Análisis de Conglomerados

Resumen

El procedimiento **Análisis de Conglomerados** esta diseñado para agrupar observaciones o variables en conglomerados basados en similaridades entre ellos. Los renglones de los datos para el procedimiento pueden estar en cualquiera de las dos formas:

1. *n* renglones o casos, cada uno conteniendo los valores de las *p* variables cuantitativas.
2. *n* renglones y *n* columnas si se conglomera a las observaciones o *p* renglones y *p* columnas si se conglomera a las variables, conteniendo una medida de "distancia" entre todos los pares de objetos.

Si un renglón de datos es la entrada, el procedimiento calculara las distancias entre las observaciones o variables.

Un número de algoritmos son dados para generar conglomerados. Algunos de estos son aglomerativos, empezando con conglomerados separados para cada observación o variable y uniéndolos de acuerdo a sus similaridades. Otros métodos empiezan con un conjunto de semillas y van uniendo otras observaciones o variables a cada semilla para formar conglomerados.

Los resultados del análisis son desplegados de distintas maneras, incluyendo un dendograma, una tabla de miembros, y una grafica icicle.

Ejemplo StatFolio: *cluster.sgp*

Datos del Ejemplo:

El archivo *cities.sf6* contiene información de *n* = 10 ciudades grandes de U.S., obtenidas de www.city-data.com. Los datos consisten de variables demográficas, económicas y ambientales. La siguiente tabla muestra una lista parcial de los datos en este archivo:

City (Ciudad)	*Population (Población)*	*Percent Female (Porcentaje Femenino)*	*Land Area (Área de Tierra)*	*Median Age (Edad mediana)*	*Median Income (Renta mediana)*	*Highest temperature (Máxima temperatura)*
New York	8008278	52.6	303.3	34.2	38293	76.9
Boston	589141	51.9	48.4	31.1	39629	72.9
Chicago	2896016	51.5	227.1	31.5	38625	74.7
Washington	572059	52.9	61.4	34.6	40127	78.2
Atlanta	416474	50.4	131.7	31.9	34770	79.7
Los Angeles	3694820	50.2	469.1	31.6	36687	72.0
San Francisco	776733	49.2	46.7	36.5	55221	63.7
Miami	362470	50.3	35.7	37.7	23483	84.3
Houston	1953631	50.1	579.4	30.9	36616	53.0
Phoenix	1321045	49.1	474.9	30.7	41207	90.7

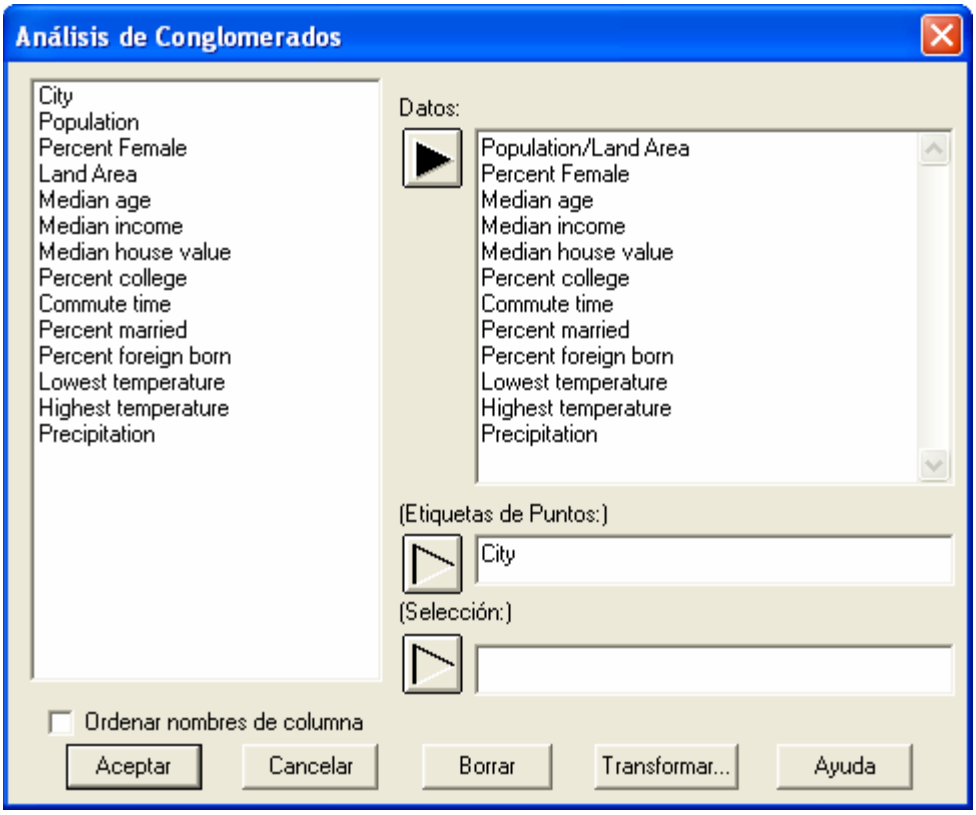
Las ciudades serán conglomeradas de acuerdo a las siguientes *p* = 12 variables:

