LiAM Laboratori de Modelització i Anàlisi de la Informació

Exploratory multidimensional statistical methods
An introduction

Anàlisi de Dades i Explotació de la Informació

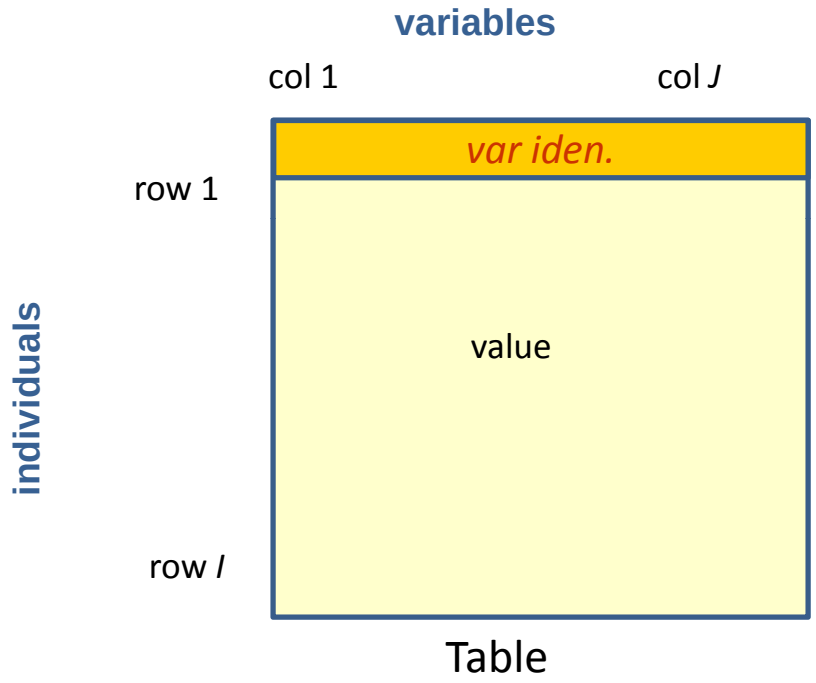
Grau d'Enginyeria Informatica.

Prof. Mónica Bécue Bertaut & Lidia Montero

Monica.becue@upc.edu *lidia.montero@upc.edu*

FIB UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Coding the data into rectangular tables



Information is stored in tables

Rows of the table

Individuals or instances , sample, example, record, …

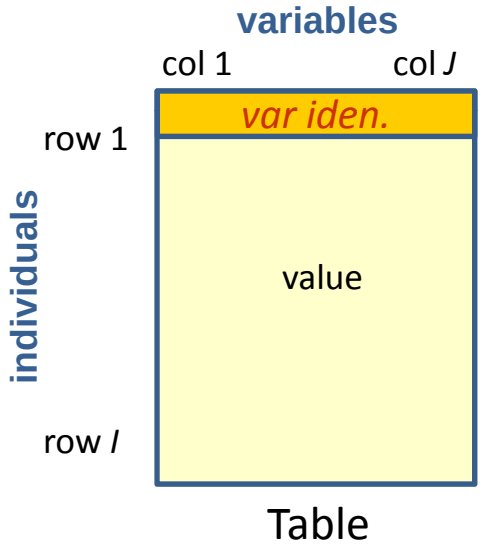
forming the sample under study, extracted from a population

Columns of the table

Variables or “attributes”.

Main attribute types : **quantitative**, binary, **categorical**, **ordinal**, interval, ratio, **textual**, …

Variables/ Attributes observed on the individuals or constructed *a posteriori*

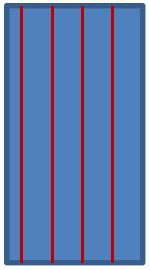


Encoding the variables

- **Continuous or quantitative**

One column

- **Categorical**
  - Binary (yes/no variable, Gender)
  - Categorical (marital status, region, ….)
  - Ordinal (health level)

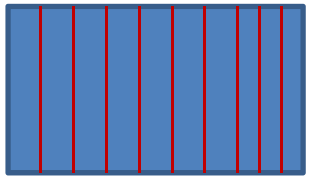


- **Frequency data**
  - Series of columns with count data that must be treated as a whole

as many columns as categories

(counts of the occurrences of the different words used to answer an open-ended question; counts of mortality data by causes; counts of the occurrences of all the different species present in an ecological site)

- Textual data



as many columns as species

Role of variables

- **Response**

  Variable to be explained or predicted

  either quantitative, categorical or frequency

- **Explanatory**

  Variables used to explain the behaviour of the response variables

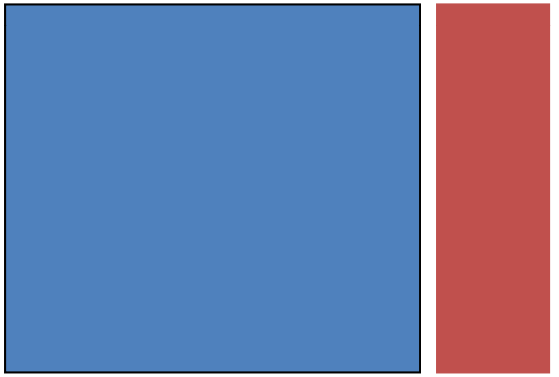
  quantitative, categorical or frequency

Types or data matrix

With or without response(s) variable

i.e. transactions data

Inputs Output(s)



Data that will be explores, describer in order to find associations (i.e. itemsets), patterns, groups of individuals, etc.

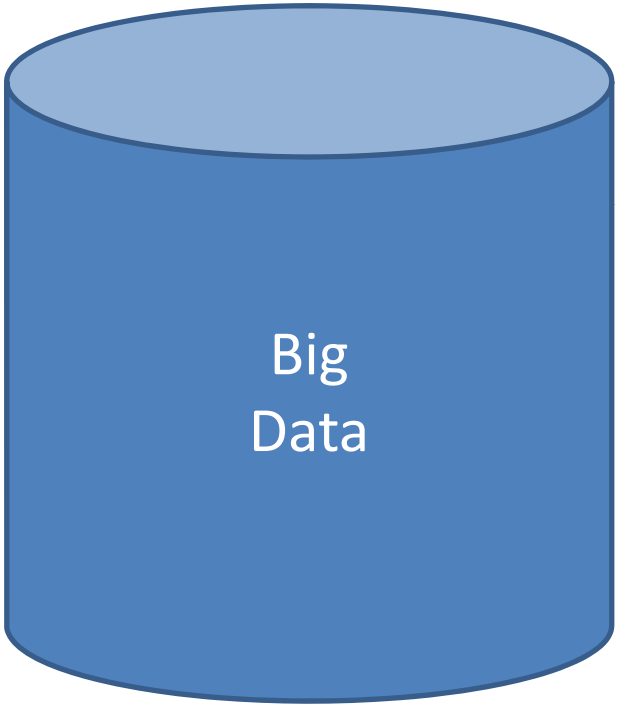
we want to **find a model to predict or explain the response**

R function available in FactoMineR

We will use FactoMineR Package (cran R)

You can also consult (and download this R function from) http://factominer.free.fr/ where a large documentation is provided, with theoretical background, examples, tutorials and so on.

