TÉCNICAS ESTADÍSTICAS MULTIVARIADAS DE AGRUPACIÓN Y CLASIFICACIÓN

Unidad 1. Técnicas exploratorias de agrupación multivariadas

Esta primera unidad está dedicado a revisar y describir ciertos métodos empleados para examinar los datos de una investigación, especialmente si el objetivo que se persigue es encontrar asociación entre observaciones. Es práctica frecuente entre investigadores y analistas de datos el omitir realizar un análisis cuidadoso de los datos debido al tiempo que se invierte para ello. El análisis previo de los datos permite entender la estructura de los mismos y conocer más a profundidad las características de cada variable bajo estudio con el propósito de determinar la técnica multivariada que permita cumplir con el objetivo planteado en el trabajo de investigación que se realiza. La facilidad de utilizar herramientas computaciones motiva a emplear las técnicas estadísticas más complejas, sin haber adquirido una visión previa de la naturaleza de los datos. Mediante esta exploración previa es posible detectar observaciones aisladas (*outliers*), aplicar algún tratamiento a los datos perdidos (*missing data*) hasta la comprobación de los supuestos requeridos por algunas de las técnicas que se estudiaran más adelante.

Las técnicas multivariadas demandan un esfuerzo importante por parte del analista para la comprensión, interpretación y articulación de resultados basados en relaciones entre variables cuya complejidad está en continuo aumento. La eficiencia de las técnicas multivariadas está directamente relacionada con el volumen de datos disponibles y el cumplimiento de supuestos más complejos que en el caso los análisis univariados. El conocer estas interrelaciones ayuda a refinar el modelo multivariante generado así como ofrecer una perspectiva razonable para la interpretación de los resultados. Además de evitar cometer errores serios cada vez que se aplique una técnica multivariante.

El propósito de esta unidad es proporcionar una visión general de ciertas técnicas gráficas multivariadas que permiten examinar los datos contenidos en la base de datos con la que se está trabajando.

1.1. Introducción al análisis exploratorio de datos

El análisis exploratorio es un proceso diseñado para examinar, dentro de grandes cantidades de datos, patrones consistentes y/o relaciones sistemáticas entre variables; para entonces, validar los hallazgos comparando dichos patrones a la luz de nueva información.

En general hay tres etapas básicas: exploración, construcción de modelos y validación. Idealmente, este proceso debería repetirse hasta identificar un modelo robusto pero, en la práctica, las opciones

para validar el modelo son típicamente limitadas. Por lo tanto frecuentemente el resultado inicial es considerado como la evidencia estadística que podría influir en los procesos de decisión.

El análisis de exploración de datos ha adquirido una creciente popularidad como herramienta de administración de información, de la cual se espera que revele conocimiento estructurado que sea guía en la toma de decisiones, bajo condiciones limitadas de certeza.

Una diferencia general importante en el propósito de la exploración de datos y el análisis tradicional de datos, es que la exploración está más orientada a conseguir aplicar los conocimientos estadísticos en la solución de un problema real, que a caracterizar la naturaleza básica del fenómeno implícito. En otras palabras, la exploración de datos es relativamente menos concerniente a la identificación de relaciones específicas entre las variables involucradas. En vez de eso, el propósito es producir una solución tal que pueda generar predicciones útiles.

El análisis exploratorio de datos multivariados no requiere aplicar modelos rígidos ya que se da importancia al despliegue visual mediante representaciones gráficas. Al realizar un análisis exploratorio se busca:

- Reconocer cualquier patrón no aleatorio o estructura que requiera explicación, y
- Generar posibles hipótesis interesantes

Opuesto a la prueba de hipótesis tradicional, diseñada para verificar una hipótesis a priori sobre relaciones entre variables, el análisis exploratorio es utilizado para identificar relaciones sistemáticas entre variables, cuando no hay (o no completamente) expectativas a priori sobre la naturaleza de tales relaciones.

