

# Datos Multivariados

## Métodos basados en Fuerza

---

- Se han desarrollado muchas técnicas para proyectar puntos de altas dimensiones en espacios 2D o 3D. El objetivo principal es mantener las características N-dimensionales y otras de los datos a través del proceso de proyección (p. ej., relaciones que existen en los datos originales también existir después de la proyección). Sin embargo, esto no es posible siempre sobre todo conforme se incrementan las dimensiones. Además, la proyección puede introducir en la visualización artefactos involuntariamente que no están presentes en los datos.

# Datos Multivariados

## Métodos basados en Fuerza

- El escalamiento multidimensional (MDS por sus siglas en inglés) es una clase importante de algoritmos para reducir dimensiones que se usan en análisis estadístico y visualización de la información. La estructura básica de un algoritmo MDS es la siguiente.
  1. Dado un conjunto de datos con  $M$  registros y  $N$  dimensiones, crear un matriz  $\mathbf{D}$  de medidas de similitud entre datos y de tamaño  $M \times M$  (p. ej., distancia Euclideana).
  2. Suponiendo que los datos se proyectan sobre  $K$  dimensiones (para un despliegue  $K = \{1, 2, 3\}$ ), se crea una matriz  $\mathbf{L}$ ,  $M \times K$ , de posiciones para los puntos proyectados. Las  $M$  posiciones iniciales se pueden seleccionar aleatoriamente o utilizando técnicas como PCA.
  3. Calcular una matriz  $\mathbf{P}$ ,  $M \times M$ , de similitudes entre todos los pares de puntos en  $\mathbf{L}$ .
  4. Calcular el valor de estrés  $\mathbf{S}$ , el cual es una medida de las diferencias entre los valores de las matrices  $\mathbf{D}$  y  $\mathbf{P}$ . Existen muchas medidas de estrés y la mayoría asume que los sistemas de coordenadas tienen que ser normalizadas de tal forma que la distancia máxima entre puntos es 1.0.
  5. Si el valor de  $\mathbf{S}$  es suficientemente pequeño (o no cambia mucho con respecto a la iteración anterior o anteriores), entonces termina el proceso.
  6. Buscar una dirección a la cual mover los puntos en  $\mathbf{L}$  de tal forma que se reducen los niveles de estrés individuales. Se debe tener cuidado de que los puntos no oscilan entre las posiciones.
  7. Regresar al paso 3.

# Datos Multivariados

## Métodos basados en Fuerza

- El escalamiento multidimensional (MDS por sus siglas en inglés) es una clase importante de algoritmos para reducir dimensiones que se usan en análisis estadístico y visualización de la información. La estructura básica de un algoritmo MDS es la siguiente.
  1. Dado un conjunto de datos con  $M$  registros y  $N$  dimensiones, crear un matriz  $\mathbf{D}$  de medidas de similitud entre datos y de tamaño  $M \times M$  (p. ej., distancia Euclideana).
  2. Suponiendo que los datos se proyectan sobre  $K$  dimensiones (para un despliegue  $K = \{1, 2, 3\}$ ), se crea una matriz  $\mathbf{L}$ ,  $M \times K$ , de posiciones para los puntos proyectados. Las  $M$  posiciones iniciales se pueden seleccionar aleatoriamente o utilizando técnicas como PCA.
  3. Calcular una matriz  $\mathbf{P}$ ,  $M \times M$ , de similitudes entre todos los pares de puntos en  $\mathbf{L}$ .
  4. Calcular el valor de estrés  $\mathbf{S}$ , el cual es una medida de las diferencias entre los valores de las matrices  $\mathbf{D}$  y  $\mathbf{P}$ . Existen muchas medidas de estrés y la mayoría asume que los sistemas de coordenadas tienen que ser normalizadas de tal forma que la distancia máxima entre puntos es 1.0.
  5. Si el valor de  $\mathbf{S}$  es suficientemente pequeño (o no cambia mucho con respecto a la iteración anterior o anteriores), entonces termina el proceso.
  6. Buscar una dirección a la cual mover los puntos en  $\mathbf{L}$  de tal forma que se reducen los niveles de estrés individuales. Se debe tener cuidado de que los puntos no oscilan entre las posiciones.
  7. Regresar al paso 3.

# Datos Multivariados

## Métodos basados en Fuerza

---

- Ya que es un método de optimización, existe la posibilidad de que la solución sea local por lo que hay varias estrategias para intentar resolver situaciones como esta. También, las soluciones finales son dependientes tanto de las condiciones iniciales o de las funciones de similitud y de estrés. De igual manera, el sistema coordenado después de la proyección no tiene significado para el usuario en términos de las dimensiones de los datos originales.

# Datos Multivariados

## Métodos basados en Fuerza

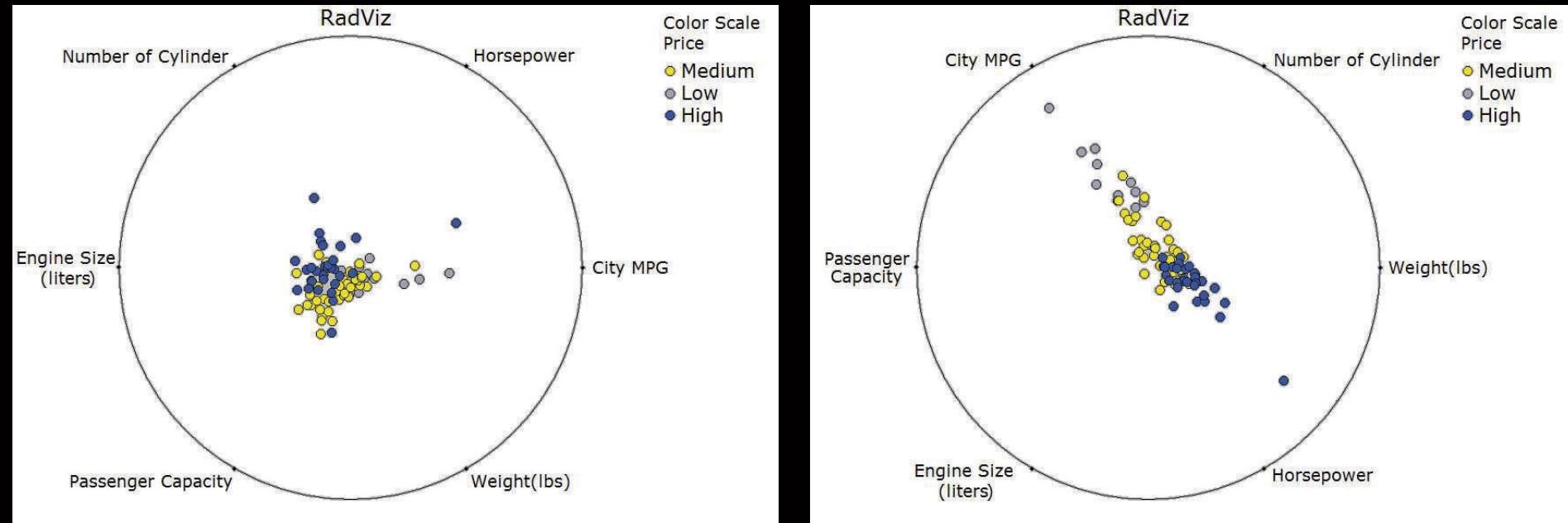


Ilustración de visualización producida por el método RadViz que se basa en la Ley de Hooke para el equilibrio. Para un conjunto de datos  $N$ -dimensional se usan  $N$  puntos ancla que se colocan alrededor de un círculo (radio unitario) para representar los puntos fijos de  $N$  resortes fijados a cada dato.

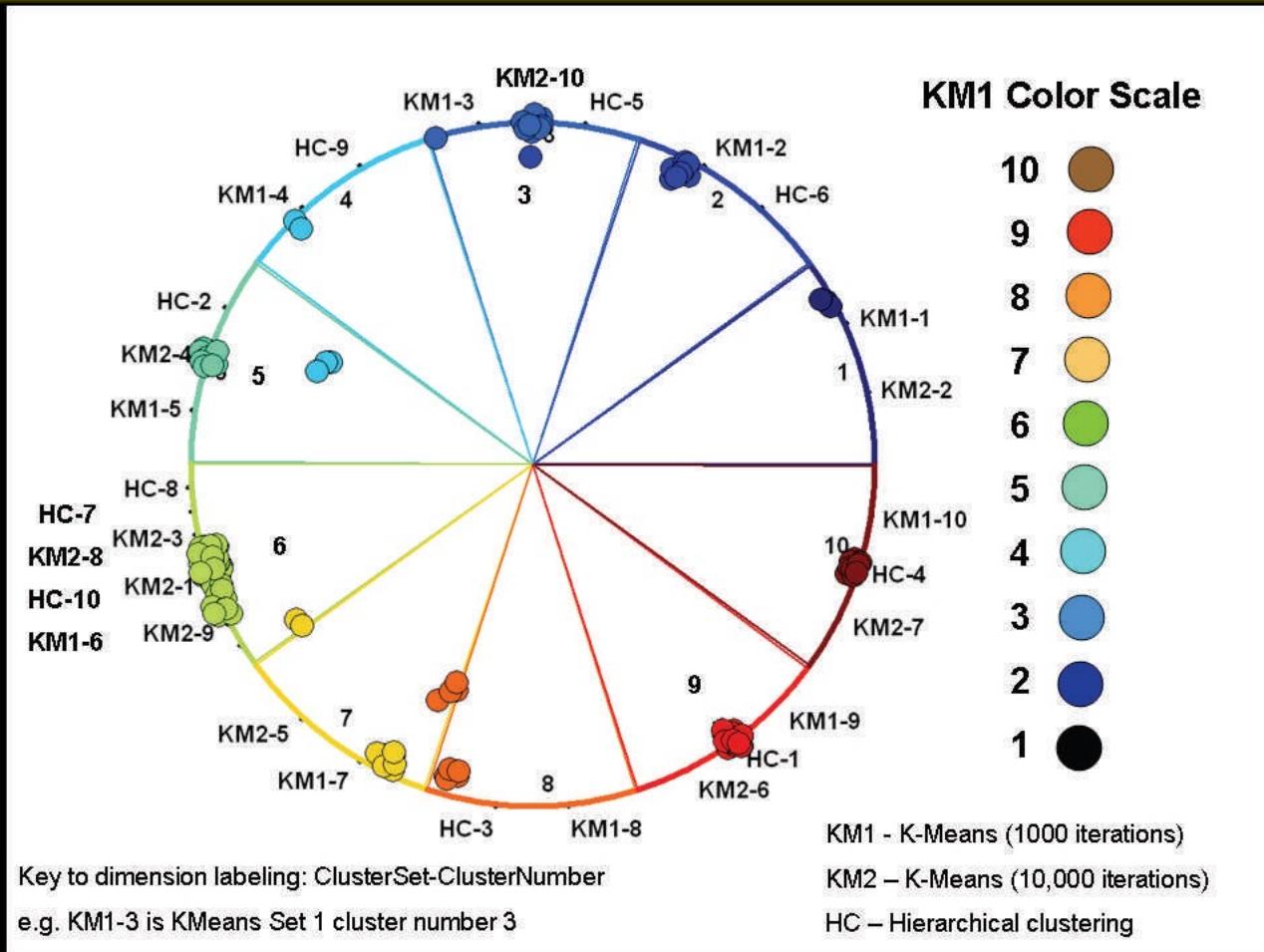
# Datos Multivariados

## Métodos basados en Fuerza

- El método de RadViz Vectorizado (VRV) construye dimensiones múltiples de dimensiones individuales a través de un proceso de aplanamiento, dividiendo cada dimensión en varias (p. ej., dividir la dimensión que representa el número de cilindros de un coche en 5 dimensiones: 1&2, 3&4, 5&6 o 7&8 cilindros). El número de dimensiones nuevas puede ser determinado manualmente o algorítmicamente. Por lo tanto, cada dimensión original se representa por un vector de dimensiones nuevas, cada coordenada nueva en dicho vector tiene un valor 0 o 1 (significa que el registro tiene el valor correspondiente a esa dimensión o no). Por lo que para cada registro cada vector de dimensiones nuevo tiene exactamente una dimensión con el valor igual a 1 con los demás igual a cero. En lugar de usar resortes, se utilizan sumas ponderadas de las posiciones ancla. Para un número grande de puntos ancla el problema es abierto.