Some details about the exploratory methods



Big data require
Exploratory multidimensional statistical methods

JOB: Data Scientist Technology – Business Intelligence | Mind Candy, London
Mind Candy – Posted by Advertiser – **London**, England, United Kingdom
Job Description
We are looking for truly talented individuals to become an integral part in driving the (…)
The Role
Due to our continued success Mind Candy is rapidly expanding and we have a truly fantastic opportunity for a Data Scientist to come on board and play a key role in (…)
Minimum Requirements:
Good business and technical skills in data analytics. Technical skills must include:
Highly proficient data mining skills in **small and very large data sets.**
Great ETL skills using a variety of languages (e.g. SQL, **R**, Python, Scala) and **big data tools** (e.g. Hive, Scalding, Pig, Elastic MapReduce).
Great statistical skills and a **passion for data** and **data visualisation**.
Ability to continuously adapt to the data needs in a rapidly changing environment. This would include quickly and efficiently integrating new data sources using various methods (from internal or external databases, using REST API, etc.).
Experience and confidence in gathering business requirements from the product teams and delivering reports, analysis and innovative, fit for purpose information solutions.
Experienced in managing your own priorities based on business goals.
Strong communication skills.
Experience of working in an Agile environment.
Preferred Requirements: (…)

Resumen de los tipos de descripciones de Big Data que Ward y Barker han descubierto de varias organizaciones influyentes:

1. **Gartner.** En 2001, un informe de Meta (hoy día Gartner) tomó nota del aumento del tamaño de los datos, la tasa de aumento a la que se producen y la creciente **variedad de formatos** y representaciones empleadas. Este informe es anterior a la expresión 'big data', pero proponía una definición triple con tres 'V': **volumen, velocidad y variedad.** Desde entonces, esta idea se ha hecho muy popular y, a veces, incluye una cuarta V: veracidad, para cubrir la cuestión de la confianza y la incertidumbre.

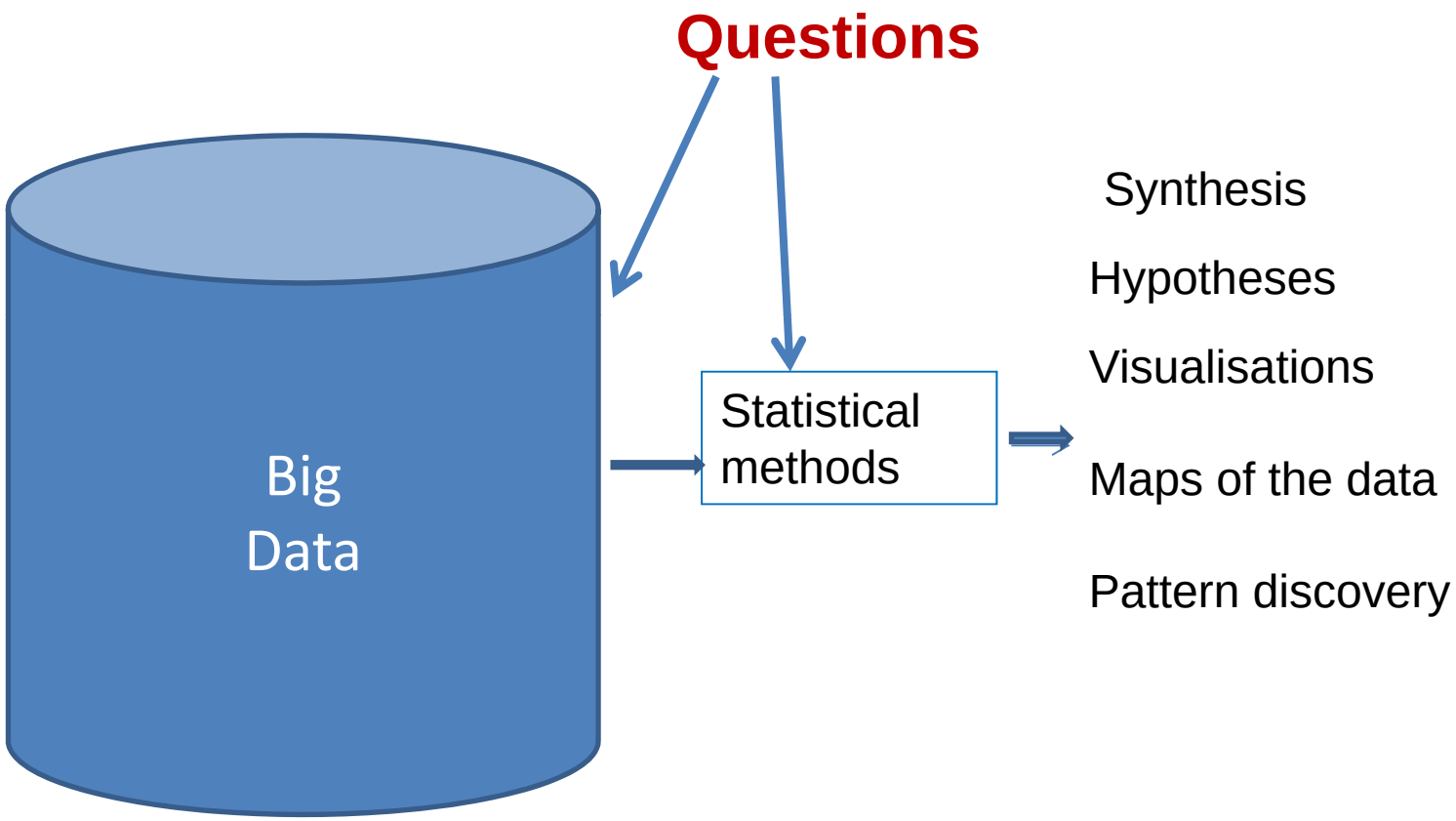
2. **Oracle.** 'Big data' es la derivación de valor a partir de la toma de decisiones de negocio en función de bases de datos relacionales tradicionales, aumentada con nuevas **fuentes de datos no estructurados**.