© 2006 por StatPoint, Inc.

Population/Land Area
Percent Female
Median age
Median income
Median house value
Percent college
Commute time
Percent married
Percent foreign born
Lowest temperature
Highest temperature
Precipitation

Entrada de Datos

La caja de dialogo de datos de entrada requiere los nombres de las columnas que contienen los datos de entrada:



- **Datos:** Si las observaciones son conglomeradas, los nombres de las *p* variables de entrada contienen los valores para los *n* casos, o una matriz de *n* por *n* contiene las distancias entre cada par de casos. Si las variables son conglomeradas, los nombres de las *p* variables de estrada contienen los valores para los *n* casos, o una matriz de *p* por *p* contiene las distancias entre cada par de variables.

- **Etiquetas de Puntos:** Etiquetas opcionales para cada renglón en la hoja de datos.

© 2006 por StatPoint, Inc.

- **Selección:** Selección de un subconjunto de los datos.

Resumen del Análisis

El *Resumen del Análisis* resume los resultados de la conglomeración.

Análisis de Conglomerados

Datos/Variables:
- Population/Land Area
- Percent Female
- Median age
- Median income (household)
- Median house value ($1,000s)
- Percent college
- Commute time (minutes)
- Percent married
- Percent foreign born
- Lowest temperature (average in month)
- Highest temperature (average in month)
- Precipitation (highest month)

Número de casos completos: 10
Método de Conglomeración: Vecino Más Lejano (Vínculo Completo)
Métrica de Distancia: Euclideana Cuadrada
Conglomeración: observaciones
Estandarizar: sí

Resumen de Conglomeración

Conglomerado	*Miembros*	*Porcentaje*
1	10	100.00

Centroides

Conglomerado	*Population/Land Area*	*Percent Female*	*Median age*	*Median income*	*Median house value*
1	10462.3	50.82	33.07	38465.8	175.27

Conglomerado	*Percent college*	*Commute time*	*Percent married*	*Percent foreign born*	*Lowest temperature*
1	29.86	30.39	40.43	28.6	47.55

Conglomerado	*Highest temperature*	*Precipitation*
1	74.61	4.83

Incluidas en la tabla están:

- **Variables de Entrada**: Identificación de las variables de entrada.

- **Numero de casos completos**: El número de casos *n* con información sobre todas las variables de entrada. Cualquier renglón en la hoja de datos con valores perdidos para alguna variable son excluidos del análisis.

- **Método de Aglomeración**: El método usado para derivar la conglomeración (ver discusión abajo).

- **Métrica Distancia**: Si los datos consisten de observaciones, la métrica usada para medir la distancia entre los conglomerados. Si la matriz de distancias ha sido, indicada por el usuario.

© 2006 por StatPoint, Inc.

- **Conglomeración**: Cualquiera *observaciones* o *variables*, dependiendo de acuerdo a que se requiera la conglomeración.

- **Estandarizados**: Si los datos fueron estandarizados antes de que las distancias fueran calculadas.

- **Resumen de Conglomeración**: El numero de conglomerados creados y el porcentaje de observaciones o variables puestas en cada conglomerado.

- **Centroides**: El valor promedio para cada variable en cada conglomerado (si las *observaciones* han sido conglomeradas).

Opciones del Análisis

- **Método**: Método usado para crear los conglomerados.

- **Numero de Conglomerados**: El numero final deseado de conglomerados.

- **Métrica Distancia** La métrica usada para medir la distancia entre los casos.

- **Conglomerar**: Si genera conglomerados para observaciones o variables.