# Datos Multivariados

## Métodos basados en Fuerza



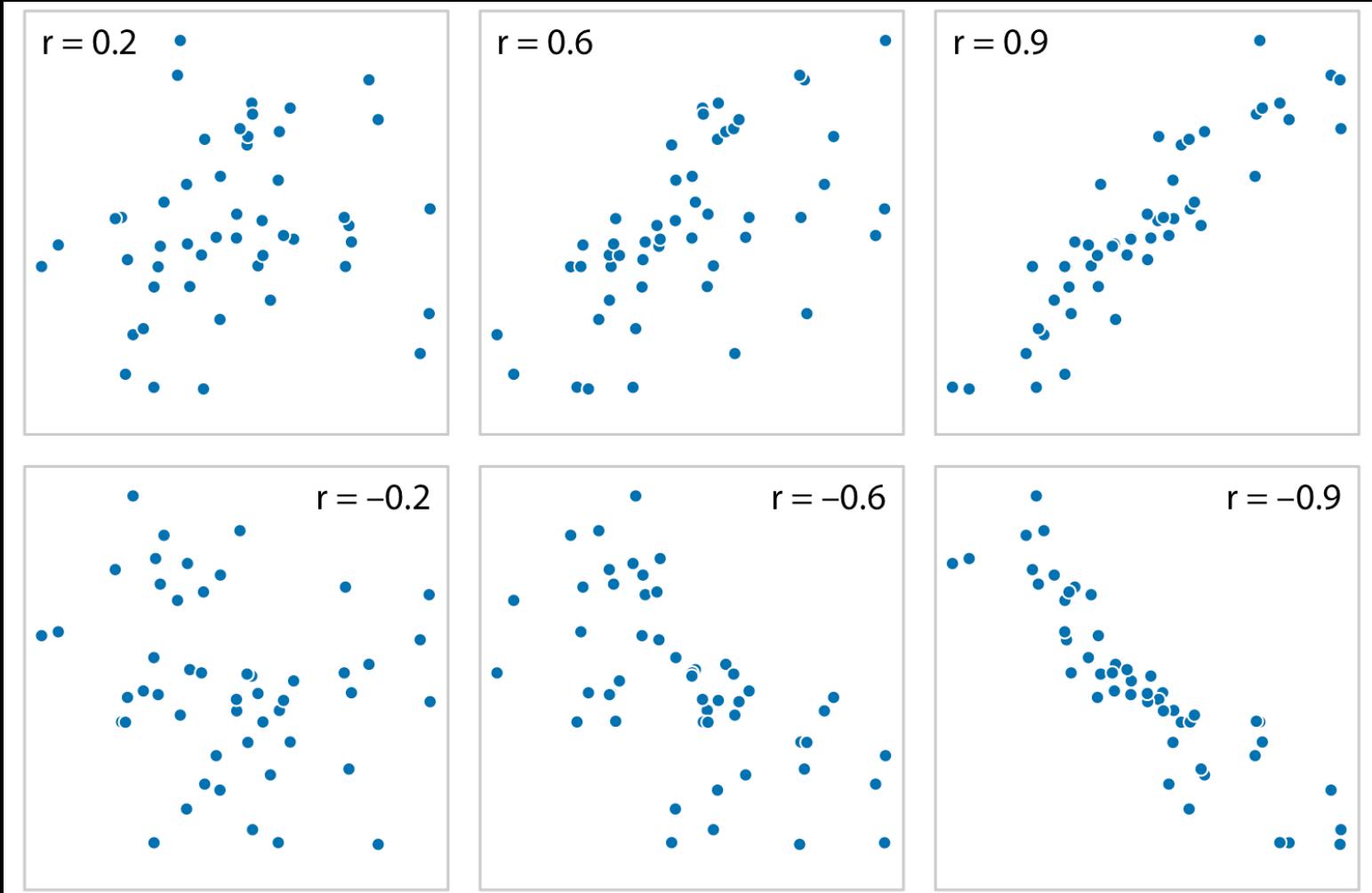
# Datos Multivariados

## Correlogramas

- Cuando hay más de 3 o 4 variables cuantitativas, las matrices de gráficas de puntos esparcidos rápidamente se vuelven pesadas o difíciles de entender. En este caso, puede ser más útil cuantificar la cantidad de asociación entre pares de variables y visualizarlas en lugar de los datos puros. Un forma común de hacer esto es por medio del cálculo de coeficientes de correlación (un valor de 0 significa que no hay correlación y un valor de  $\pm 1$  significa correlación perfecta).
- La visualización de coeficientes de correlación se conocen como correlogramas.

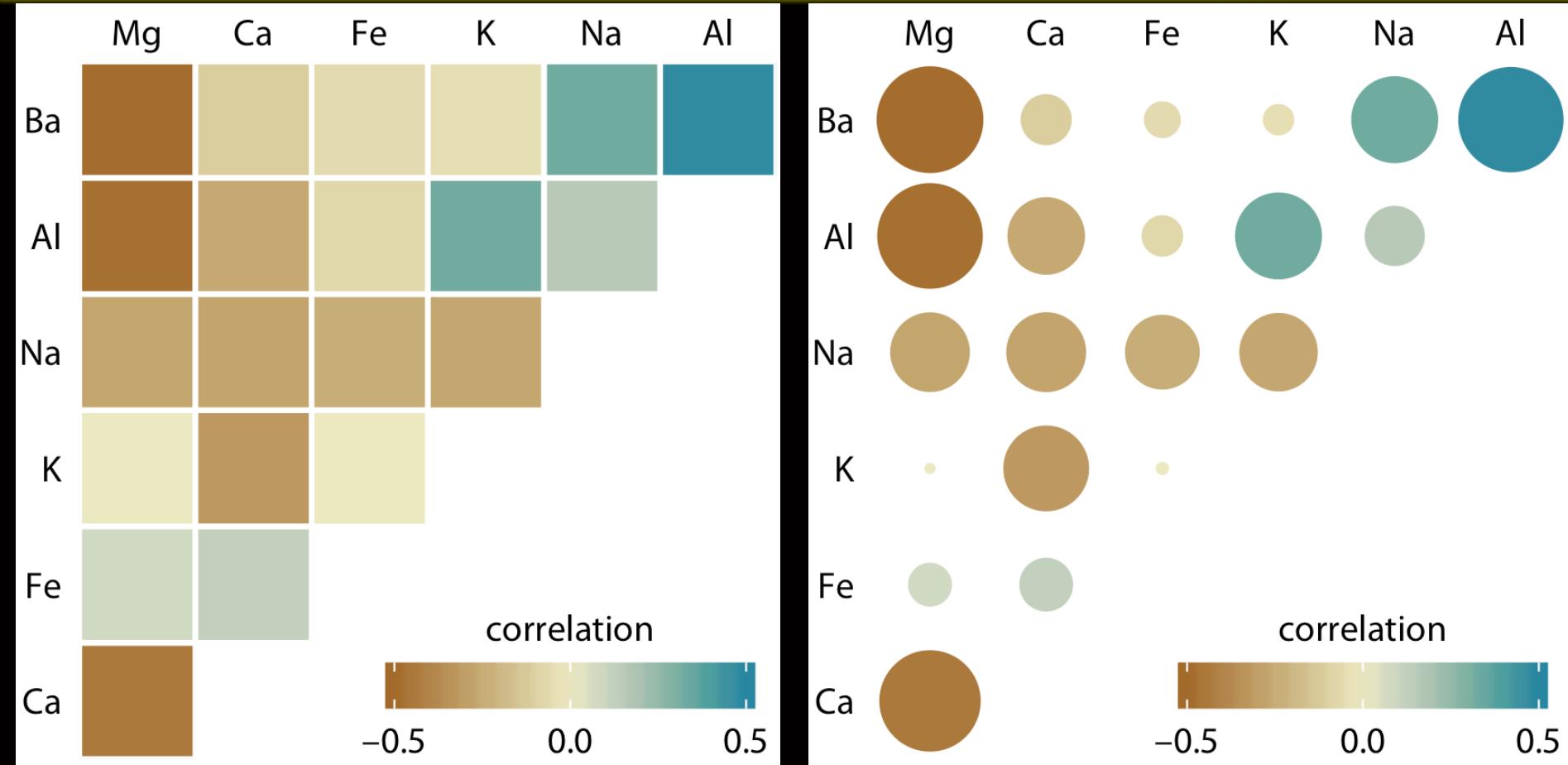
# Datos Multivariados

## Correlogramas



# Datos Multivariados

## Correlogramas



En algunos casos es necesario resaltar las correlaciones aún más y se pueden utilizar círculos para las correlaciones y usar los valores absolutos de la correlación para escalar sus radios.

# Datos Multivariados

## Correlogramas

- Todos los correlogramas tienen una limitación importante: son bastante abstractos. Es cierto que muestran patrones importantes en los datos, también es cierto que ocultan los datos originales y pueden causar que se obtengan conclusiones erróneas. Siempre es mejor visualizar los datos “duros” en lugar de las cantidades abstractas derivadas. Las técnicas reducción de dimensiones es una forma de hallar un punto intermedio.

# Datos Multivariados

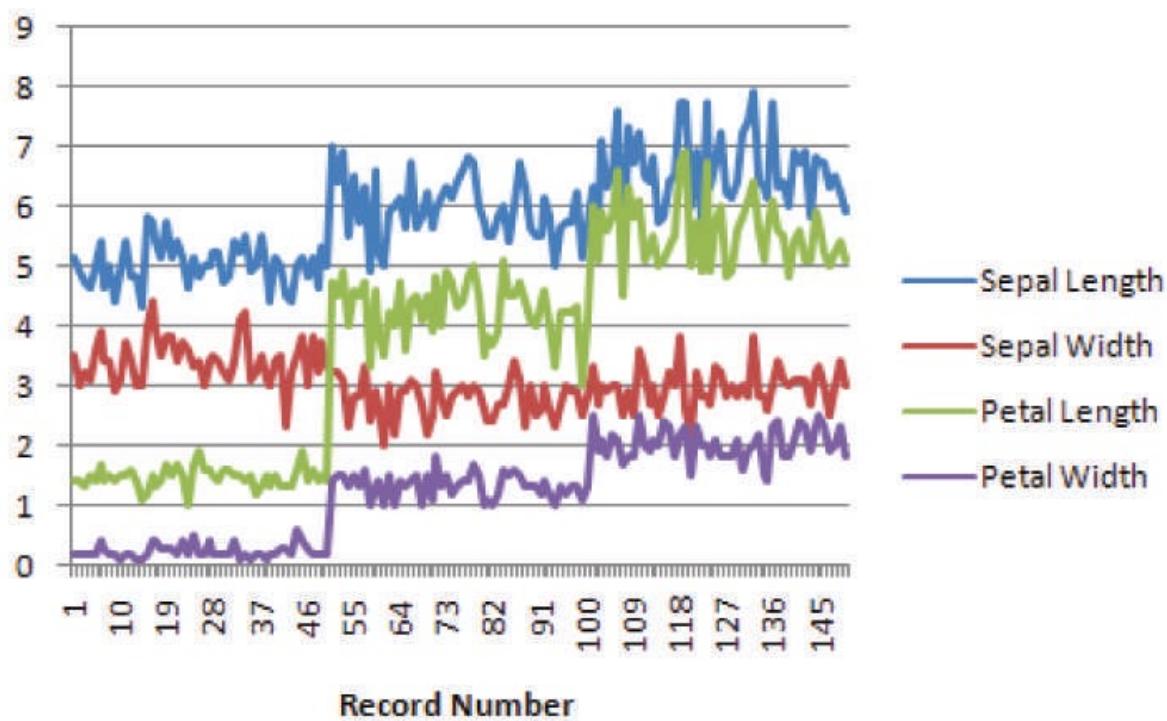
## Técnicas basadas en Líneas

---

- Los métodos basados en puntos representan cada valor en los datos o registro con una marca. En los métodos basados en líneas, los puntos que corresponden a un registro o dimensión particular están ligados a líneas (curvas o rectas). Estas líneas refuerzan la relación entre los valores en los datos además de transmitir características de los datos vía pendientes, curvatura, cruces y otros patrones de línea.

# Datos Multivariados

## Gráficas de Líneas



Para un número modesto de dimensiones en los datos, las gráficas de líneas se pueden trazar sobre un conjunto de ejes comunes, diferenciando las dimensiones usando color, tipo de línea, ancho u otro atributo gráfico.