3. **Intel.** Las oportunidades de trabajo con grandes volúmenes de datos surgen en organizaciones que generen un **promedio de 300 terabytes de información** a la semana. La clase de datos más común es la de las transacciones comerciales almacenadas en bases de datos relacionales, seguida de **documentos**, correo electrónico, datos de sensores, **blogs y redes sociales.**

4. **Microsoft.** "**'Big data'** es un término cada vez más utilizado para describir el proceso de aplicación de una significativa potencia de computación (lo último en el aprendizaje de máquinas e inteligencia artificial) a conjuntos de información de **enorme tamaño** y, a menudo, de **alta complejidad**".

5. El proyecto de código abierto **MIKE** (siglas en inglés de **Method for an Integrated Knowledge Environment**). El proyecto MIKE argumenta que l**os grandes volúmenes de datos no tienen que ver con el tamaño sino con la complejidad.** Por consiguiente, lo que define un conjunto de datos como 'big data' es su alto grado de permutaciones e interacciones.

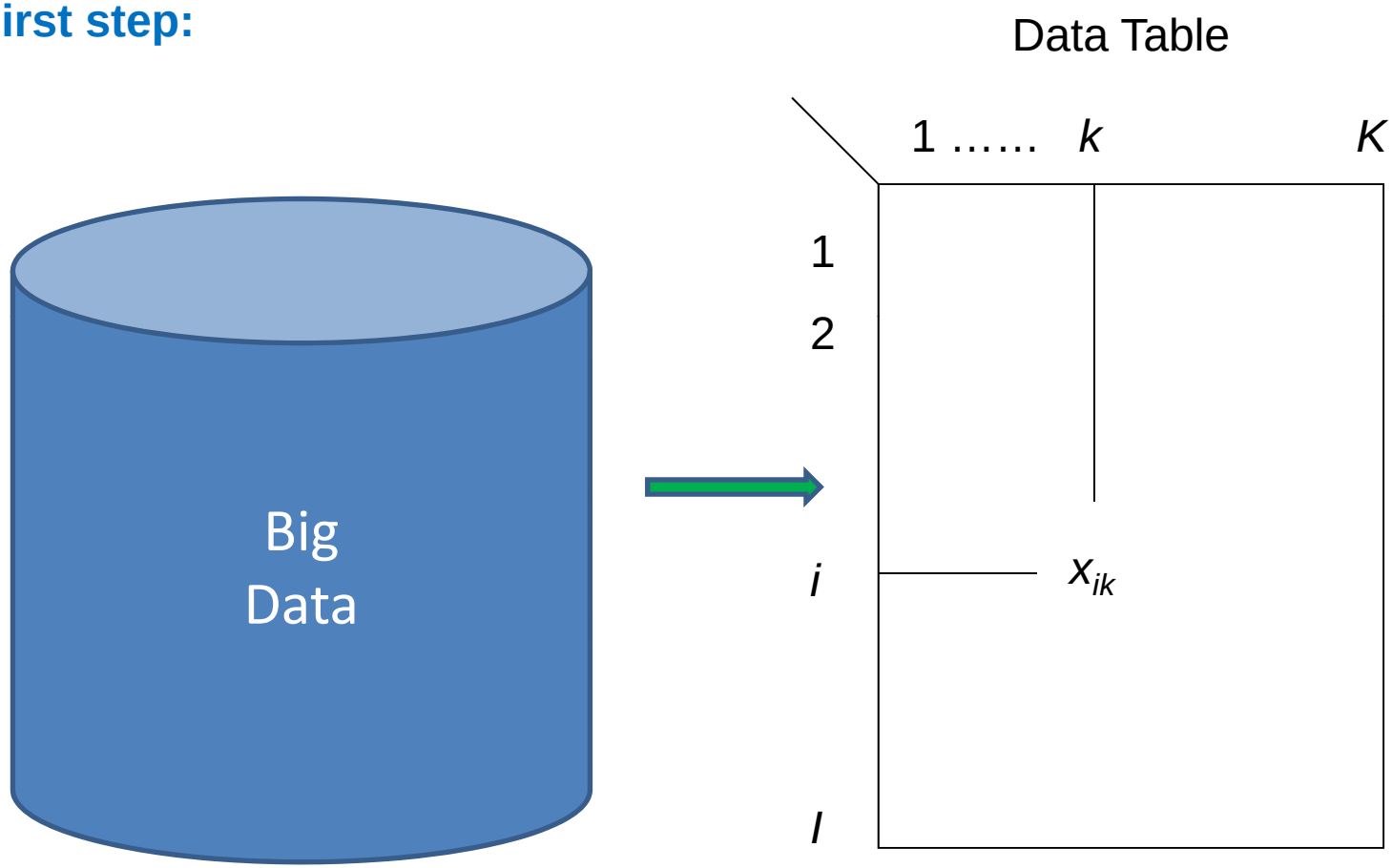
6. El **Instituto Nacional de Estándares y Tecnología** de EEUU. El Instituto afirma que los grandes volúmenes de datos se refieren a aquellos que "superan la capacidad o la habilidad de los métodos y sistemas actuales o convencionales". En otras palabras, la noción de 'grande' está relacionada con el estándar de computación actual.


Questions
Big Data
Statistical methods
Synthesis
Hypotheses
Visualisations
Maps of the data
Pattern discovery

Among the Statistical methods: **multidimensional exploratory statistical methods**