1.2. Gráficos exploratorios para varias variables simultáneas

En muchas ocasiones el analista puede desear comparar observaciones caracterizadas por más de tres variables, necesitando para ello un medio de representación el perfil multivariante de una observación, tanto si es para propósitos descriptivos como si es un complemento para procedimientos analíticos. Para esta situación se han elaborado varios métodos de gráficos multivariantes por medio de figuras planas. A continuación, se describen algunas de estas técnicas. Un aspecto importante para emplear este tipo de técnicas es la estandarización previa de las variables que serán analizadas.

1.2.1. Gráfico de perfiles multivariados

Los gráficos de perfiles multivariados son gráficos en los cuales se coloca en el eje horizontal cada una de las variables del estudio y en el eje vertical se coloca el valor estandarizo de las mediciones obtenidas en cada una variable de la observación que se está graficando. Dependiendo del paquete computacional empleado se muestran los gráficos de forma separa como en la Figura 1 o todos los

concentra en una misma gráfica. Esto último en caso de tener muchas observaciones puede dificultar la interpretación del mismo. Este gráfico es la forma más simple para representar las mediciones de varias variables.

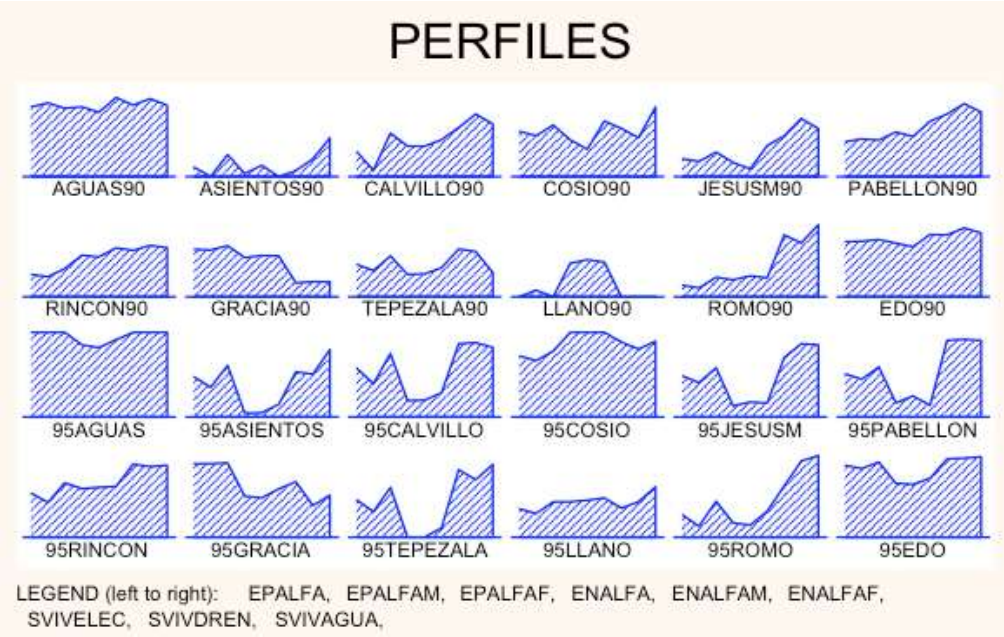


Figura 1 Gráfico de perfiles multivariados

Ejemplo 1.1. Evaluación de automóviles. Una revista especializada en automóviles publicó los resultados de una encuesta realizada a 40 de sus suscritores para valorar 23 modelos de carros. Cada suscriptor evaluó 8 características otorgando una calificación entre 1 (muy bueno) a 6 (muy mala). Las variables son:

x_1: No depreciación del valor

x_2: Precio compra (1 = muy barato, 6 = muy caro)