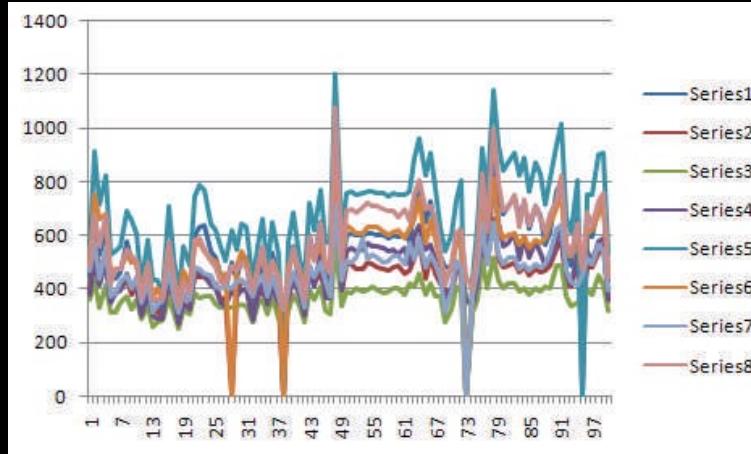
# Datos Multivariados

## Gráficas de Líneas

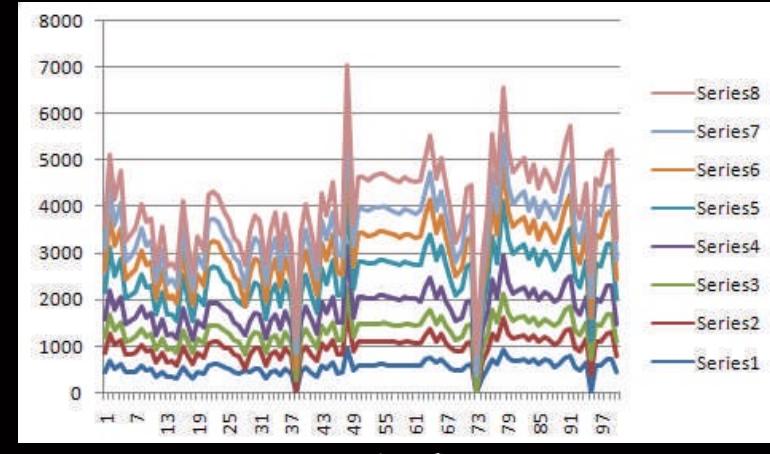
- Para un número modesto de dimensiones en los datos, las gráficas de líneas se pueden trazar sobre un conjunto de ejes comunes, diferenciando las dimensiones usando color, tipo de línea, ancho u otro atributo gráfico.
- Cuando el número de dimensiones se incrementa o las dimensiones se traslanan significativamente en sus rangos de datos, superponer se vuelve más problemático. Alternativamente, se puede usar una línea base, cada dimensión usa la gráfica previa como base; aunque esto puede reducir el número de occlusiones, hace más difícil el entendimiento de los valores de los datos. Adicionalmente, se pueden ordenar los datos con respecto a una de las dimensiones.
- Cuando las variables tienen unidades diferentes, la visualización se vuelve más complicada de diseñar. Un método común para resolver este problema es incluir varios ejes verticales, cada uno etiquetado y marcado independientemente. Ambos lados de una gráfica se pueden usar para evitar el desorden.
- Claramente, si las marcas de valores no están alineadas no es sabio utilizar una rejilla en la gráfica, a menos que se puedan cambiar fácilmente las marcas de uno de los ejes. Otra alternativa es crear un conjunto de gráficas, una por cada dimensión y apilarlas verticalmente (usualmente después de escalarlas en la dimensión vertical para permitir a todas, o a la mayoría de, las gráficas ser visibles al mismo tiempo).
- Alternativamente se puede agregar un sistema para visualizar valores en otras dimensiones que corresponden a una característica de interés en alguna de las gráficas.

# Datos Multivariados

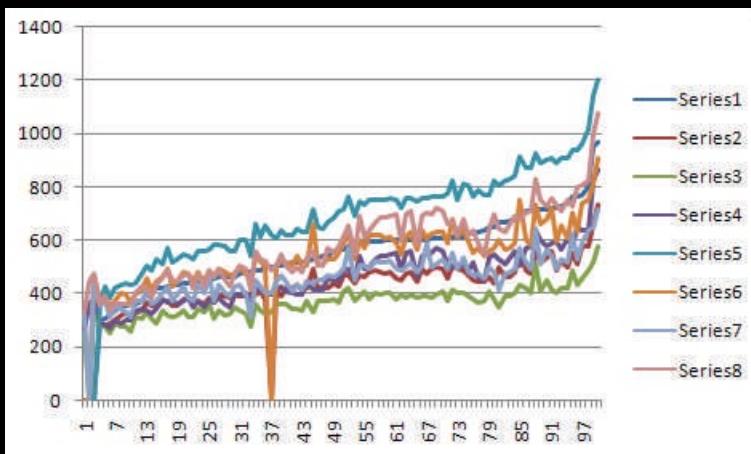
## Gráficas de Líneas



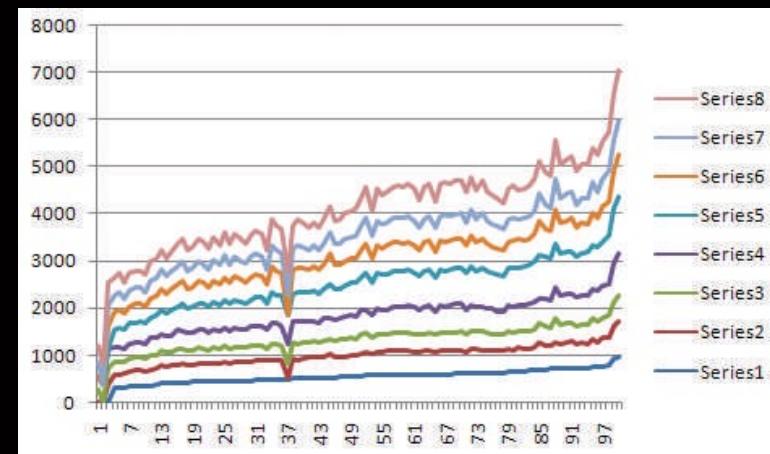
superpuestas



apiladas



superpuestas ordenadas



apiladas ordenadas

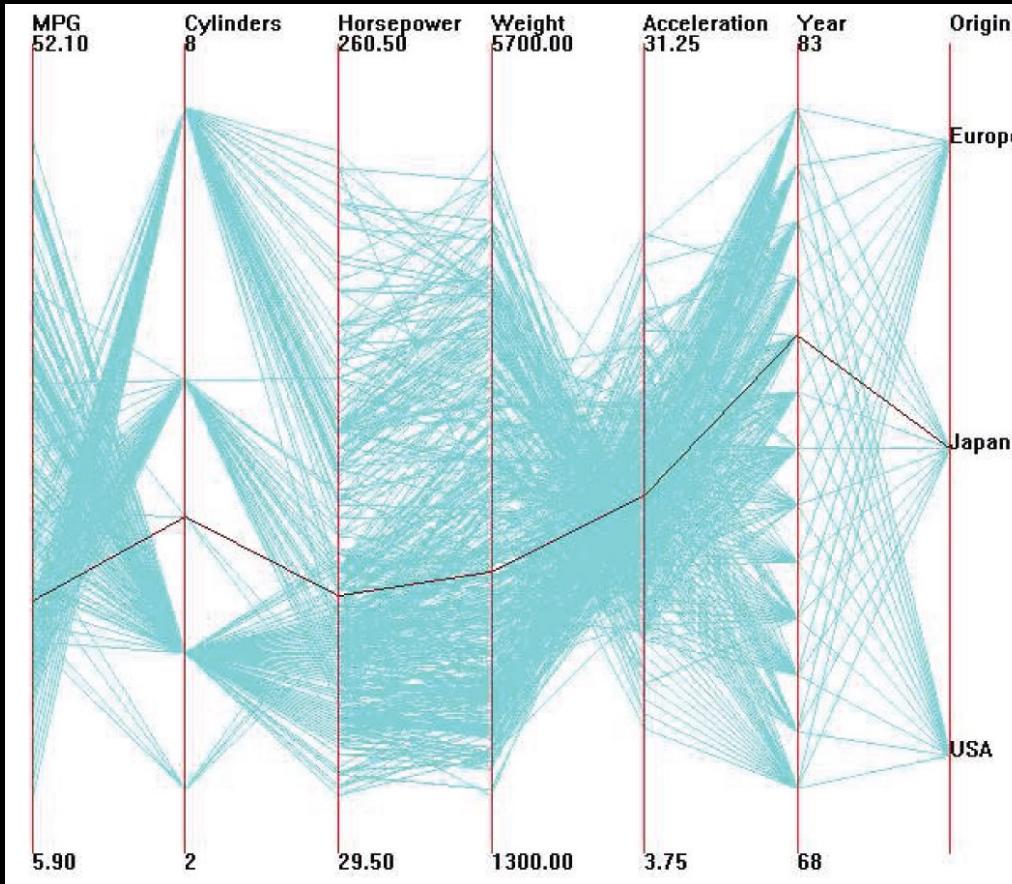
# Datos Multivariados

## Coordenadas Paralelas

- Las gráficas por coordenadas paralelas (PCP por siglas en inglés) fueron introducidas en 1985 como una forma de estudiar geometría de dimensiones altas y desde entonces se han hecho mucho trabajo con respecto a este tipo de gráficas. La idea básica es que los ejes en lugar de ser ortogonales son paralelos, con líneas espaciadas verticalmente u horizontalmente representando un ordenamiento particular de las dimensiones.
- Un dato puntual es graficado como una polilínea que cruza cada eje en una posición proporcional a su valor para esa dimensión.
- Una gráfica de coordenadas paralelas se puede considerar como una gráfica de líneas después de rotar los datos, ya que los valores de un registro están ligadas entre sí en lugar de estar asociados a valores de una dimensión.

# Datos Multivariados

## Coordenadas Paralelas



Un solo dato (*data point*) se representa por medio de la polilínea oscura.

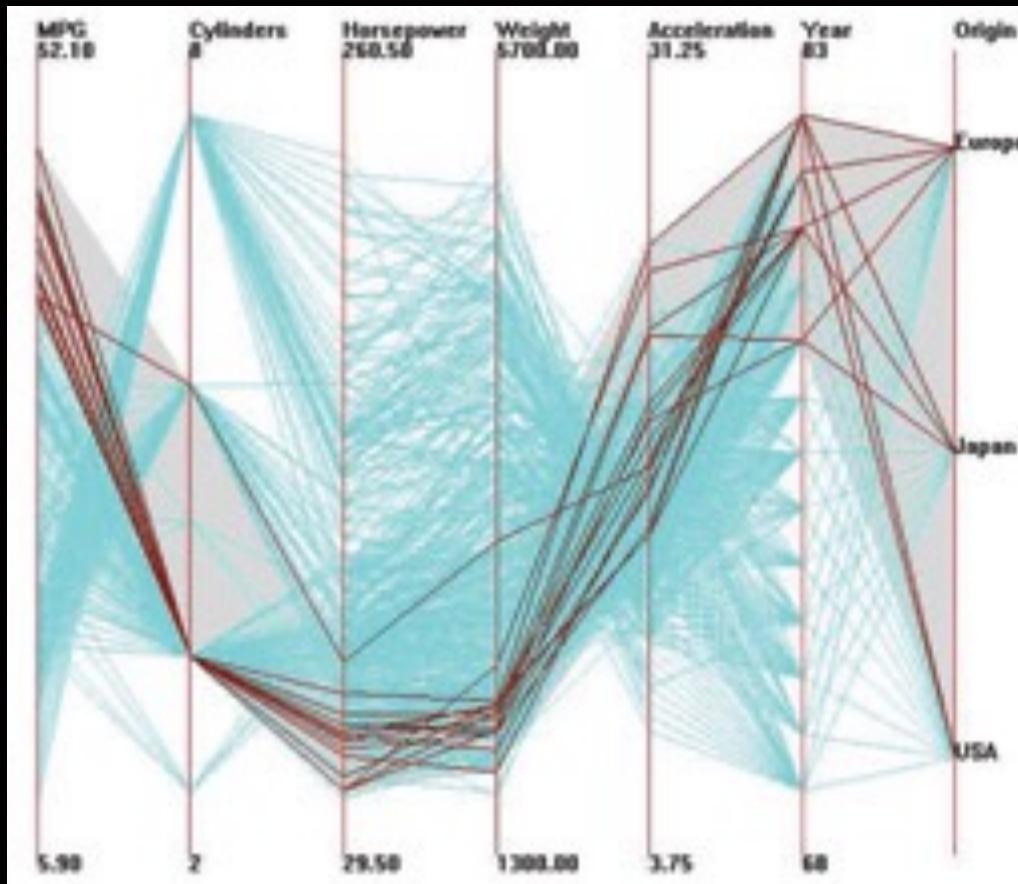
# Datos Multivariados

## Coordenadas Paralelas

- Para interpretar la gráfica, se buscan los grupos (*clusters*) de líneas similares (indicando una correlación parcial entre pares de dimensiones), puntos de cruce similares (indicando correlaciones negativas parciales) y líneas que están aisladas o tienen una pendiente que es significativamente diferente a la de sus vecinos (indicando valores atípicos).
- Un problema es que similar al caso de puntos esparcidos, las coordenadas paralelas son un método muy bueno para mostrara relaciones entre pares de dimensiones. Para extender esta característica, la selección interactiva y resaltar los registros permite a los usuarios ver relaciones que se extienden en todas las dimensiones.

# Datos Multivariados

## Coordenadas Paralelas



Misma gráfica que antes pero seleccionando los valores altos de MPG.