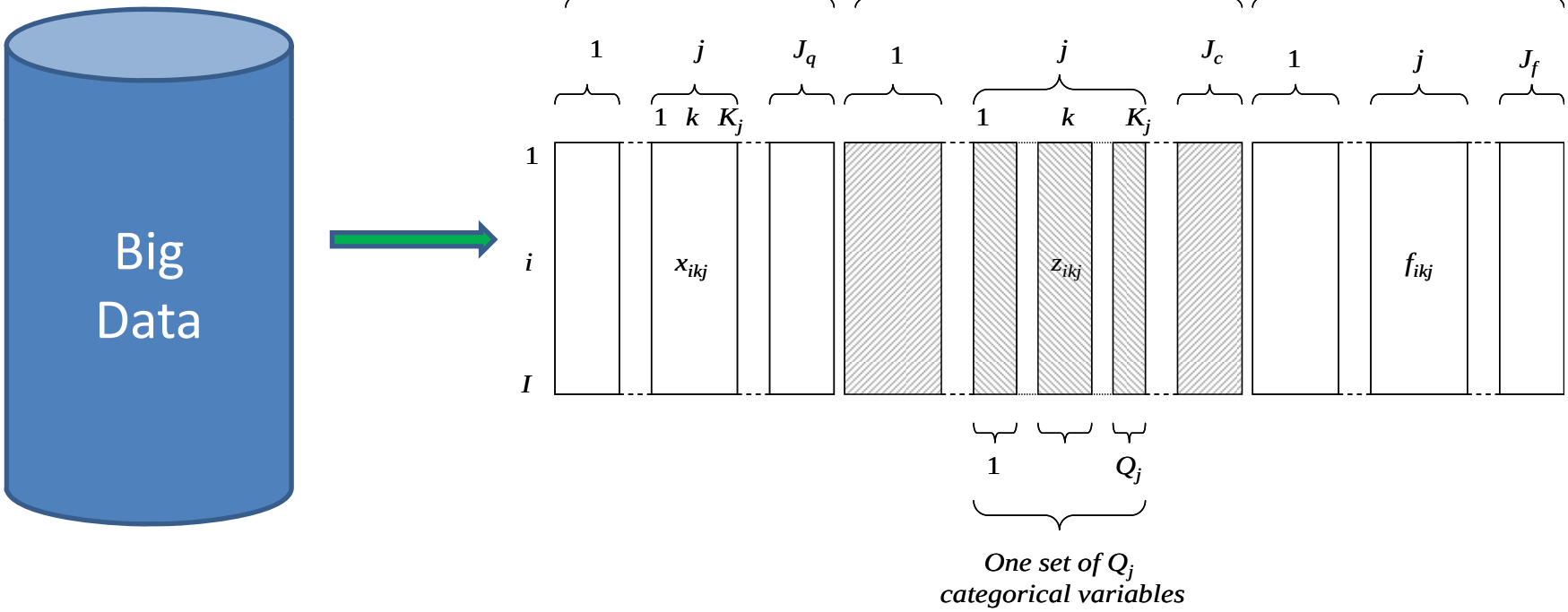
What are they for?

First step:



It tabular data led to data analysis, it can also be pointed out that tabular data led to the computer (Fionn Murtagh *Electronic Journ@l for History of Probability and Statistics*- Vol 4, n°2; Décembre/December 2008 **www.jehps.net**

Or First step:



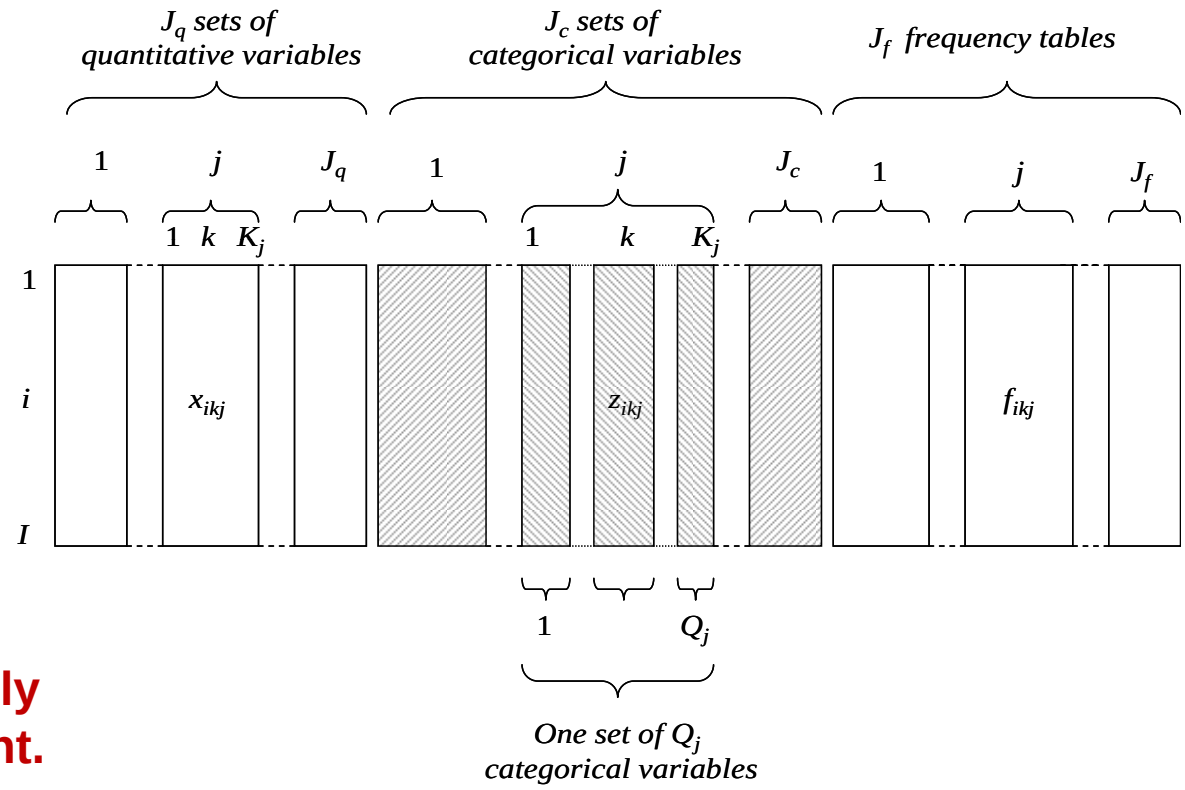
Or First step:

Big Data

Data Coding

Extremely important. Coding conditions the results

Multiple Table



And now exploratory data analysis gets going!

To answer the questions

"In data analysis numerous disciplines have to collaborate.
The role of mathematics, although essential, remains modest in the sense that classical theorems are used almost exclusively, or elementary demonstration techniques.

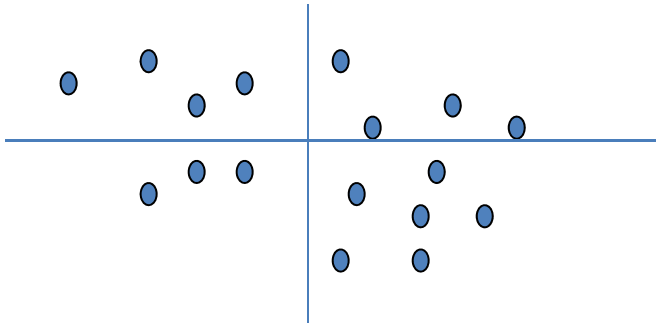
But it is necessary that certain abstract conceptions penetrate the spirit of the users, who are the specialists collecting the data and having to orientate the analysis in accordance with the problems that are fundamental to their particular science."