- **Estandarizar**: Si se selecciona, las variables serán estandarizadas antes de hacer la conglomeración. Si se conglomeran observaciones, cada variable es estandarizada sustrayendo su media muestral y dividiendo por su desviación estándar muestral. Si se conglomeran variables, la conglomeración se basa en la matriz de correlaciones muestrales en lugar de en la matriz de covarianzas muestrales.

- **Semilla**: Cuando usamos el método de *k-medias*, se muestra una caja de dialogo para introducir las k semillas.

© 2006 por StatPoint, Inc.

Metodología Estadística

Con el objetivo de crear conglomerados de observaciones o variables, es importante tener una medida de" cercanía" o "similaridad" tal que los objetos parecidos puedan ser juntados. Cuando observaciones son conglomeradas, la cercanía es típicamente medida por la distancia entre observaciones en el *p*-dimensional espacio de variables. El procedimiento *Análisis de Conglomerados* provee 3 diferentes métricas para medir la distancia entre 2 objetos, representados por *x* y *y*:

1. Distancia Euclidiana Cuadrada: $d(x,y) = \sum_{i=1}^{p} (x_i - y_i)^2$ (1)

2. Distancia Euclidiana : $d(x,y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$ (2)

3. Distancia de bloque ciudad : $d(x,y) = \sum_{i=1}^{p} |x_i - y_i|$ (3)

Cuando se conglomeran variables, la distancia es definida similarmente excepto que *x* y *y* representan la localización de 2 variables en el *n*-dimensional espacio de las observaciones, y la suma es sobre las observaciones en lugar de sobre las variables.

Si alguna otra métrica de distancia es preferida, el usuario puede introducir la matriz de distancias directamente en lugar de introducir las observaciones originales.

Hay dos tipos básicos de métodos para conglomerar objetos:

1. *Métodos Jerárquicos Aglomerativos*: Métodos de conglomeración jerárquicos aglomerativos inician poniendo cada observación en un conglomerado separado. Conglomerados son unidos, dos cada vez, hasta que el número de conglomerados es reducido a un objetivo deseado. En cada etapa, los conglomerados son unidos en pares de acuerdo a su cercanía.

2. *Método de k-Medias*: Este método inicia identificando *k* objetos como semillas iniciales para cada conglomerado. Los objetos son adheridos a el conglomerado más cercano.

<u>Métodos Aglomerativos</u>

Los métodos aglomerativos inician poniendo cada objeto en un conglomerado separado y después combinando conglomerados de acuerdo a sus distancias con todos los demás. El proceso continúa hasta que el número deseado de conglomerados es alcanzado. Donde los métodos difieren es en como estos definen la distancia entre dos conglomerados cuando uno o ambos de los conglomerados contienen más que un miembro:

© 2006 por StatPoint, Inc.

1. *Vecino mas cercano (liga simple)*: Define la distancia entre 2 conglomerados como el mínimo de las distancias entre cualquier miembro de un conglomerado con cualquier miembro del otro conglomerado.

2. *Vecino mas lejano (liga compuesta)*: Define la distancia entre 2 conglomerados como el máximo de las distancias entre cualquier miembro de un conglomerado con cualquier miembro del otro conglomerado.

3. *Centroide*: Define la distancia entre 2 conglomerados como la distancia entre los centroides de cada conglomerado, donde el centroide es localizado en el valor promedio de cada variable sobre todos los miembros del conglomerado.

4. *Mediana*: Define la distancia entre 2 conglomerados como la distancia entre las medianas de cada conglomerado, donde la mediana es localizada en al valor mediano de cada variable sobre todos los miembros del conglomerado.

5. *Promedio de Grupo (liga promedio)*: Define la distancia entre 2 conglomerados como la distancia promedio entre todos los miembros de un conglomerado a todos los miembros del otro.

6. *Método de Ward*: Define la distancia entre 2 conglomerados en términos del incremento en la suma de las desviaciones cuadradas alrededor de la media del conglomerado que ocurriría si los dos conglomerados estuvieran unidos.

Método de k-Medias

El método de *k-medias* trabaja como sigue:

1. *k* objetos son seleccionados para ser las semillas iniciales (*semillas)* para los *k* conglomerados deseados.