# Datos Multivariados

## Coordenadas Paralelas

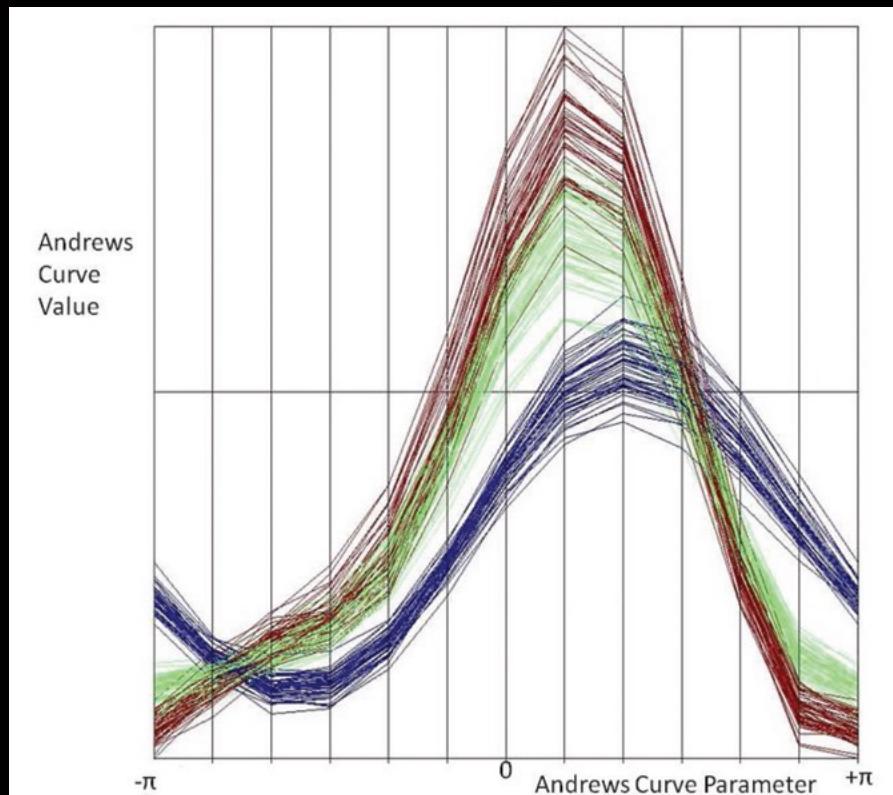
- Este tipo de gráficas se han sido extendidas por varios autores:
  - Coordenadas paralelas jerárquicas que muestran grupos de datos y no los datos originales.
  - Usar líneas semi-transparentes para revelar grupos en grandes conjuntos de datos.
  - Agrupamiento, reordenamiento y espaciado de ejes basándose en correlación.
  - Reordenamiento de ejes para reducir el desorden visual.
  - Agrupamiento de datos en bandas de grupos con tratamiento especial de los datos atípicos.
  - Incorporando histogramas en los ejes para transmitir mejor distribuciones univariadas.
  - Ajuste de curvas para la intersección de puntos para mejorar transmitir continuidad a través de los ejes.

# Datos Multivariados

## Curvas de Andrew

- Son gráficas que usan los registros de un dato son utilizados para aproximar una curva con segmentos de línea recta (polilínea) cuyos puntos de control están determinados por sinusoidales.

Longitud de sépalo  
Ancho de sépalo  
Longitud de pétalo  
Ancho de pétalo



# Datos Multivariados

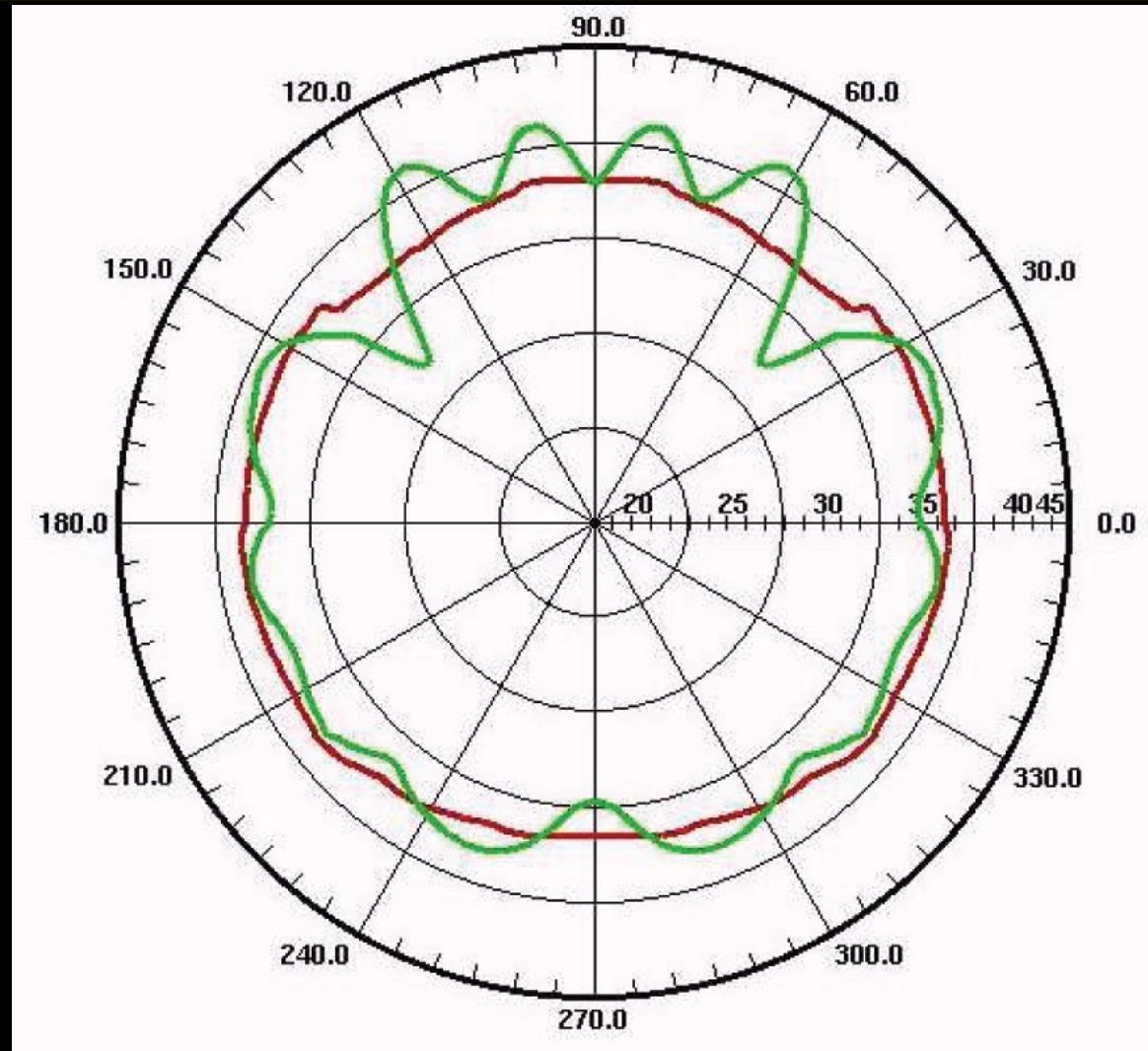
## Técnicas de Ejes Radiales

- Por cada técnica que orienta los sistemas coordenados horizontalmente y/o verticalmente, existe una técnica equivalente que usa orientación radial.
- Una gráfica larga puede ser anidada dividiéndola en segmentos homogéneos y mapeando cada uno a una base de radio diferente. Esta técnica es potencialmente útil para estudiar eventos cíclicos. Variaciones de las gráficas circulares de líneas incluyen gráficas de radar y de estrellas.
- Se han desarrollado otras gráficas circulares:
  - Gráficas polares (gráficas de puntos usando coordenadas polares).
  - Gráficas circulares de barras (similar a gráficas de líneas circulares, pero usando barras en la línea base).
  - Gráficas circulares de área (similar a gráficas de línea, pero con área bajo la línea rellenada con un color o textura).
  - Gráficas circulares de barra (se usan barras que son arcos circulares con un punto central común y línea base)

# Datos Multivariados

## Curvas de Andrew

Gráfica circular  
de línea



# Datos Multivariados

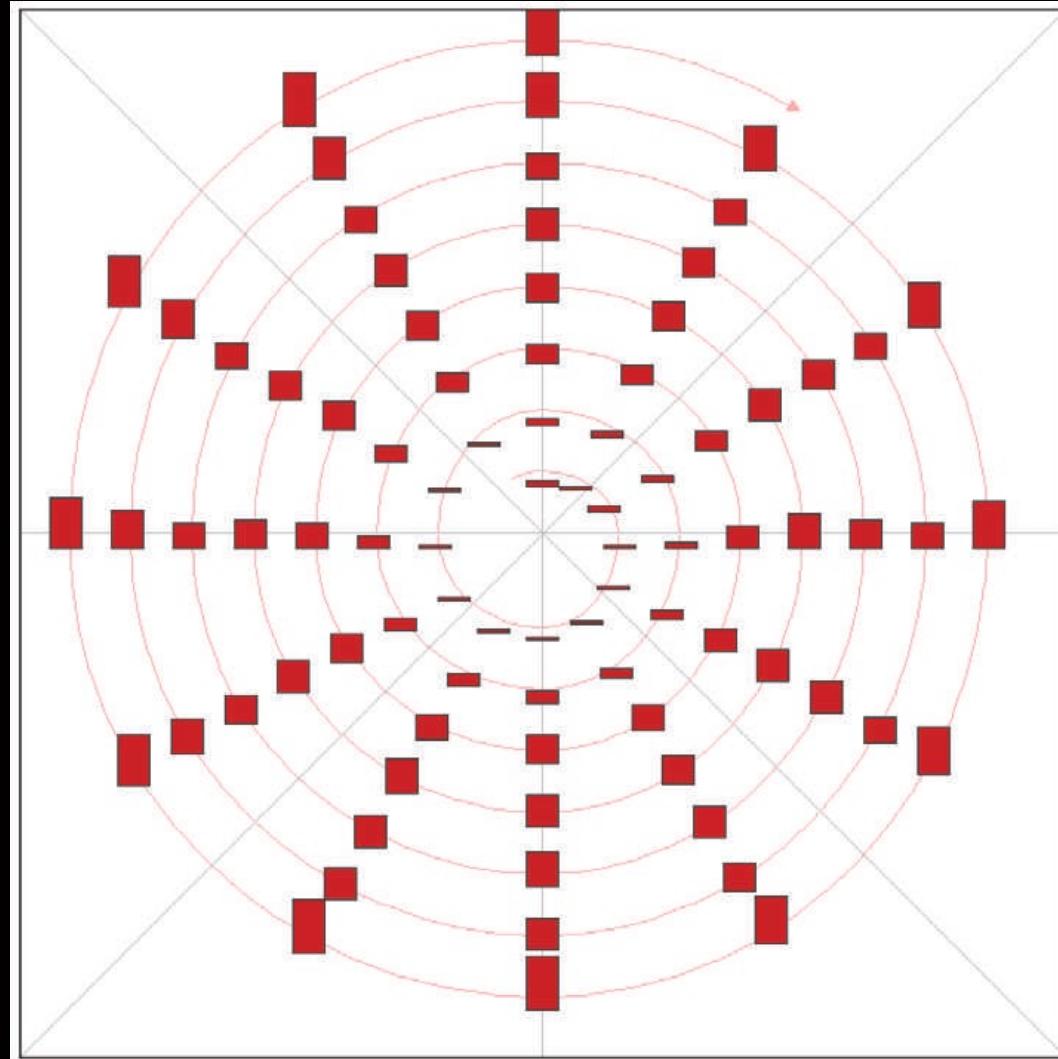
## Técnicas de Ejes Radiales

- Cada una de las técnicas que usan ejes radiales e involucran más de un ciclo pueden arreglarse sobre anillos concéntricos o una espiral continua.
- Los métodos de espiral no sufren de la discontinuidad al final de cada ciclo que se presentan en los arreglos con círculos concéntricos.
- Las comparaciones entre ciclos al interior y entre ciclos son fáciles de realizar, especialmente con las barras orientadas a lo largo del eje vertical en lugar de estar perpendiculares a la base de la espiral. Debido a la percepción humana, es más medir diferencias entre elementos adyacentes que tienen una línea base; sin embargo, una gráfica de barras tradicional no permite ver patrones entre elementos en la misma posición de ciclos diferentes.

# Datos Multivariados

## Curvas de Andrew

Gráfica sobre  
una espiral.