Fionn Murtagh
Electronic Journ@l for History of Probability and Statistics
Vol 4, n°2; Décembre/December 2008
www.jehps.net

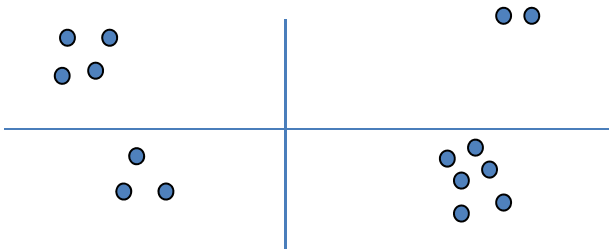
Among the Statistical methods: **multidimensional exploratory statistical methods**

What are they for?

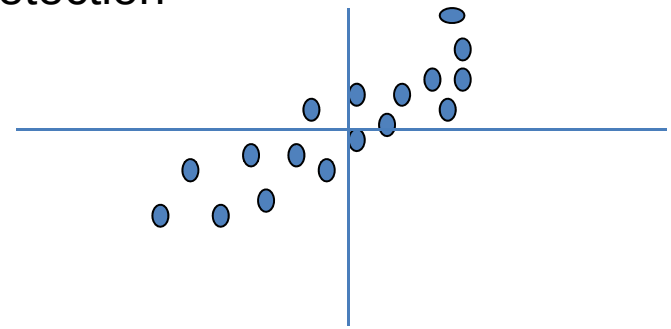
Visualisation of data



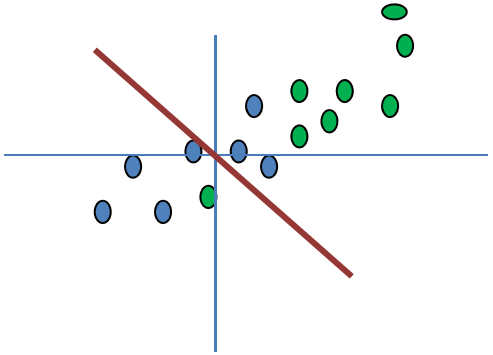
Mapping the individuals from their similarity



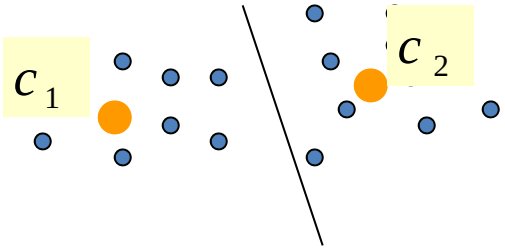
pattern detection



For classifying the individuals
Supervised classification



or clustering them
non-supervised classification



Examples

1. Method: PCA

Ejemplo: cata de chocolates

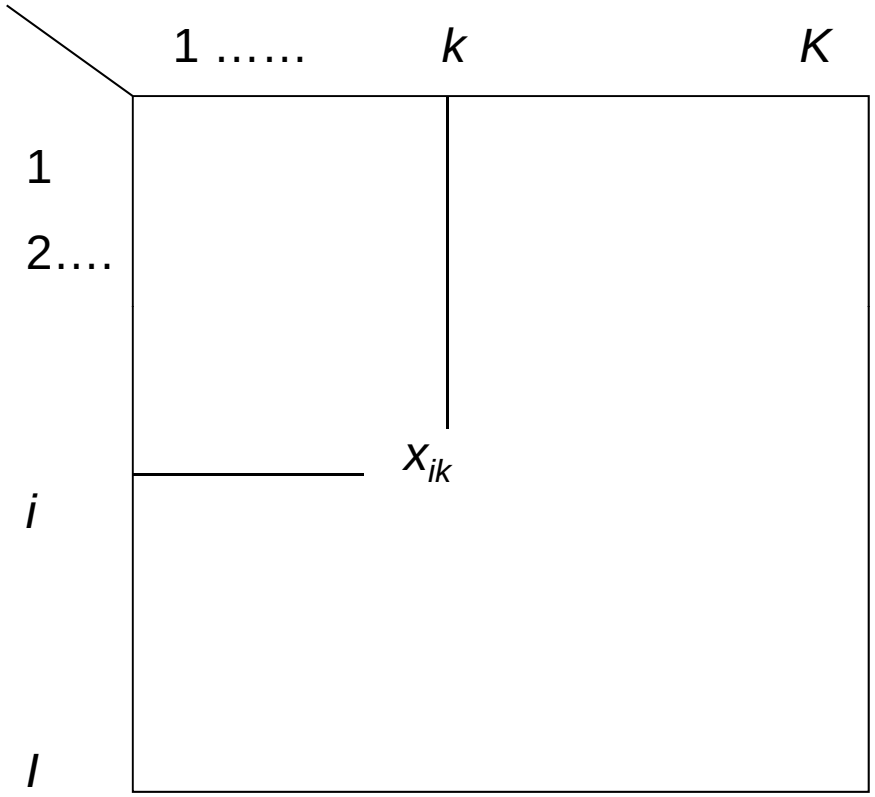
Ejemplo: Chocolates

- 10 chocolates negros (individuos)
  - 3 marcas: Lindt, Valrhona y Hacendado
  - Porcentaje de cacao entre 55% y 85%
- Método de recogida de datos: QDA (Quantitative Descriptive Analysis)
- 16 panelistas-jueces y 2 sesiones
- 14 descriptores o atributos
  - Olor: cacao, leche
  - Sabor: azúcar, ácido, amargo, cacao, leche, caramelo, vainilla
  - Características: astringencia, crujiente, fusión en la boca, pegajoso, granuloso
- Notas entre 0 y 10

- Diseño de experimentos completo balanceado para los rangos y efectos de arrastre de orden 1

23

Building the products × attributes table



Objetivos de ACP

Análisis de los individuos-chocolates (A)