x_3: Economía

x_4: Servicio

x_5: Diseño

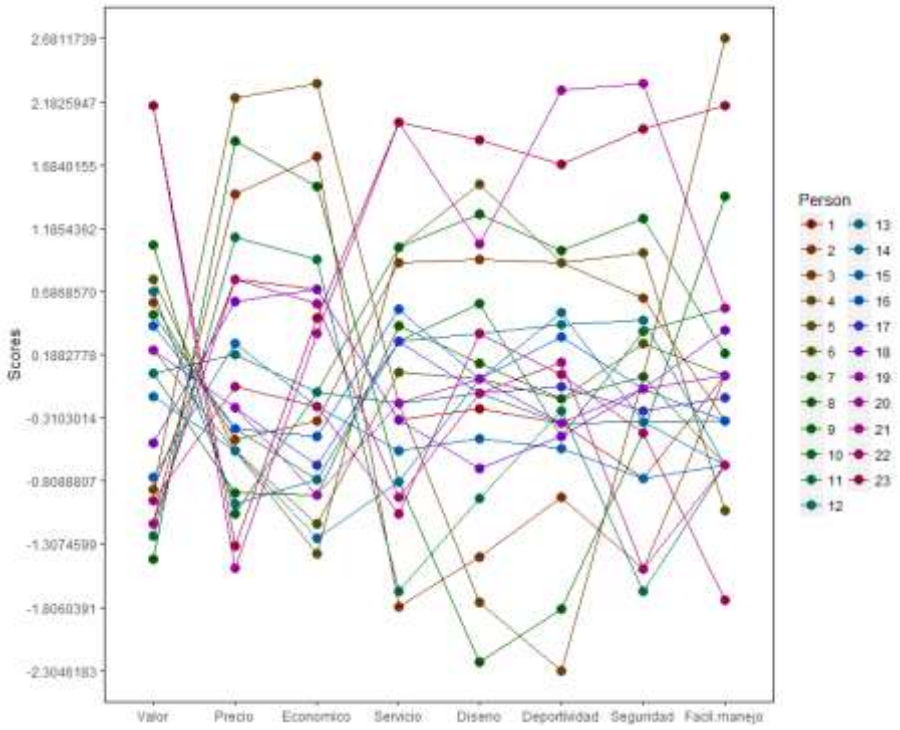
x_6: Deportividad

x_7: Seguridad

x_8: Fácil manejo

```r
library(readxl)
vehiculos <- read_excel("G:/Mi unidad/vehiculos.xlsx")

install.packages("profileR")
library(profileR)
profileplot(vehiculos, standardize = TRUE, interval = 10, by.pattern = TRUE,
            original.names = TRUE)
```



Como se aprecia en el gráfico de perfiles es posible identificar la posición que guarda cada una de las observaciones en cada una de las variables. Es evidente que, a una mayor cantidad de observaciones que son graficadas se complica la visualización del comportamiento de los individuos y por lo tanto su interpretación.

1.2.2. Gráficos de estrella

Al asociamos cada variable observada a un rasgo de una figura plana, es posible representar cada individuo en la muestra por una figura geométrica. En estas representaciones gráficas, las similitudes entre figuras indican las similitudes entre los individuos de la muestra por lo que los valores atípicos aparecerán como figuras discordantes con el resto. Una figura plana muy utilizada para representar