2. Los restantes objetos son asignados a el conglomerado cuya semilla este mas cercana de estos.

3. Los centroides de cada conglomerado son calculados.

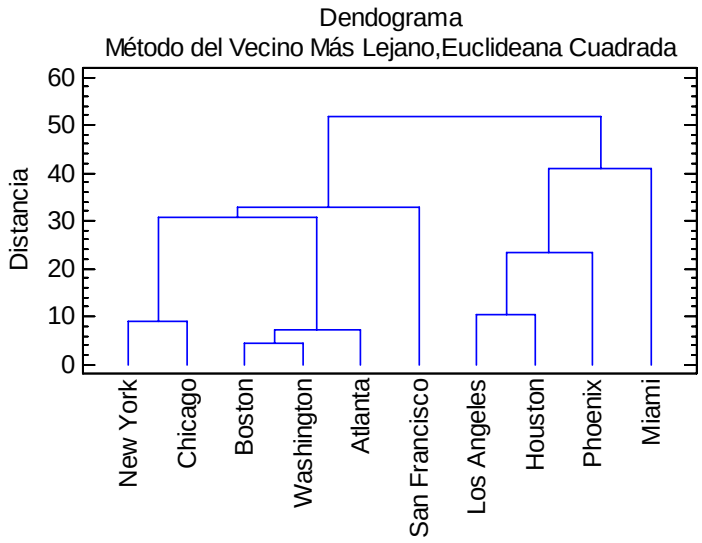
4. Cada objeto es revisado para determinar si es más cercano al centroide de otro que al centroide del conglomerado que esta actualmente asignado. Si es así este se asigna al otro y ambos centroides son recalculados.

5. El paso 4 es repetido hasta que no hay cambios de lugar.

© 2006 por StatPoint, Inc.

Dendograma

El mejor modo para ver la salida del análisis de conglomerados es usualmente un *Dendograma*:



Trabajando con este dendograma muestra la sucesión de uniones que fueron hechas entre conglomerados. Líneas son dibujadas conectando las conglomeraciones unidas en cada paso, mientras que el eje vertical muestra las distancias a las que fueron unidos los conglomerados.

Por ejemplo, el dendograma anterior muestra el resultado de conglomerar las $n = 10$ ciudades en el archivo del ejemplo usando el método *vecino mas lejano* y la distancia *cuadrada Euclidiana*. En el inicio cada una de las 10 ciudades forma un conglomerado separado. Los primeros conglomerados unidos fueron aquellos que contenían *Boston* y *Washington*, en una distancia de aproximadamente 4. Después, *Atlanta* fue unida al conglomerado que contiene *Boston* y *Washington*. En un tercer paso, *New York* y *Chicago* fueron unidas en un solo conglomerado, y entonces *Los Ángeles* y *Houston* fueron unidos. El procedimiento continúa hasta que un solo conglomerado es formado.

La forma general de un dendograma sugiere agrupar las ciudades en dos grupos:

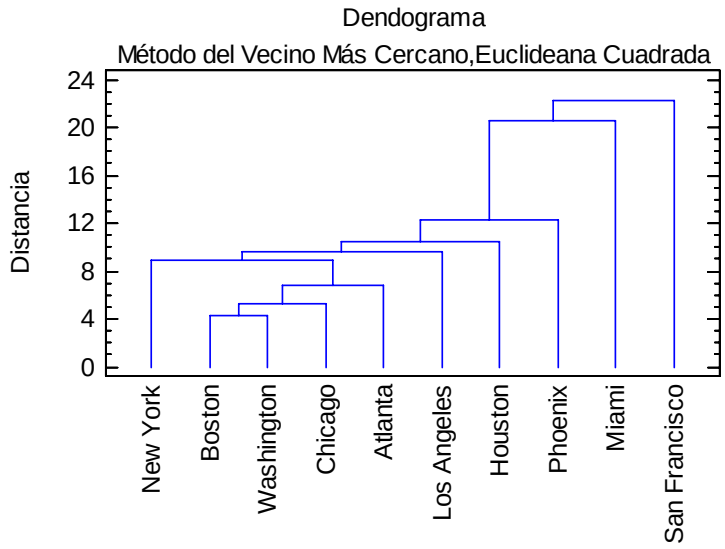
Grupo #1: New York, Chicago, Boston, Washington, Atlanta y San Francisco.

Grupo #2: Los Angeles, Houston, Phoenix, y Miami.

Ya que el Grupo #2 contiene ciudades que tienden a estar localizadas en áreas mas calientes, hace parecer que el clima juega un papel importante en el agrupamiento de las ciudades cuando el método *vecino más lejano* es usado.

Algo diferente se obtiene usando el método *vecino más cercano*:

© 2006 por StatPoint, Inc.



