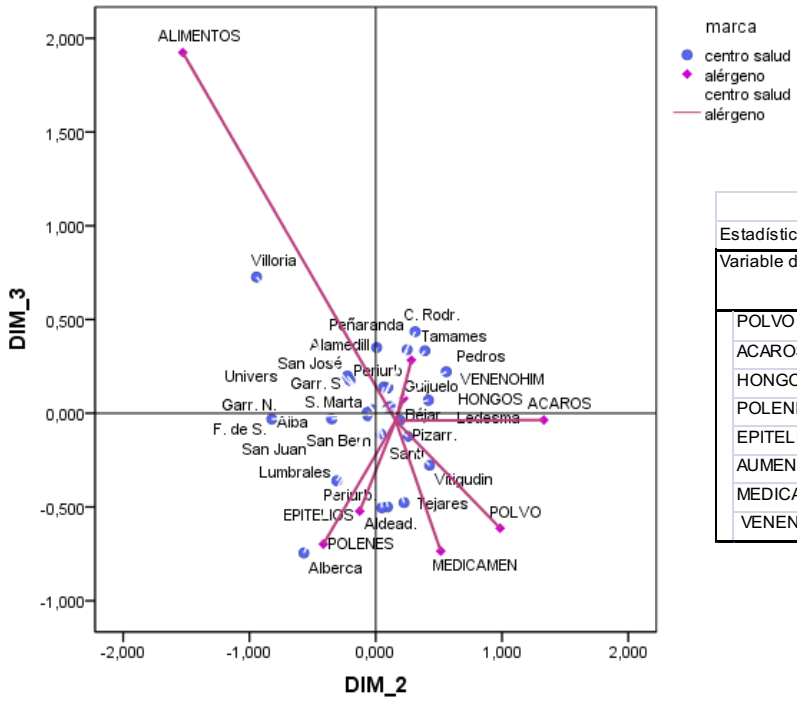
Estimaciones de los parámetros				
Estadísticos= Sig.				
Variable dependiente		Parámetro		
		DIM_1	DIM_2	DIM_3
	POLVO	,148	,063	,283
	ACAROS	,896	,015	,949
	HONGOS	,542	,702	,902
	POLENES	,000	,301	,122
	EPITELIOS	,144	,817	,395
	AUMENTOS	,481	,000	,000
	MEDICAMENTOS	,228	,344	,226
	VENENO HIM.	,440	,622	,655

Figura 39



Estimaciones de los parámetros			
Estadísticos= Sig.			
Variable dependiente	Parámetro		
	DIM_1	DIM_2	DIM_3
POLVO	,148	,063	,283
ACAROS	,896	,015	,949
HONGOS	,542	,702	,902
POLENES	,000	,301	,122
EPITELIOS	,144	,817	,395
AUMENTOS	,481	,000	,000
MEDICAMENTOS	,228	,344	,226
VENENO HIM.	,440	,622	,655

Figura 40



Estimaciones de los parámetros			
Estadísticos= Sig.			
Variable dependiente	Parámetro		
	DIM_1	DIM_2	DIM_3
POLVO	,148	,063	,283
ACAROS	,896	,015	,949
HONGOS	,542	,702	,902
POLENES	,000	,301	,122
EPITELIOS	,144	,817	,395
AUMENTOS	,481	,000	,000
MEDICAMENTOS	,228	,344	,226
VENENO HIM.	,440	,622	,655

Figura 41

3.- REPETIR EL MODELO VECTORIAL, INCLUYENDO TAMBIÉN LOS SÍNTOMAS