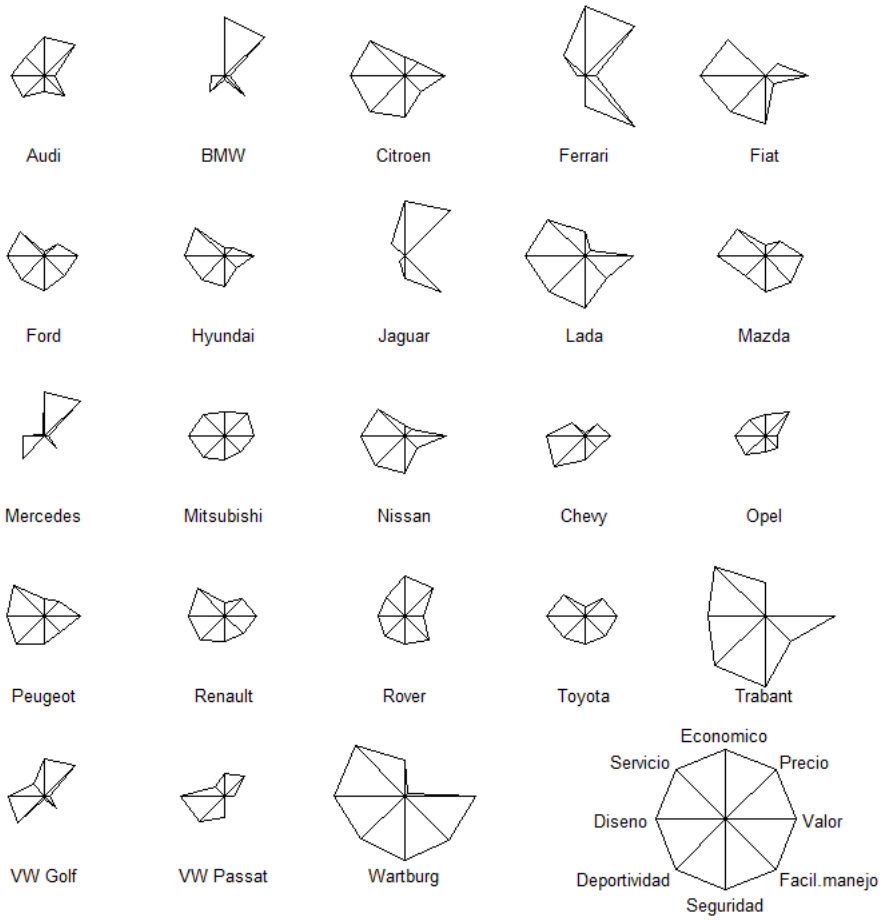
valores multivariantes es la estrella. Las estrellas proveen una mejor representación que los perfiles ya que facilita la comparación entre individuos. Pero al igual que el caso anterior, mientras más crece el número de variables a representar más se complica su análisis. Normalmente las variables se estandarizan de manera que tengan media cero y desviación típica unitaria. Entonces, se marca el cero sobre cada eje y se representa el valor de la variable en unidades de desviaciones típicas.

Por ejemplo, para representar cinco variables, se escoge una estrella de cinco radios y se asocia cada variable a cada uno de estos radios o ejes. Cada observación dará lugar a una estrella como se parecía en la Figura 2.

Figura 2 Gráfico de estrella

Ejemplo 1.2. Evaluación de automóviles (*Continuación*).

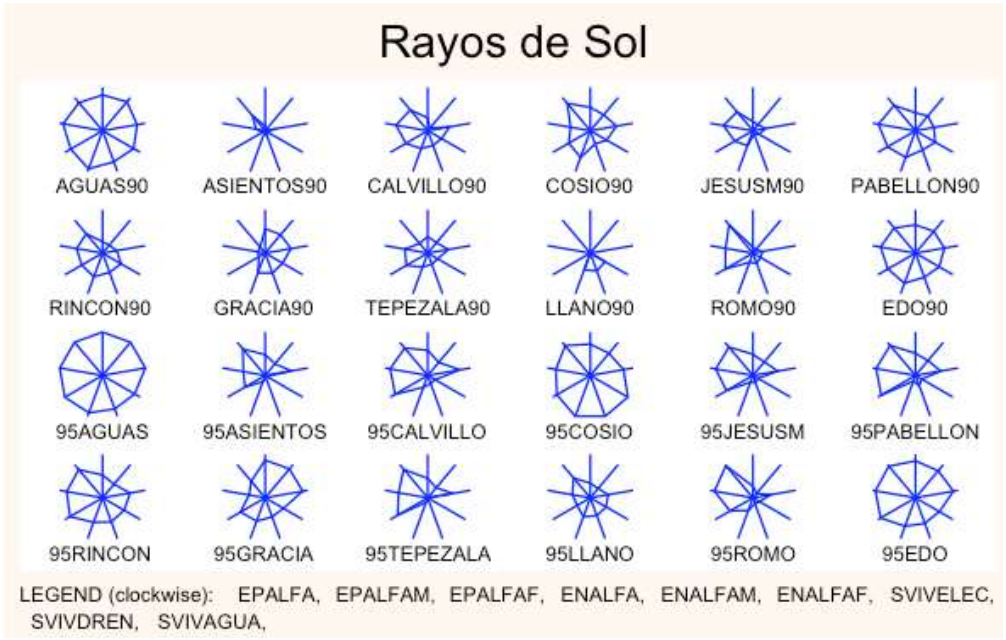
```
install.packages("symbols")
library(symbols)
stars(scale(vehiculos), key.loc = c(11, 2), radius = TRUE,
      flip.labels = FALSE, len = 0.9)
```