Particularmente los saltos son los cambios de localización entre Los Ángeles y San Francisco. Los Ángeles parecen unirse a otras "grandes" ciudades mas pronto que con el método anterior.

Tabla de Miembros

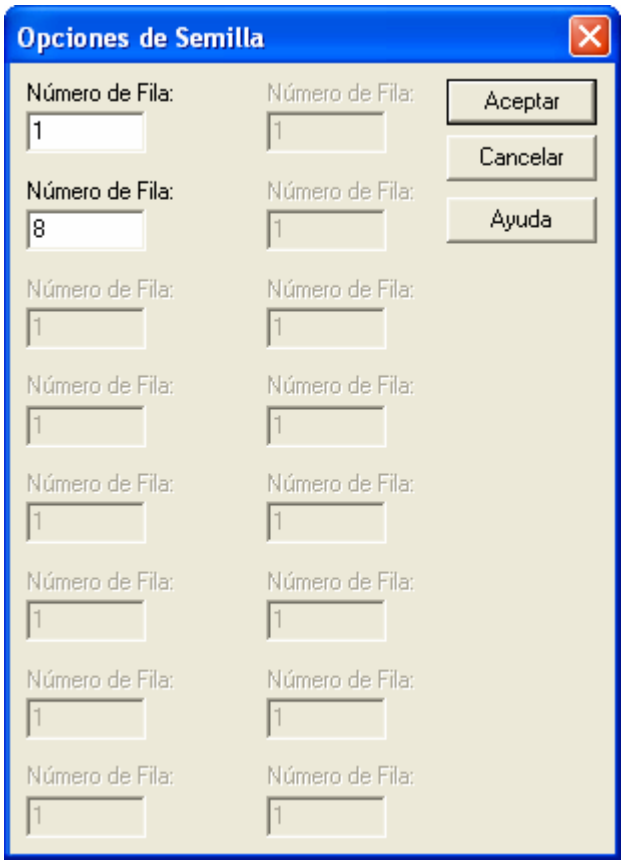
La *Tabla de Miembros* muestra la asignación de las observaciones o variables a cuales conglomerados. Su uso será ilustrado en el siguiente ejemplo.

Ejemplo – Método de k-medias

Ambos métodos usados anteriormente indican que *New York* y *Miami* son muy diferentes entre ellas. Es interesante ver que agrupación ocurriría si uno pide crear 2 conglomerados, usando esas ciudades como semillas. Para hacer esto, ingresamos a *Opciones del Análisis*:

Selecciona *k-medias* y escribe 2 en *Numero de Conglomerados Clusters*. Luego presiona *semilla* e introduce los números de renglón de *New York* y *Miami*:

© 2006 por StatPoint, Inc.



Presiona *OK* un par de veces para generar el análisis. Aunque el dendograma no esta disponible cuando se usa el método de *k-medias* (ya que la conglomeración no es), la *Tabla de Miembros* muestra las asignaciones finales de los conglomerados:

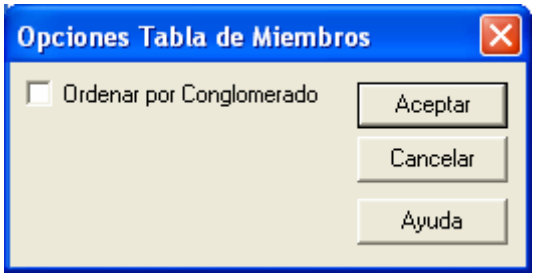
Tabla de Miembros
Método de Conglomeración: k-Medias
Métrica de Distancia: Euclideana Cuadrada

Fila	*Etiqueta*	*Conglomerado*
1	New York	1
2	Boston	1
3	Chicago	1
4	Washington	1
5	Atlanta	1
6	Los Angeles	1
7	San Francisco	1
8	Miami	2
9	Houston	2
10	Phoenix	1

La única cuidad que es puesta con Miami es Houston. Todas las demás caen en el conglomerado de New York.

© 2006 por StatPoint, Inc.

Opciones del Panel



Selecciona *Ordenar por Cluster* para ordenas los objetos por numero de conglomerado.