# Datos Multivariados

## Técnicas Basadas en Regiones

- La técnicas basadas en regiones rellenas polígonos para transmitir los valores, con base en sus tamaños, forma, colores u otros atributos. Hay que recordar que los humanos somos peores para medir áreas que para medir longitudes, pero se han desarrollado métodos efectivos para usar regiones como base de visualización. Para algunos, el objetivo no es mostrar los datos puros sino resúmenes o distribuciones de los valores. Varios de los métodos basados en regiones fueron diseñados originalmente para datos univariados, tales como gráficas de pastel o de barras pero que han sido extendidas a dimensiones múltiples.

# Datos Multivariados

## Histogramas y Gráficas de Barras

- Uno de los métodos más comunes para visualizar es la gráfica de barras en donde las barras rectangulares se usan para transmitir valores numéricos. Recordando que los humanos son buenos para comparar longitudes de características lineales, por lo tanto, las gráficas de barras son una opción natural para visualizar varios tipos de datos.
- Se usan rutinariamente barras horizontales y verticales para visualizar porque son fáciles de interpretar, lo que significa que los diseñadores tienen algo de flexibilidad en la forma en que se integran en una aplicación. Incluso, se puede considerar etiquetar la barras con lo que se puede ayudar a la percepción de longitud.

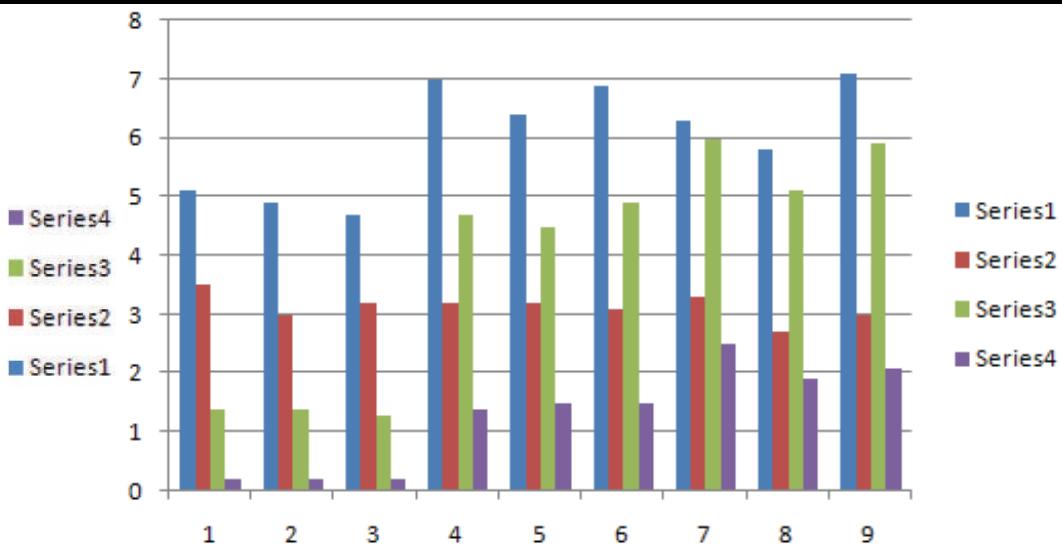
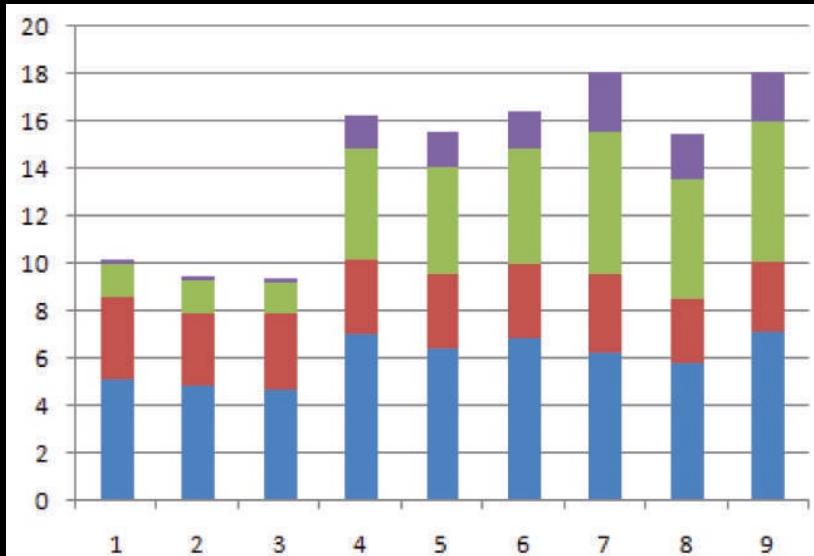
# Datos Multivariados

## Histogramas y Gráficas de Barras

- Una de las decisiones más importantes en la utilización de gráficas de barras es ver la cantidad de barras que son necesarias para mejor representar los datos. Siempre y cuando el número de variables sea relativamente pequeño, se puede usar una correspondencia uno a uno con el número de barras. Si el objetivo es representar un resumen o una distribución del conjunto de datos, se puede utilizar un histograma para informar sobre la ocurrencia de los valores.
- Si los datos son nominales o son un conjunto modesto de números enteros bien definidos, se puede considerar usar una correspondencia uno a uno con el número de barras.
- Para el caso de datos continuos o variables enteras con un rango amplio, es posible dividir los datos en subrangos y asignar cada subrango a una barra. Si los datos son multivariados, se pueden usar varias opciones tales como barras apiladas o barras contiguas.

# Datos Multivariados

## Gráficas de Barras/Histogramas

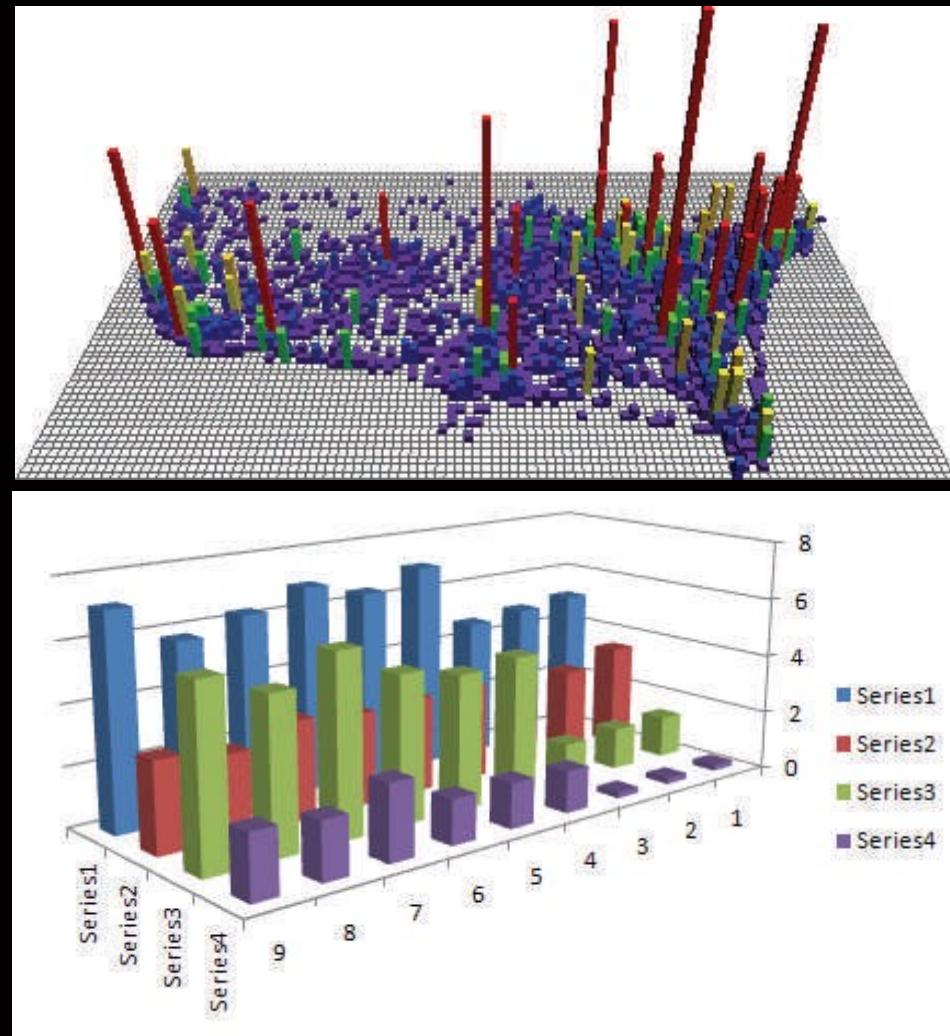


Los humanos tenemos una alta agudeza visual cuando se compara la longitud de características lineales. Como ya vimos, no hay que utilizar un valor de variables muy grande para lograr una buena visualización.

# Datos Multivariados

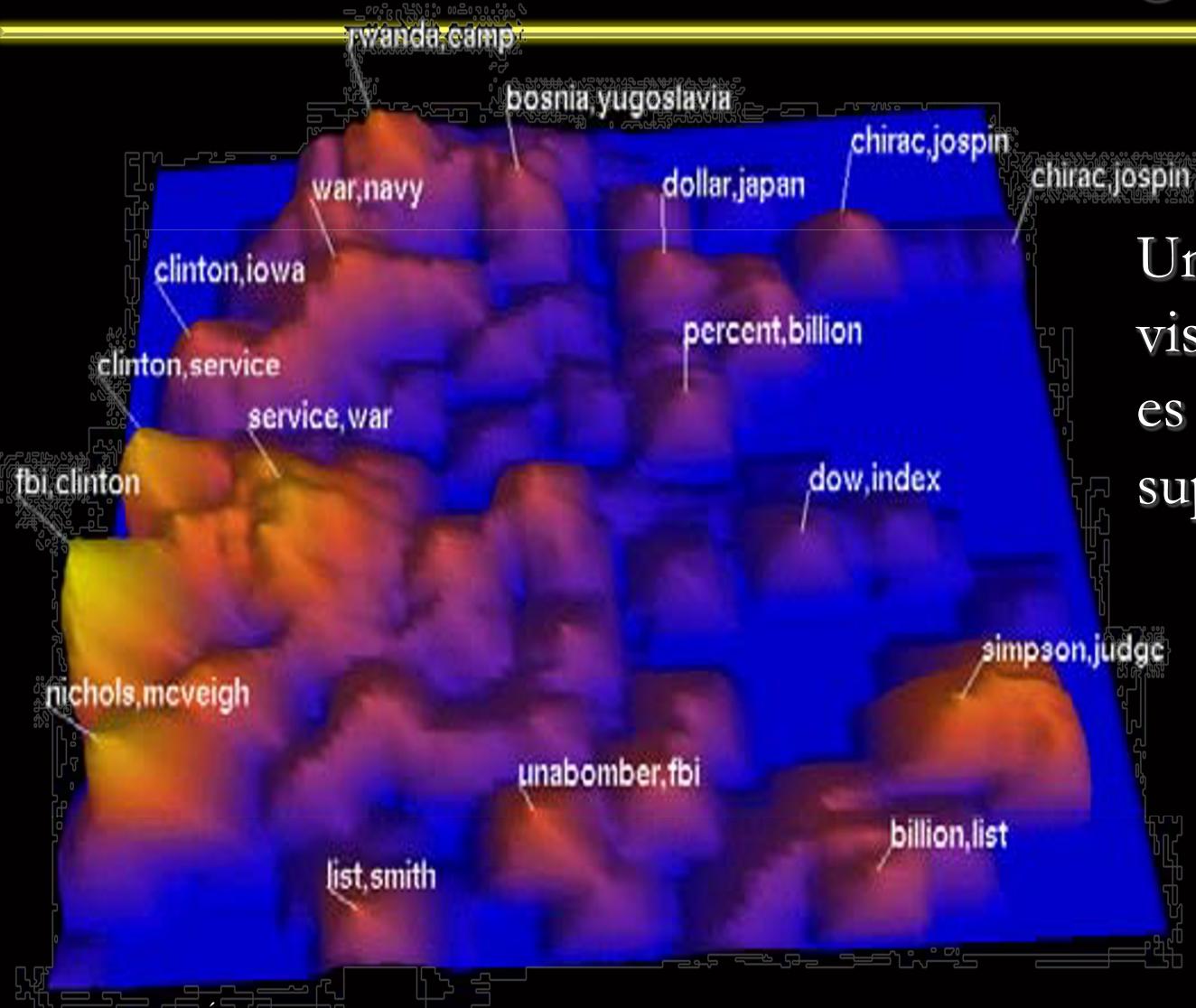
## Gráficas de Barras/Histogramas

Una opción en 3D es la gráfica conocida como *cityscape* (paisaje citadino, las barras parecen edificios) en donde se usan barras 3D en lugar de rectángulos sobre una rejilla. Algunas de las variables controlan altura y color. Como ya hemos visto, este tipo de gráficas tienen el problema de oclusión y existen varias estrategias para minimizar este problema (cambio interactivo de punto de vista, transparencia, área de la base).