Al comparar la figuras (estrellas) del ejemplo anterior se observa que BMW y Mercedes tienen muchas similitudes en cuanto la forma de la estrella y un poco menos de parecido el VW Golf. De la misma manera, Hyundai, Nissan y Citroen, aunque de tamaño diferente el último, las estrellas que los representan tienen un comportamiento similar.

1.2.3. Gráfico de rayos de sol y gráfico de polígonos

El gráfico de rayos de sol (o gráfico de radar) y el grafico de polígonos son parecidos al gráfico de estrellas en cuanto a su construcción e interpretación. En el gráfico de rayos de sol los ejes empelados para representar a cada variable permanecen en la gráfica como referencia mientras en el gráfico de polígonos lo que se obtiene es una superficie poligonal en donde se indica el punto central. Estas dos representaciones gráficas se muestran en la Figura 3.



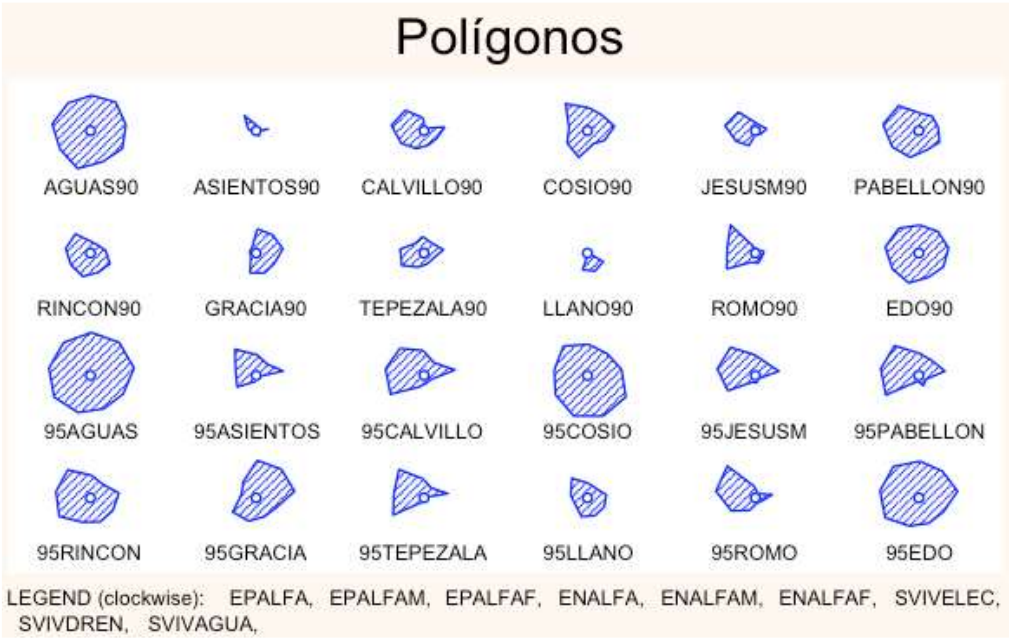
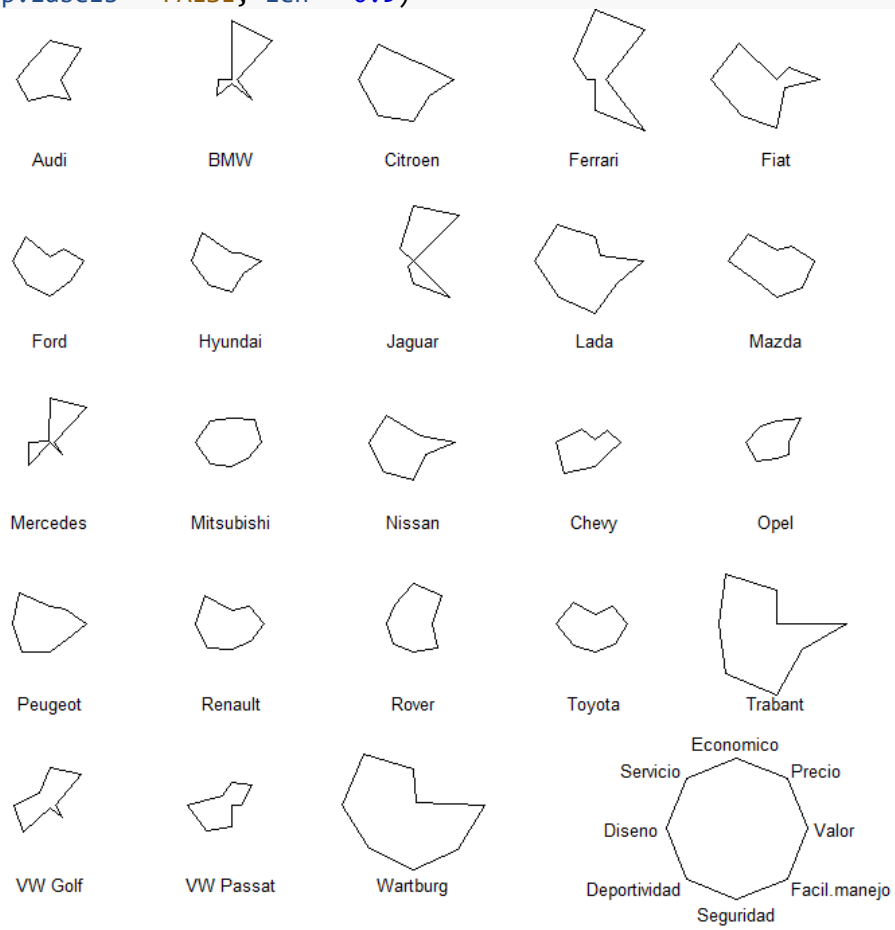