Gráfico Icicle

El *Gráfico Icicle* provee un modo adicional para ilustrar el conglomerado que ha ocurrido. Esto es muy útil cuando el número de objetos es pequeño:

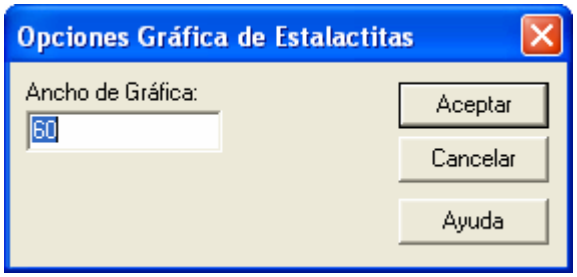
Gráfica de Estalactitas
Método de Conglomeración: Vecino Más Lejano (Vínculo Completo)
Métrica de Distancia: Euclideana Cuadrada

```
                  Número de Conglomerados
                                1
Etiqueta    Fila    1234567890
---------- ------ ----------
New York   1      XXXXXXX
                  XXXXXXX
Chicago    3      XXXXXXX
                  XXXX
Boston     2      XXXXXXXXX
                  XXXXXXXXX
Washington 4      XXXXXXXXX
                  XXXXXXXX
Atlanta    5      XXXXXXXX
                  XXX
San Franci 7      XXX
                  X
Los Angele 6      XXXXXX
                  XXXXXX
Houston    9      XXXXXX
                  XXXXX
Phoenix    10     XXXXX
                  XX
Miami      8      XX
```

Bajo cada *Numero de Conglomerados* esta un renglón de X's. Cualesquiera objetos conectados por X's contiguas son contenidas en el mismo conglomerado. Por ejemplo, el renglón abajo del "2" muestra que cuando las ciudades son divididas en dos conglomerados, los conglomerados consisten de las primeras 6 ciudades y de las ultimas 4.

© 2006 por StatPoint, Inc.

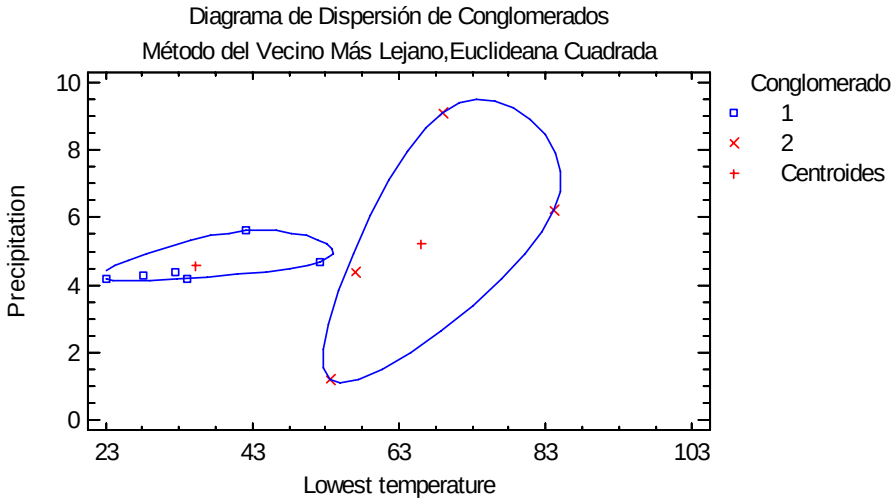
Opciones del Panel



- **Ancho del Gráfico**: el máximo número de caracteres a ser mostrados en una sola pagina.

Gráfico de Dispersión 2D

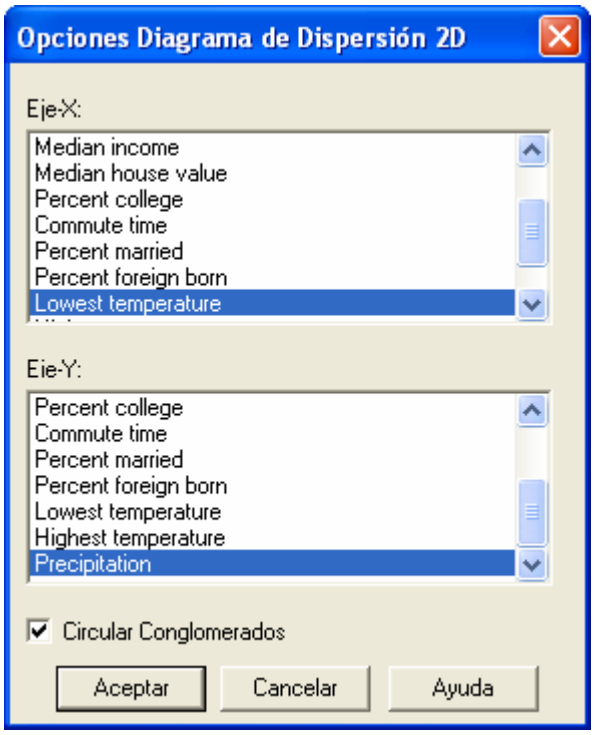
El *Gráfico de Dispersión 2D* muestra la conglomeración con respecto a algún par de variables de entrada:



Cada observación en la hoja de datos es graficada, junto con los centroides de los conglomerados. Si se desea, una curva puede ser usada para conectar observaciones en los bordes de cada conglomerado. En los datos del ejemplo, los conglomerados son bastante bien separados en el espacio de *Lowest temperature* y *Precipitation*.

© 2006 por StatPoint, Inc.

Opciones del Panel

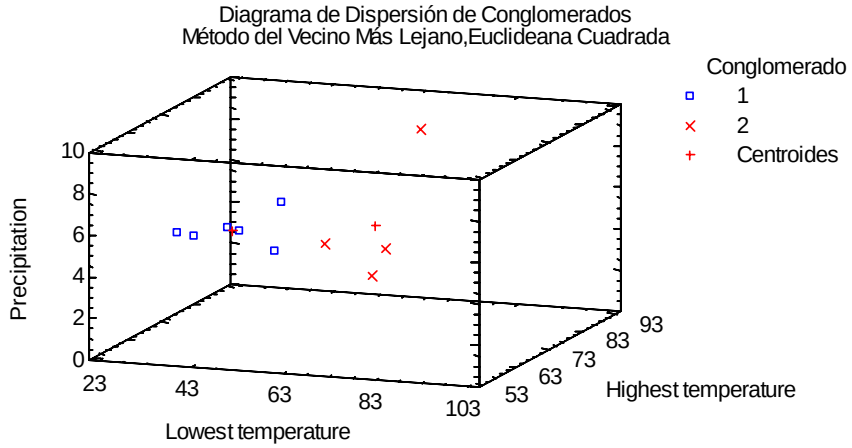


- **Ejes X y Y:** las variables a ser graficadas en el eje horizontal y vertical.

- **Conglomerados en Círculos:** si se selecciona, una curva será usada para conectar las observaciones alrededor de los bordes de cada conglomerado.

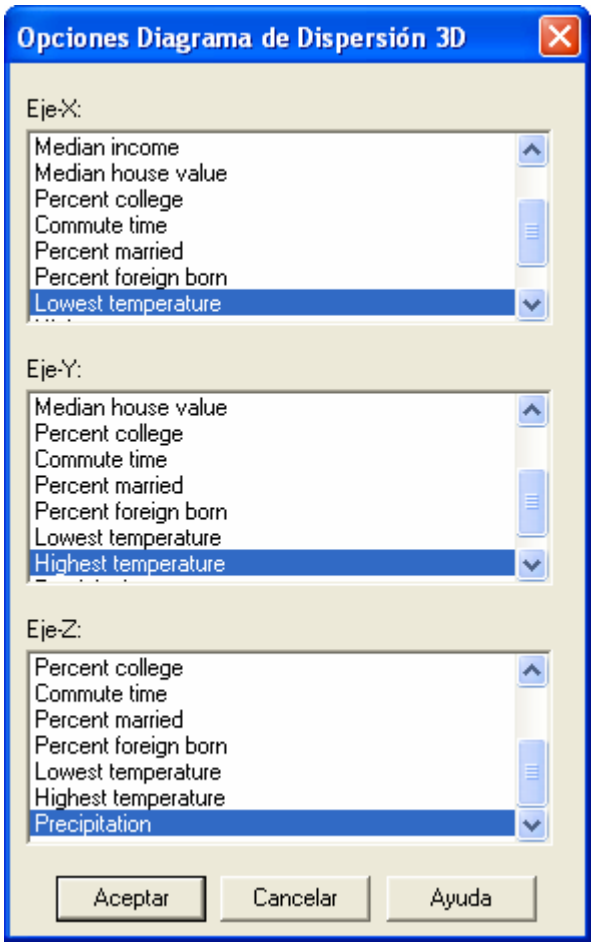
Gráfico de Dispersión 3D

El *Gráfico de Dispersión 3D* muestra la conglomeración con respecto a cualesquiera 3 variables de entrada:



© 2006 por StatPoint, Inc.