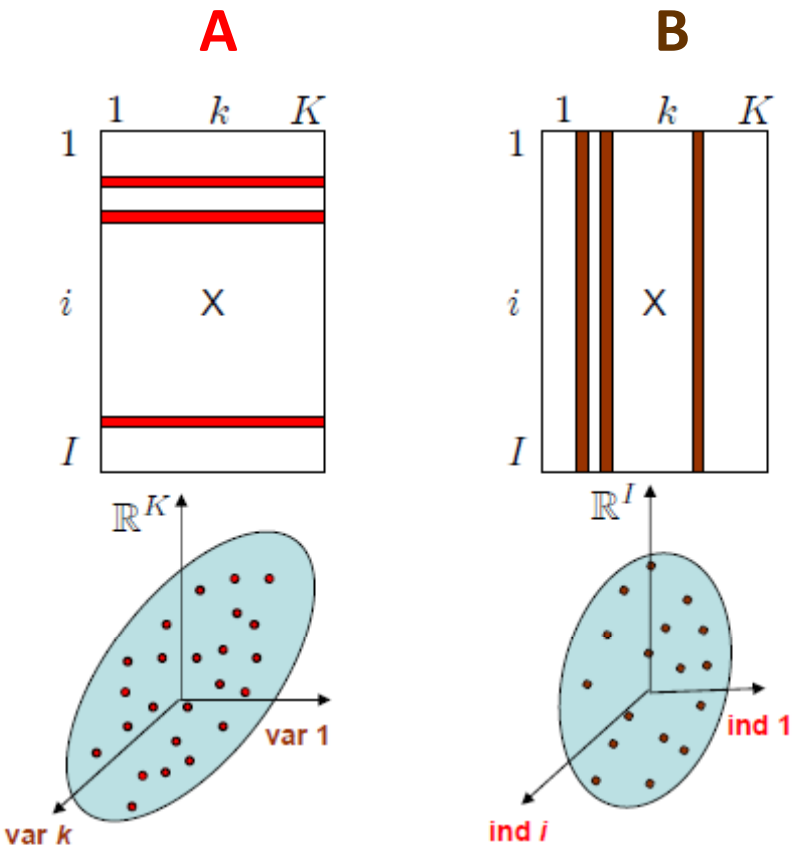
Similaridad entre los individuos respecto a todas las variables

→ partición entre los individuos

- ## Análisis de las variables-atributos (B)

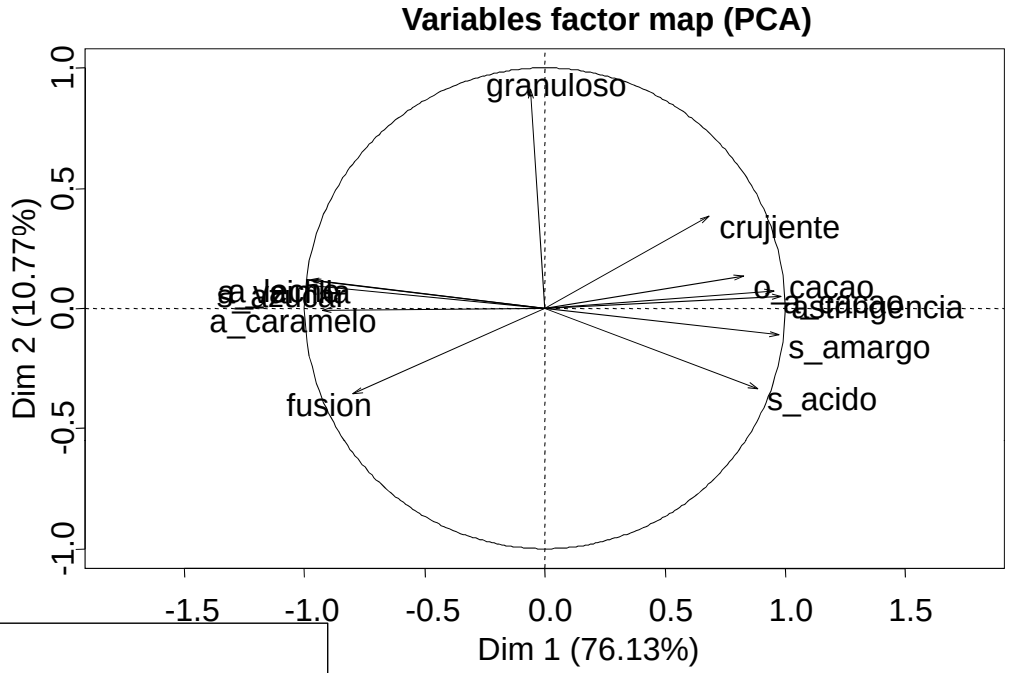
  Similaridad entre los atributos (correlación fuerte)