Figura 3 Gráfico de rayos de sol y grafico de polígonos

Ejemplo 1.3. Evaluación de automóviles (*Continuación*).

```
stars(scale(vehiculos), key.loc = c(11, 2), radius = FALSE,
      flip.labels = FALSE, len = 0.9)
```



1.2.4. Transformaciones de Andrew Fourier

Andrews (1972) propuso un método basado en la transformación de Fourier para representar datos multivariantes en dos dimensiones. Básicamente es una representación funcional alterna de senos y cosenos, de cada observación. La transformación se define como

$$f_{\mathbf{X}}(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \cdots$$

Cada variable de cada observación se representa por una componente individual en la suma de la transformada de Fourier. Tradicionalmente t varía entre $-\pi$ y π. La magnitud de cada variable de un sujeto afecta la frecuencia, la amplitud y la periodicidad de f, dando de esta forma una representación única a cada sujeto. Al igual que el gráfico de perfiles si se grafican en una mismo plano todas las

observaciones se hace más notorio las diferencia o similitudes entre ellos. Pero, si la cantidad de observaciones es muy grande, el ojo humano no puede apreciar con facilidad el comportamiento de la función de cada individuo y por consiguiente de complica su interpretación (Figura 4). También es posible graficar cada curva de forma separada.

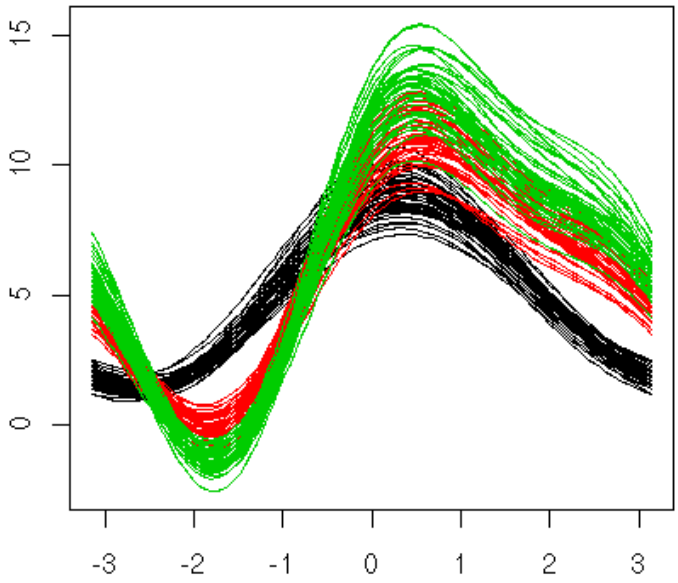
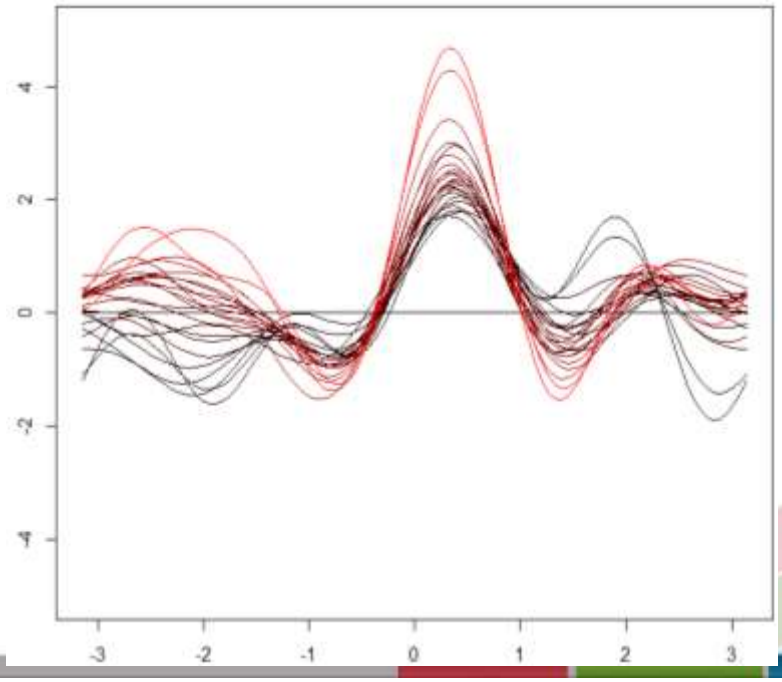


Figura 4 Transformaciones de Andrew Fourier