# Datos Multivariados

## Gráficas de Barras/Histogramas



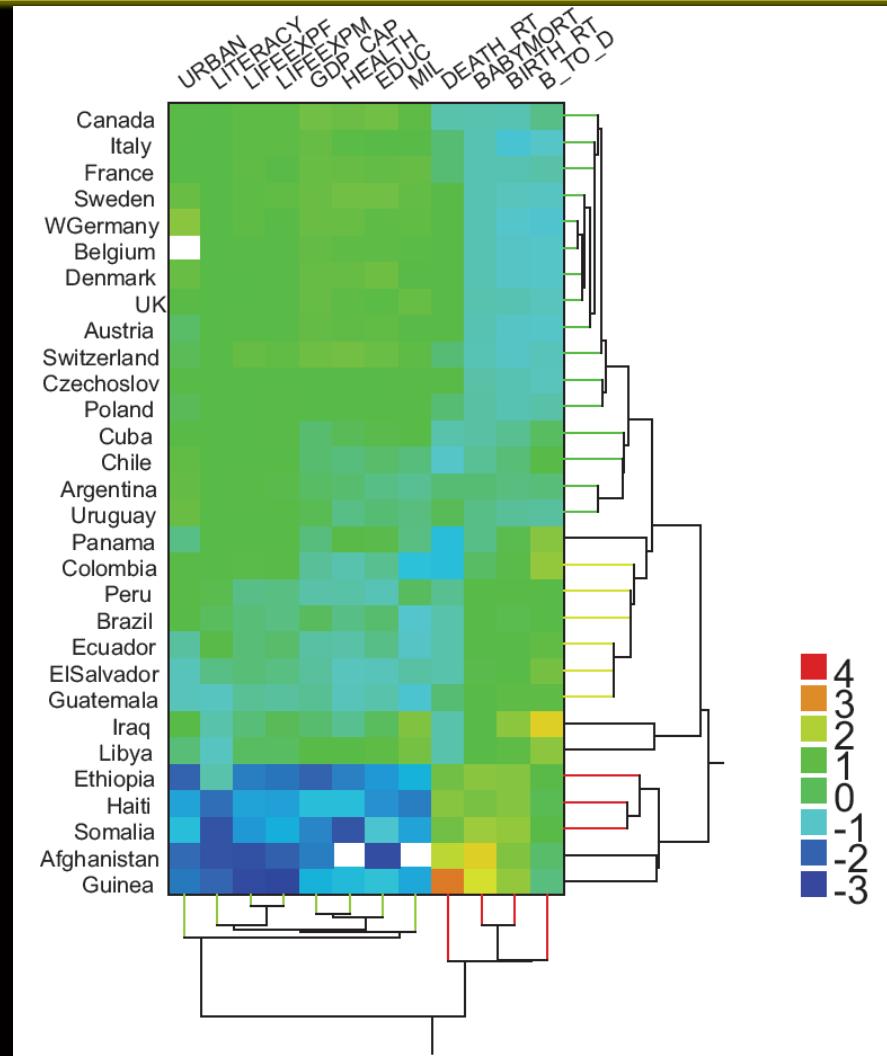
Una opción más visualmente agradable es la gráfica de superficie o paisaje.

# Datos Multivariados

## Gráficas de Tabulares

Es frecuente que los datos multivariados estén almacenados en tablas y hay varios métodos para visualizar esta estructura, (varían en el tipo de interacción que permiten).

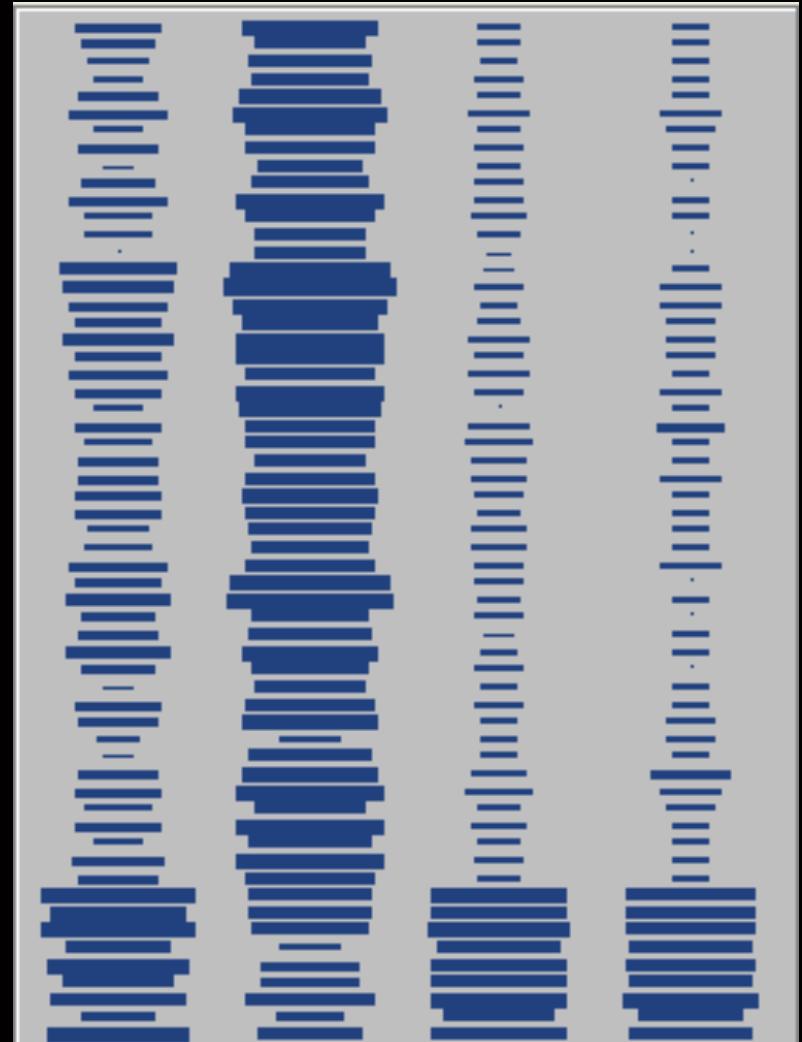
Los mapas de calor (*heatmaps*) despliegan los registros de la tabla usando colores, usando un mapa de color normalizado y cada registro se representa como un rectángulo con un color. Para resaltar los datos o sus relaciones, se pueden permutar las fileras y/o columnas.



# Datos Multivariados

## Gráficas de Tabulares

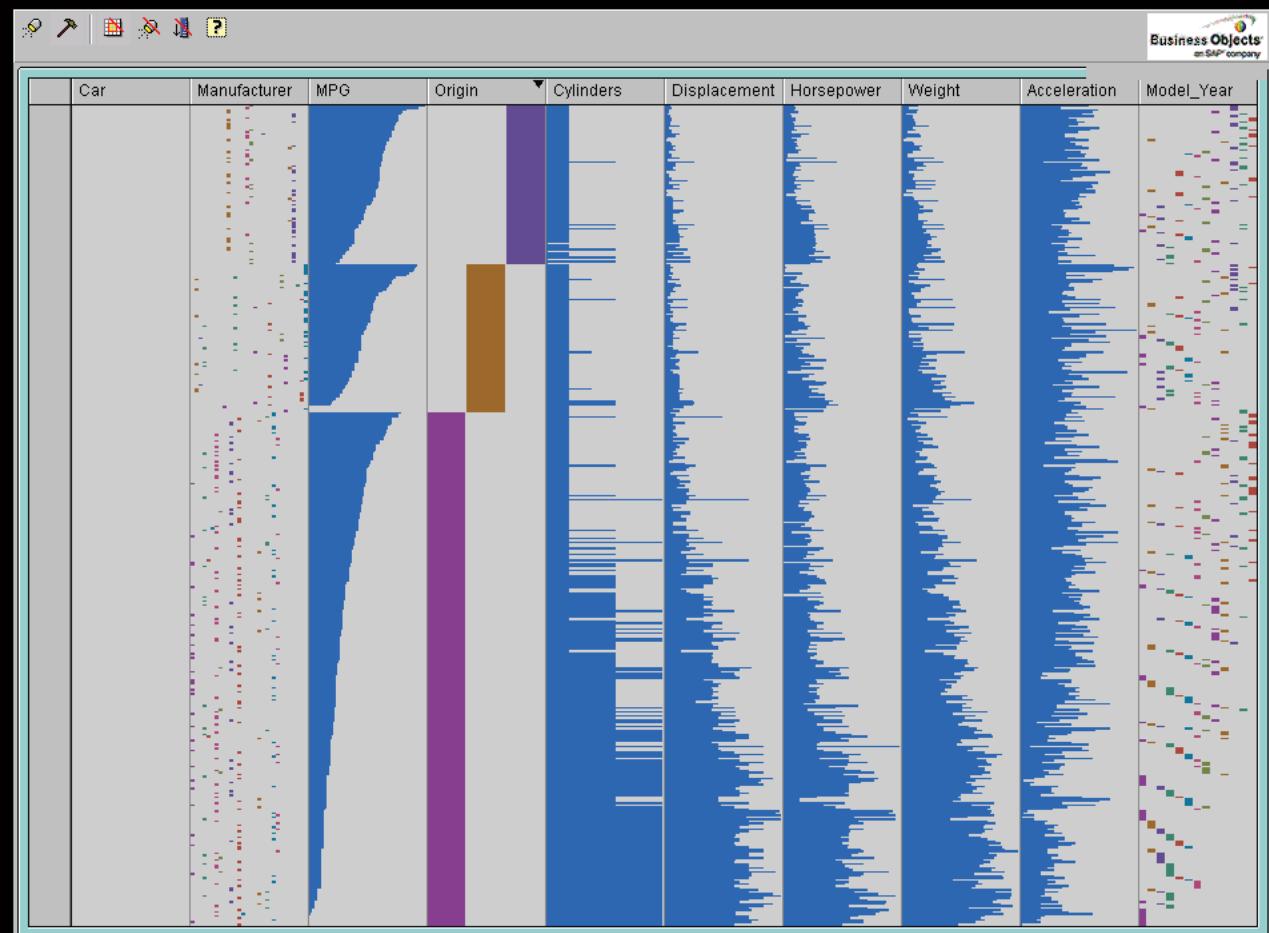
Para evitar el sesgo en la percepción del color (la apreciación de un color depende de sus vecinos) se puede variar el tamaño de las celdas en lugar del color, aunque este método tiene el problema de que los humanos percibimos peor las áreas.



# Datos Multivariados

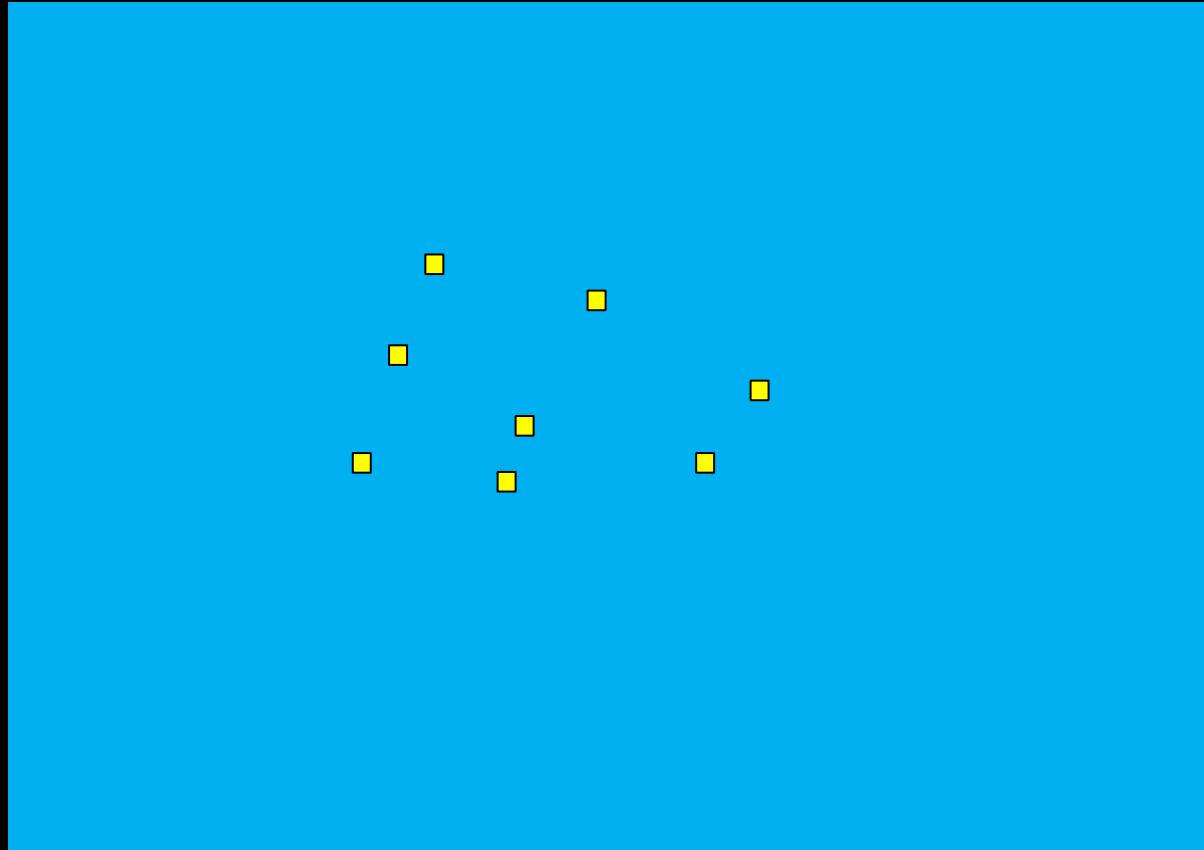
## Gráficas de Tabulares

Otra técnica es *lente tabular* la cual combina las ideas previas e incluye un mecanismo nivel-de-detalle para proveer acercamientos y desplazamiento. Los datos se presentan de varias formas dependiendo del espacio de despliegue, el ordenamiento de columnas e hileras permite analizar los datos (tendencias, correlaciones, datos atípicos).