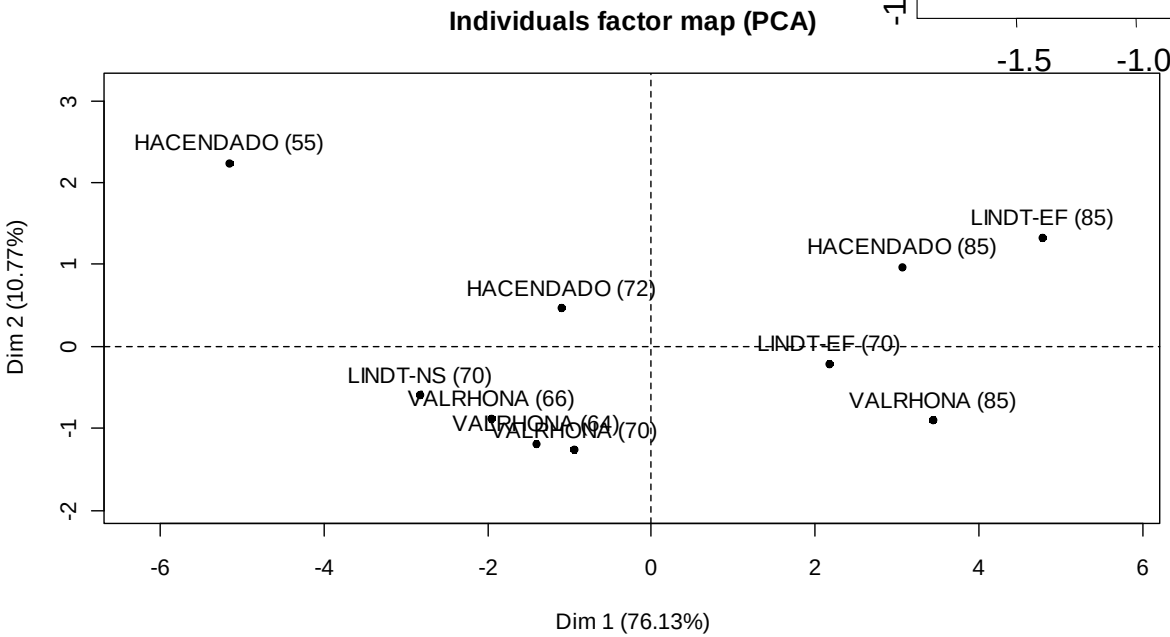
Relación entre los dos análisis



25

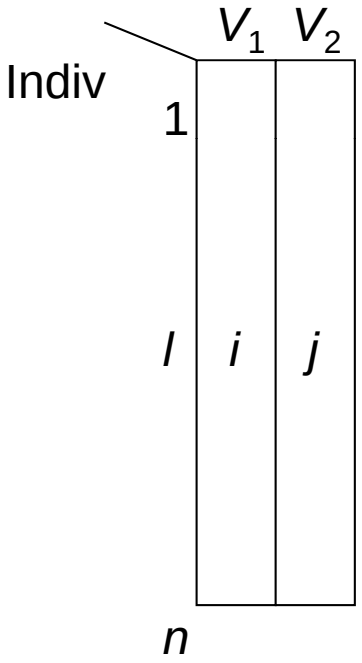
Analysis by PCA



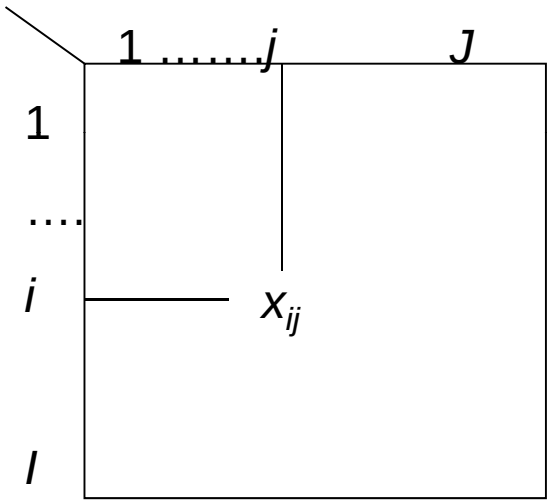


2. Method: CA

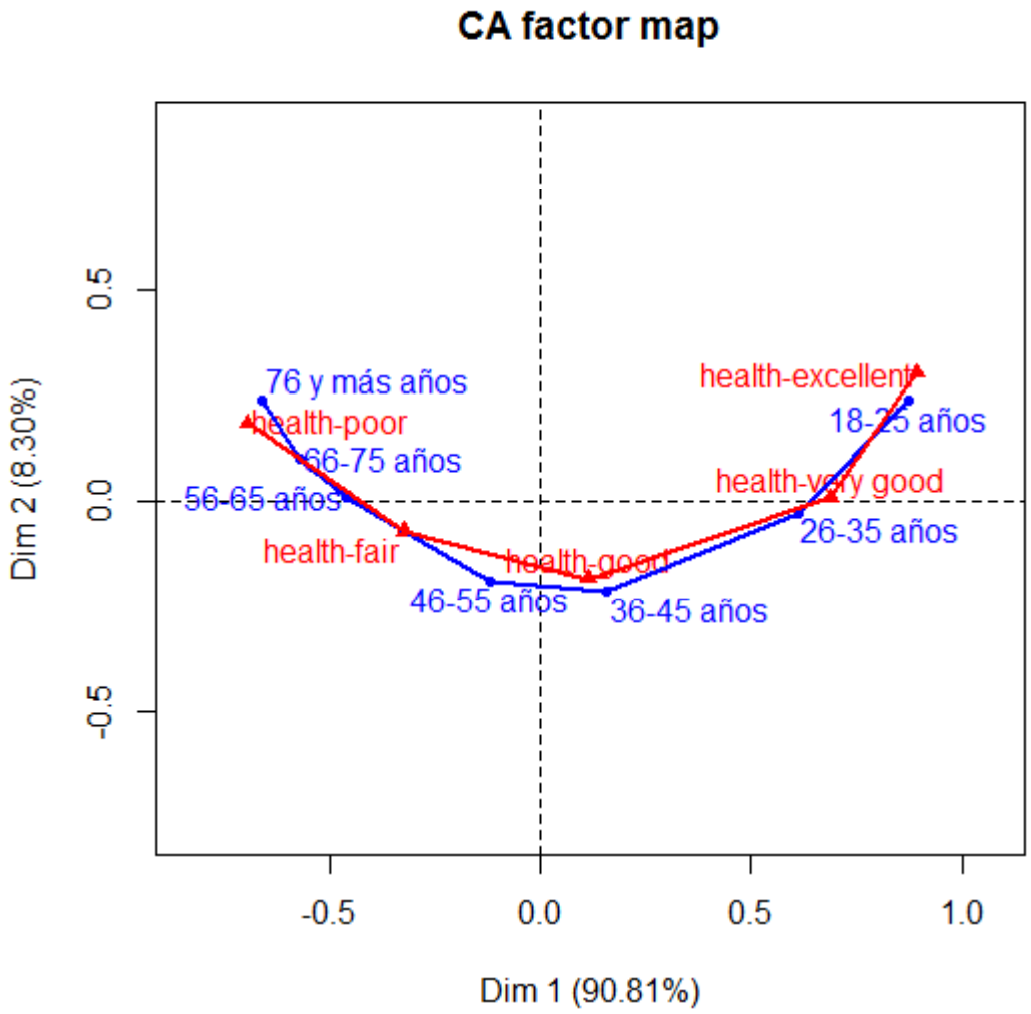
Relationship between categorical variables