Ejemplo 1.4. Evaluación de automóviles (*Continuación*).

```
install.packages("andrews")
library(andrews)
andrews(scale(vehiculos), clr = 1, ymax = 5)
```



1.2.5. Caras de Chernoff

En 1973, Chernoff propuso la utilización de caras, tomando ventaja de la facilidad inherente del ser humano para reconocer patrones, es decir, es posible relacionar expresiones faciales con estados emocionales ya que una persona amigable puede representarse por una cara sonriente mientras que una cara triste puede asociarse con alguien depresivo. Usar caras es relevante por:

- Cada uno de los rasgos faciales (ojos, boca, nariz, etc.) son asociados a cada una de las variables o características observadas de los objetos de interés.
- Como humanos, se está acostumbrado a distinguir con cierta precisión caras con diferentes características.

Sobre lo anterior, es necesario tener cuidado la asociación que se haga ya que la boca y la forma de la cabeza son rasgos más llamativos que las orejas o la longitud de la nariz, y el mismo conjunto de datos puede sugerir distintos patrones de similitud entre las observaciones según la asociación elegida entre rasgos y variables. Es práctica común que variables que se infiera relación sean asociadas a un área específica de la cara como los ojos y las cejas. Los rasgos a manipular pueden variar dependiendo el paquete estadístico que se esté utilizando. La Figura 5 representa un ejemplo de esta técnica.