# Datos Multivariados

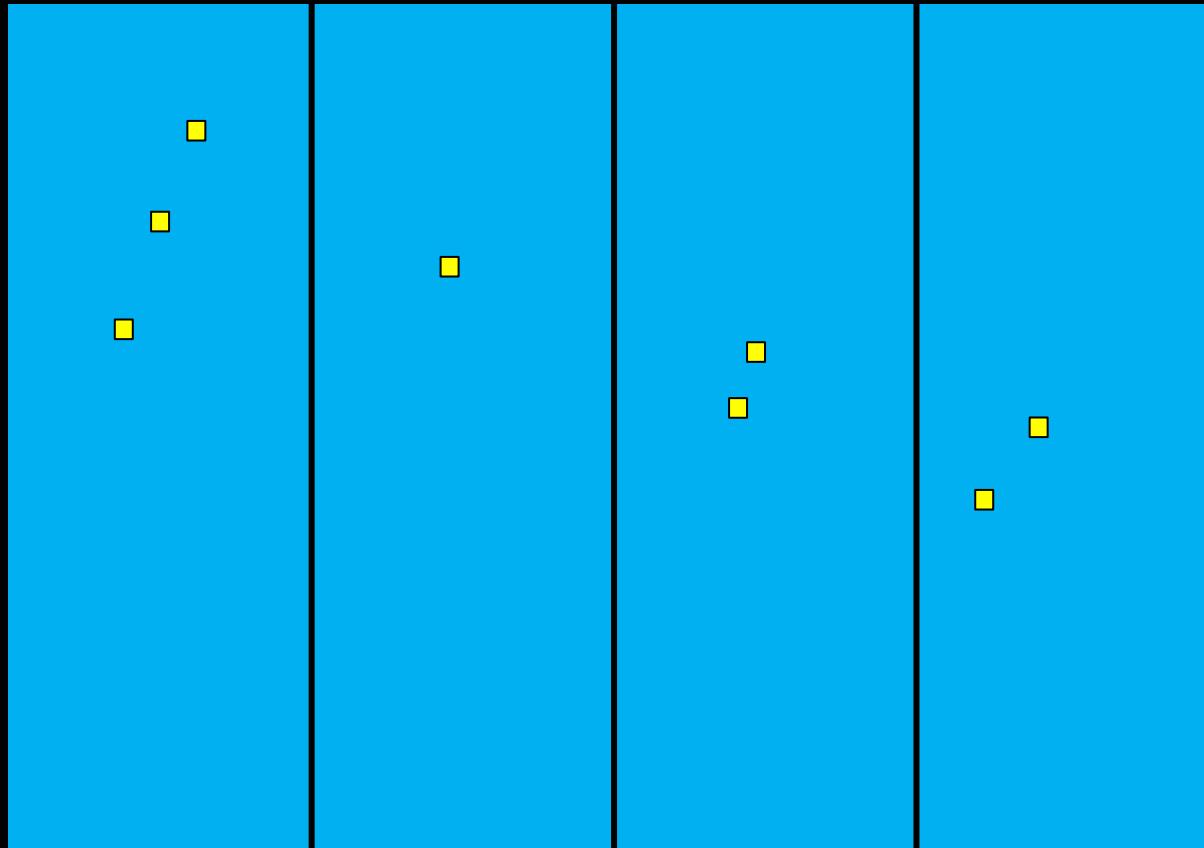
## Apilado Dimensional



Cada registro de datos con cuatro atributos/valores ( $N = 4$ ) se coloca sobre un espacio 2D.

# Datos Multivariados

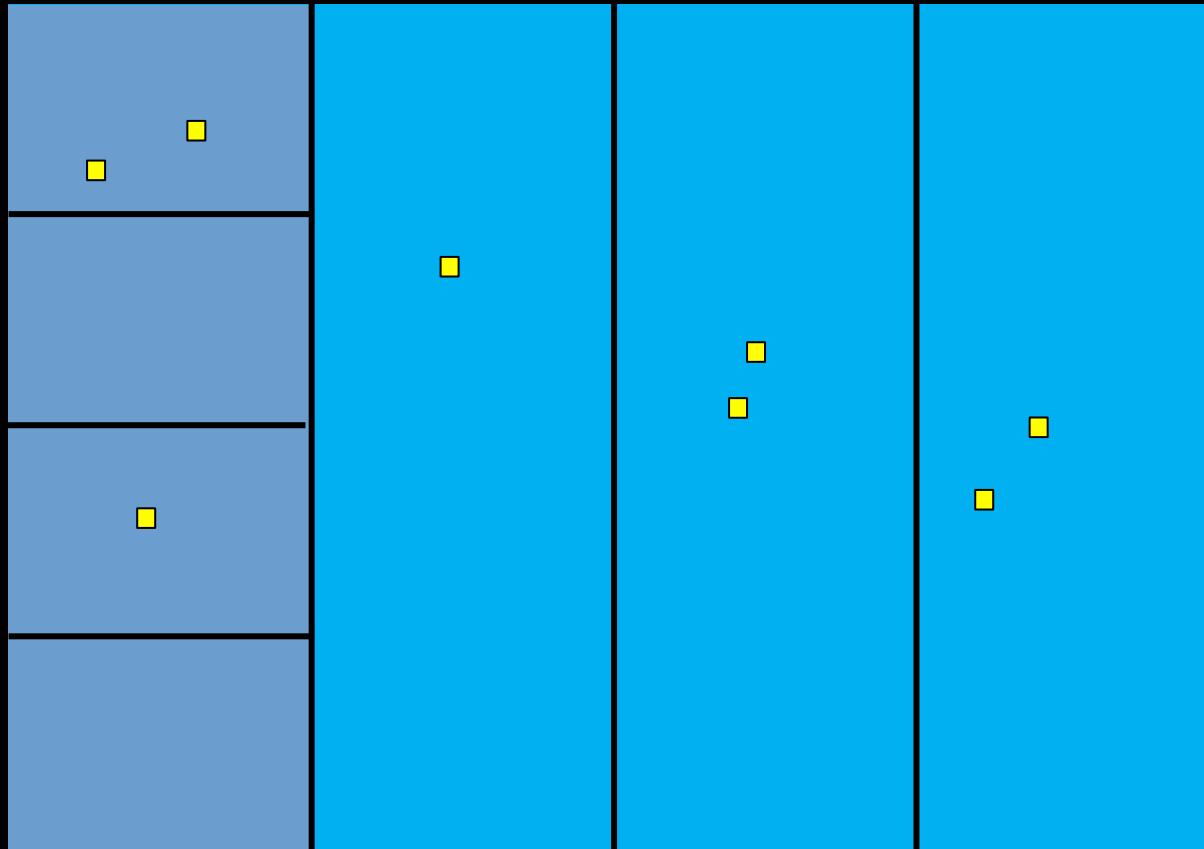
## Apilado Dimensional



Se agregan  $N$  columnas con separación homogénea y se ordenan los datos con respecto al primer atributo.

# Datos Multivariados

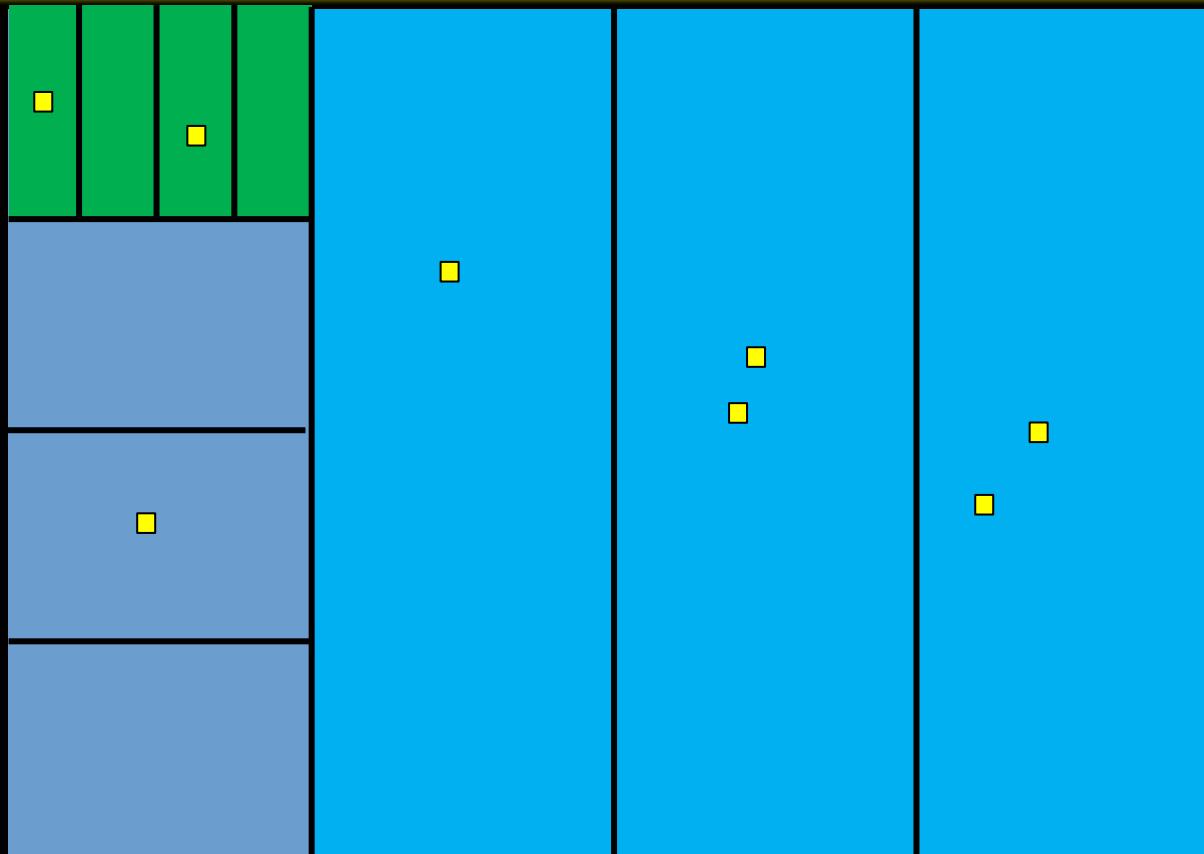
## Apilado Dimensional



Se divide cada columna con  $N$  hileras y los datos se ordenan con respecto al segundo atributo.