Opciones del Panel



- **Ejes X, Y y Z:** las variables son graficadas en 3 ejes.

© 2006 por StatPoint, Inc.

Esquema de Aglomeración

El *Esquema de Aglomeración* provee un resumen de cada paso en el algoritmo de conglomeración aglomerativo:

Programación de la Aglomeración
Método de Conglomeración: Vecino Más Lejano (Vínculo Completo)
Métrica de Distancia: Euclideana Cuadrada

	Conglomerado 1	*Conglomerado 2*		*Etapa Previa*	*Etapa Previa*	*Etapa*
Etapa	*Combinado*	*Combinado*	*Distancia*	*Conglomerado 1*	*Conglomerado 2*	*Siguiente*
1	2	4	4.33537	0	0	2
2	2	5	7.24389	1	0	6
3	1	3	8.94147	0	0	6
4	6	9	10.4417	0	0	5
5	6	10	23.4536	4	0	8
6	1	2	30.6609	3	2	7
7	1	7	32.8948	6	0	0
8	6	8	40.9814	5	0	0

Conglomerado	*Menor*
Número	*Fila*
1	1
2	6

La sección superior muestra:

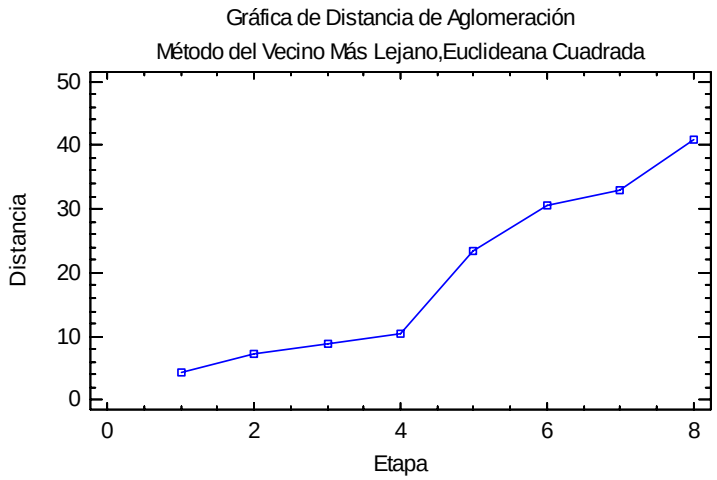
- **Etapas**: El numero de pasos en el algoritmo.
- **Conglomerados Combinados**: El número de observaciones o variables combinadas en cada etapa. Por ejemplo, en la etapa 1, ciudades #2 y #4 fueron combinadas para formar un solo conglomerado. El conglomerado retiene el mas pequeño de los dos números para combinar conglomerados (i.e., “2”). En la segunda etapa, las ciudades en el conglomerado fueron combinados con la ciudad #5.
- **Distancia**: La distancia entre los conglomerados cuando ellos estas unidos.
- **Etapa Previa**: El numero de la etapa en la cual cada conglomerado ha aparecido por última vez, o 0 si estos no han sido unidos en algún conglomerado en una etapa más temprana.
- **Etapa Siguiente**: La etapa próxima en la cual el conglomerado nuevamente aparece.

La sección inferior de la salida exhibe el número más pequeño de la fila de la hoja de datos entre los miembros de cada conglomerado.

© 2006 por StatPoint, Inc.

Gráfico de Distancia de Aglomeración

El *Gráfico de Distancia de Aglomeración* muestra la distancia mínima entre conglomerados cuando ellos son combinados:



Nótese que en los datos del ejemplo las distancias a través de la etapa 4 son pequeñas. Las primeras 4 uniones evidentemente suceden entre ciudades que son muy similares entre ellas:

Etapas 1 y 2: Boston, Washington y Atlanta
Etapa 3: New York y Chicago
Etapa 4: Los Ángeles y Houston

Después de esto, los conglomerados combinados están a distancias considerables uno de otro.

La grafica de distancia de aglomeración pueda ser de ayuda para determinar cuantos conglomerados naturales existen en los datos.

Guardar Resultados

Los siguientes resultados pueden ser guardados en una hoja de datos:

1. *Números de Conglomerado*: Los números de conglomerado asignados a los datos en cada renglón de las variables de entrada.

2. *Matriz de Distancia*: La matriz de distancias derivada entre objetos que son conglomerados.

© 2006 por StatPoint, Inc.