Uno de los mayores inconvenientes de esta técnica es la subjetividad para realizar la agrupación de observaciones al solo guiarse por la apreciación de los rasgos faciales más si el número de variables es muy grande. Además, es prácticamente imposible representar las caras de Chernoff sin el uso de un paquete computacional.

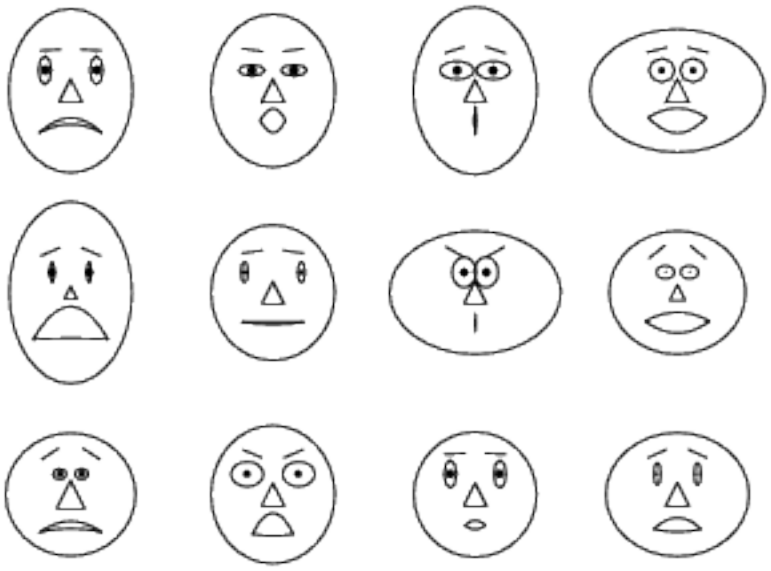
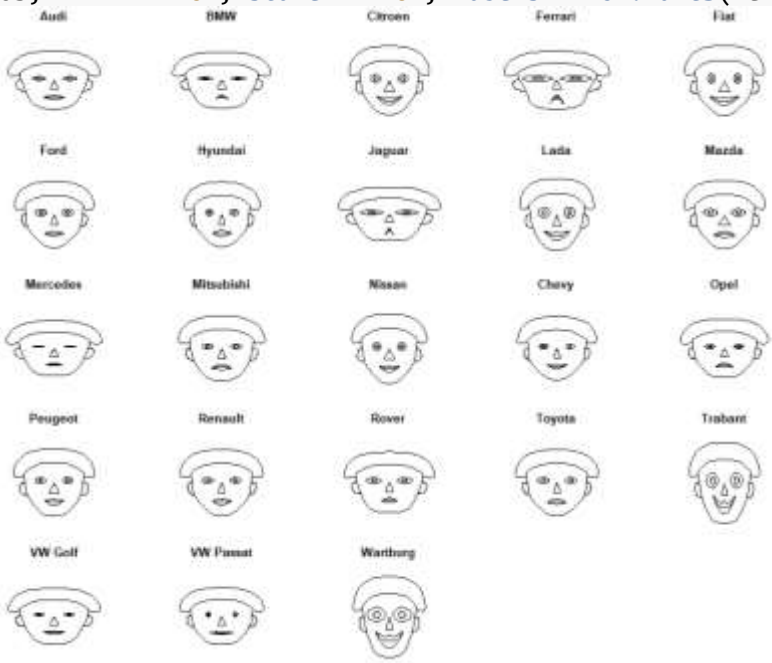


Figura 5 Caras de Chernoff

Ejemplo 1.5. Evaluación de automóviles (*Continuación*).

```
# Caras de Chernoff (opción 1)
install.packages("TeachingDemos")
library(TeachingDemos)
faces(vehiculos, fill = TRUE, scale = TRUE, labels = row.names(vehiculos))
```




```
# Caras de Chernoff (opción 2)
faces2(vehiculos, scale = "center", labels = row.names(vehiculos), cex = 1.5)
```