# Datos Multivariados

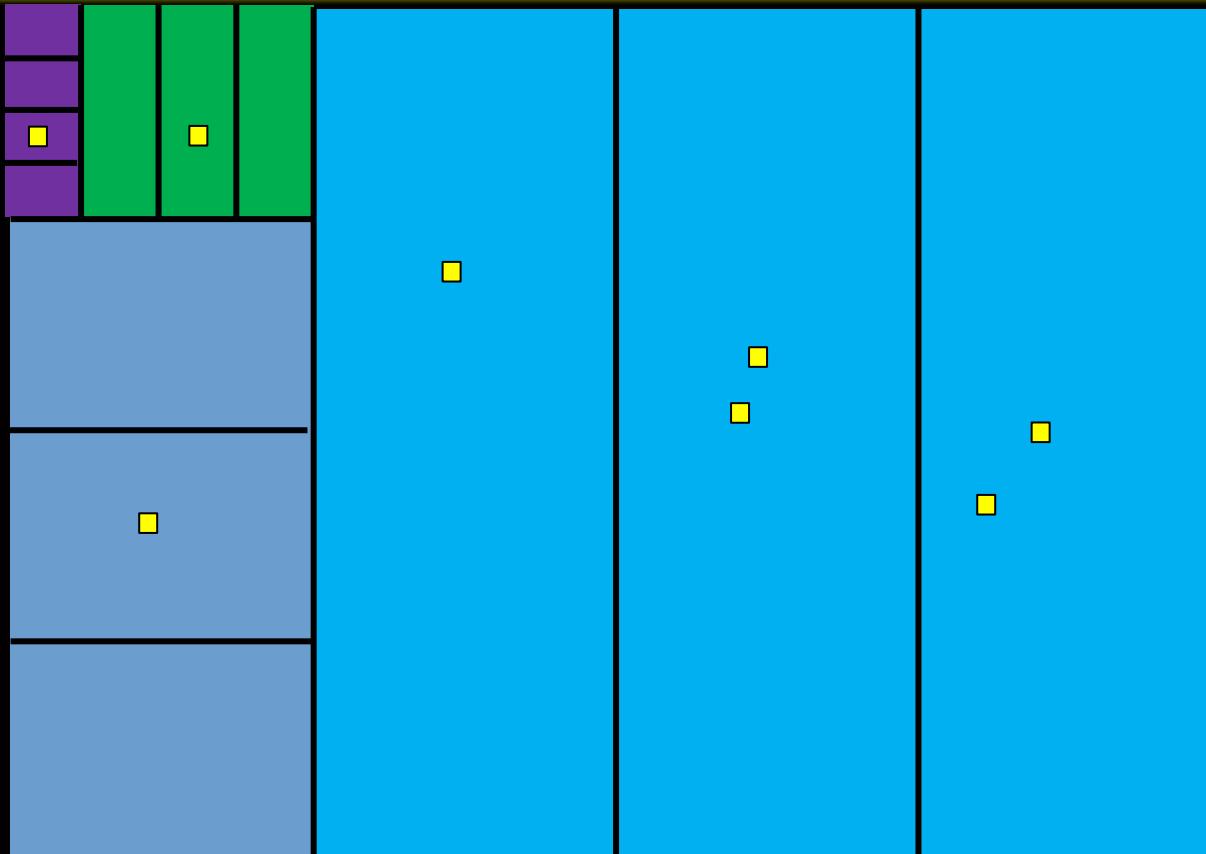
## Apilado Dimensional



Cada celda se divide con  $N$  columnas y los datos se ordenan con respecto al tercer atributo.

# Datos Multivariados

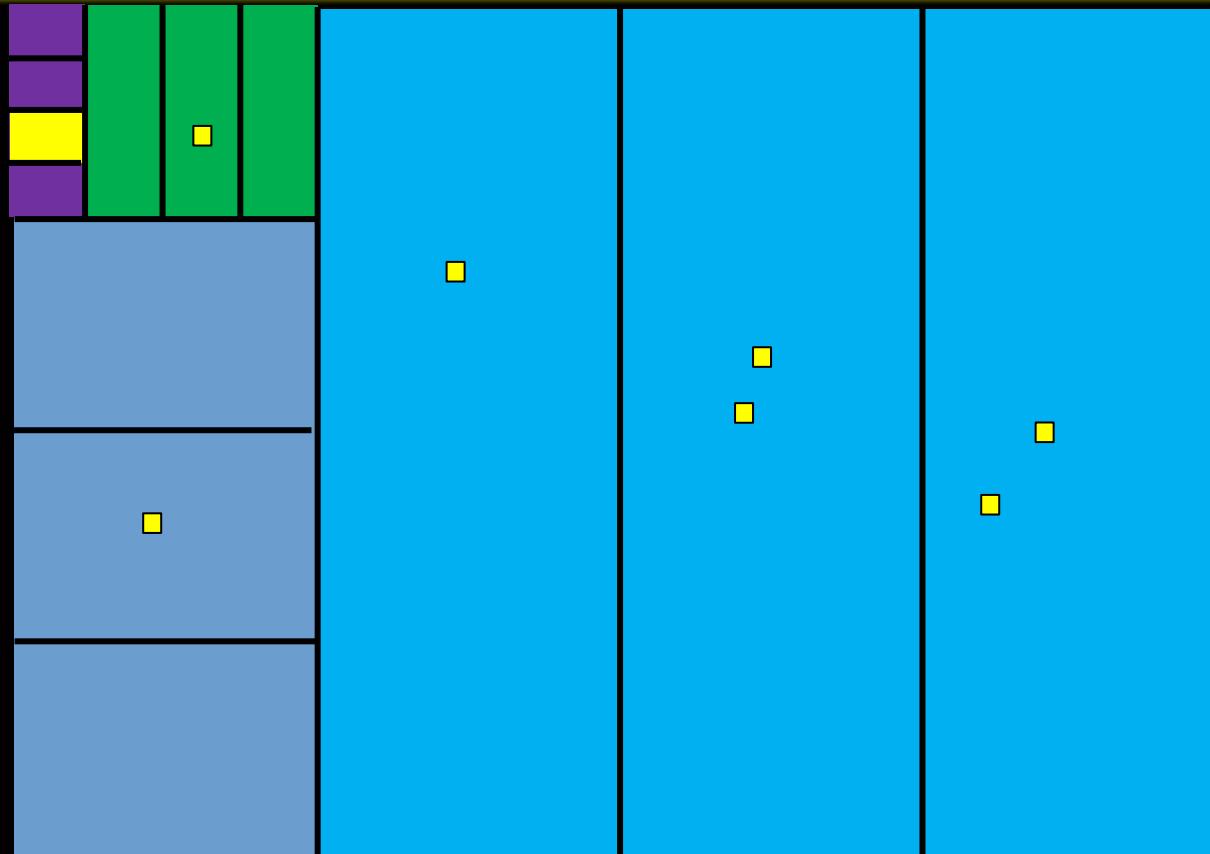
## Apilado Dimensional



La última división en este ejemplo consiste en dividir cada columna nueva con  $N$  fileras y los datos se ordenan con respecto al cuarto atributo.

# Datos Multivariados

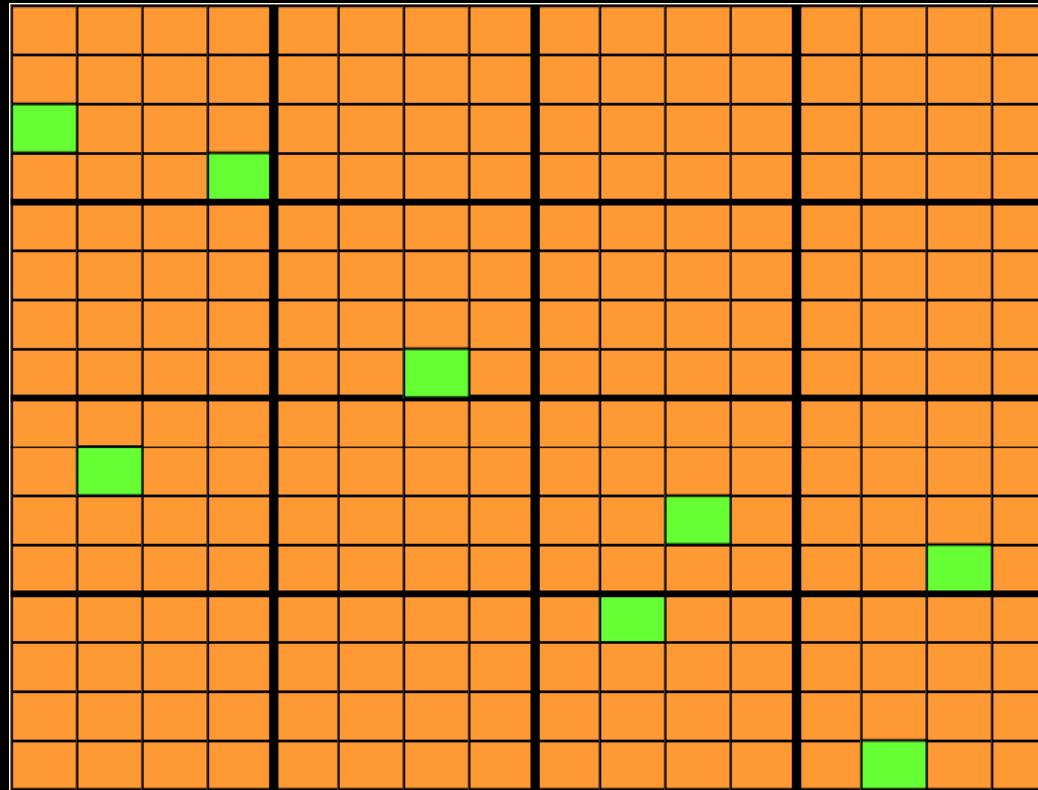
## Apilado Dimensional



Finalmente, el punto se expande para ocupar todo la celda.

# Datos Multivariados

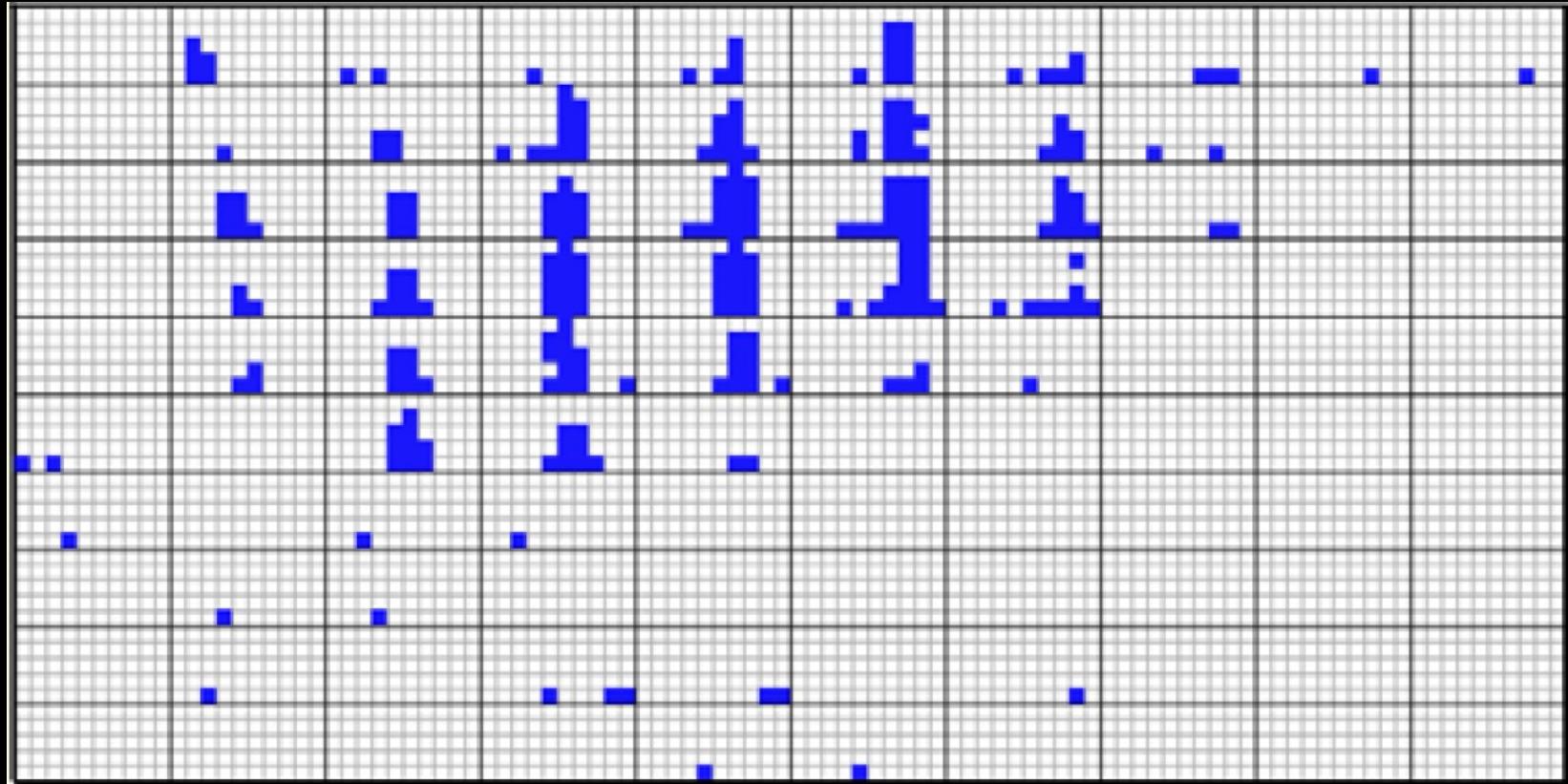
## Apilado Dimensional



Se repite el proceso para todos los datos para obtener el resultado final.

# Datos Multivariados

## Apilado Dimensional



Datos de extracción de petróleo con la longitud y latitud mapeadas sobre los ejes x e y externos, el grado y la profundidad sobre los ejes x e y internos.