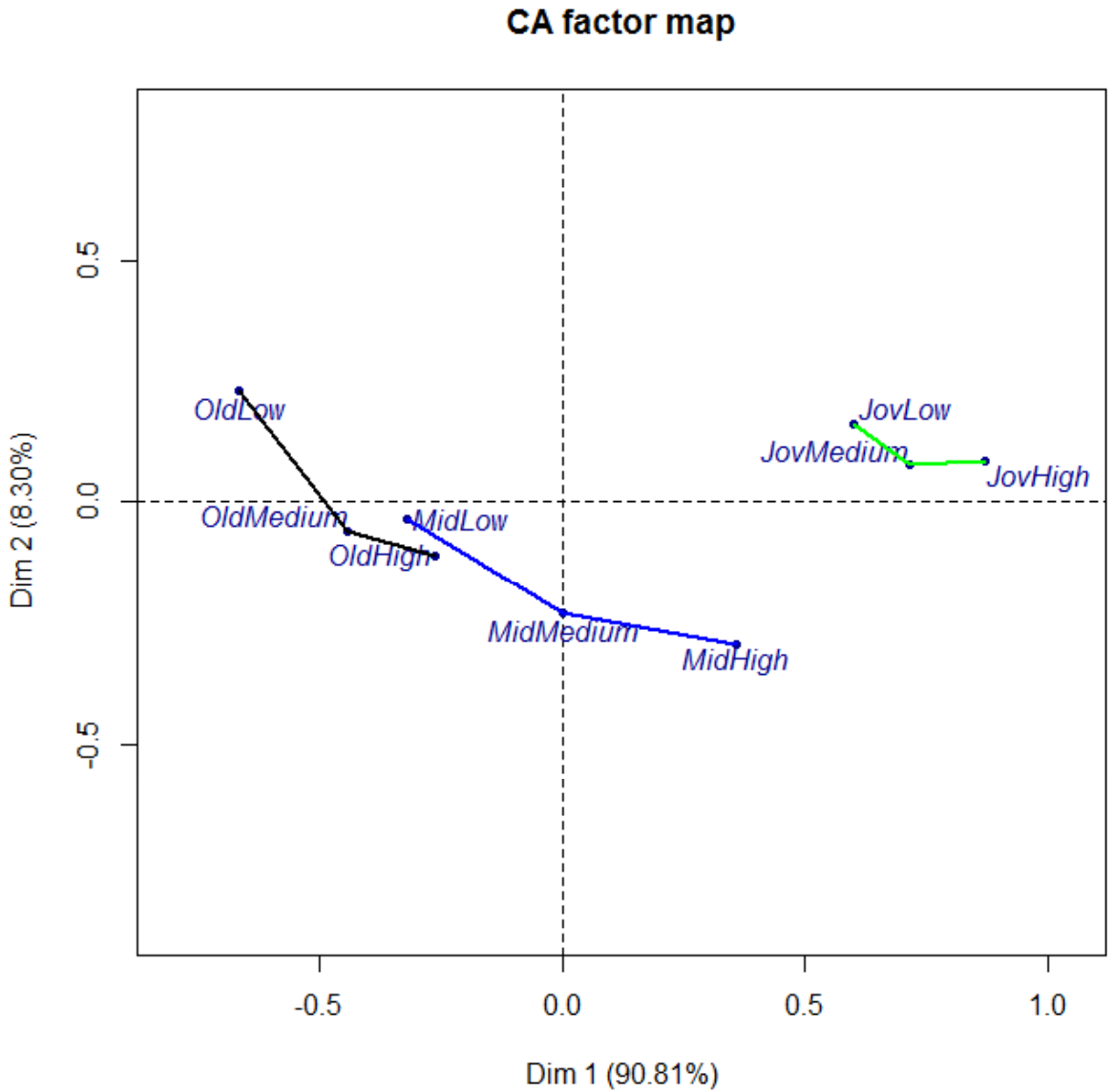


Contingency table



x_{ij} : respondents who present category i of V_1 and category j of V_2


CA factor map
76 y más años
health-poor
66-75 años
56-65 años
health-fair
46-55 años
health-good
36-45 años
26-35 años
health-very good
18-25 años
health-excellent
Dim 2 (8.30%)
0.5
0.0
-0.5
-0.5
0.0
0.5
1.0
Dim 1 (90.81%)


CA factor map
OldLow
OldMedium
OldHigh
MidLow
MidMedium
MidHigh
JovLow
JovMedium
JovHigh
0.5
0.0
-0.5
Dim 2 (8.30%)
-0.5
0.0
0.5
1.0
Dim 1 (90.81%)

Textual Analysis

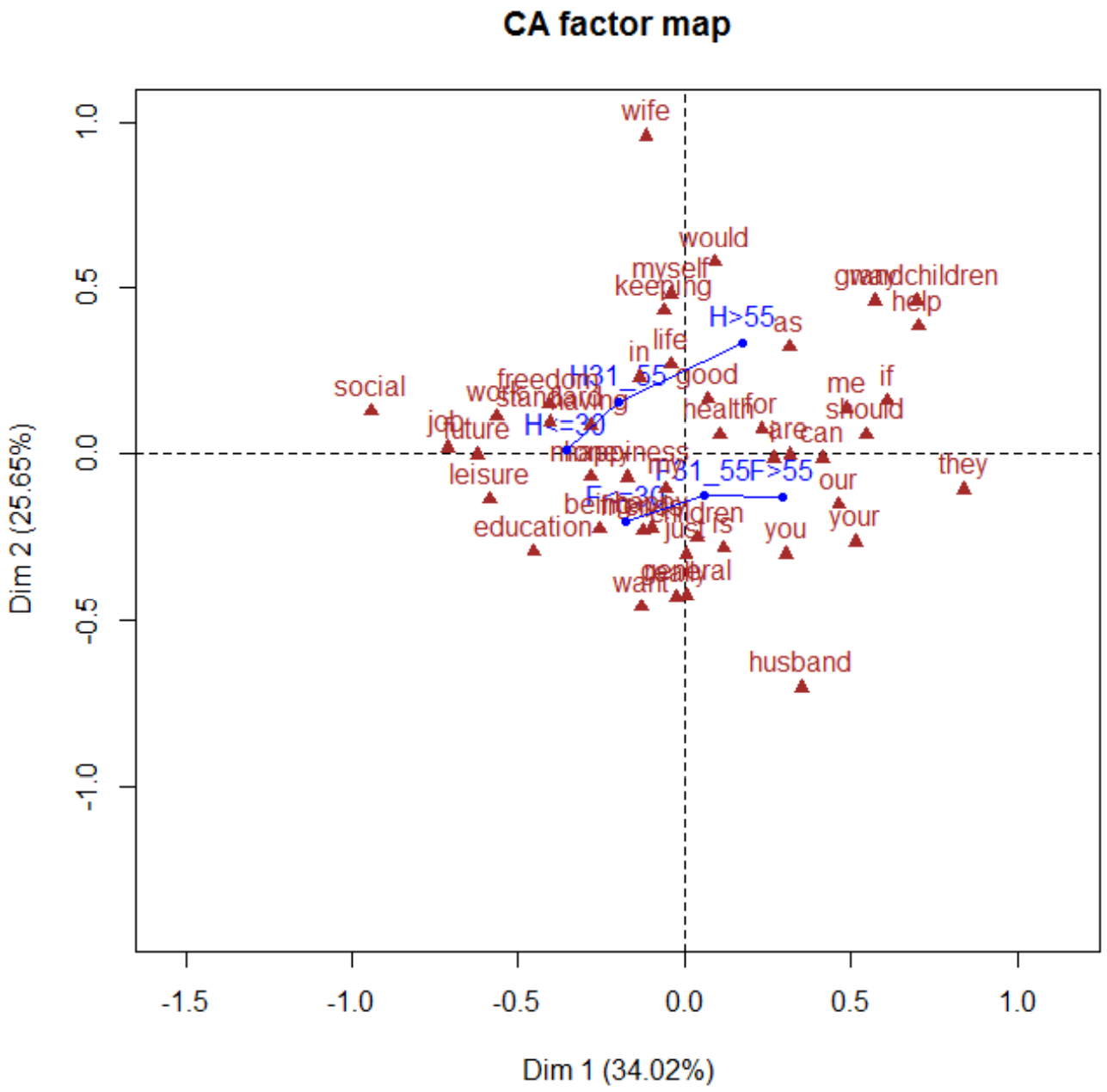
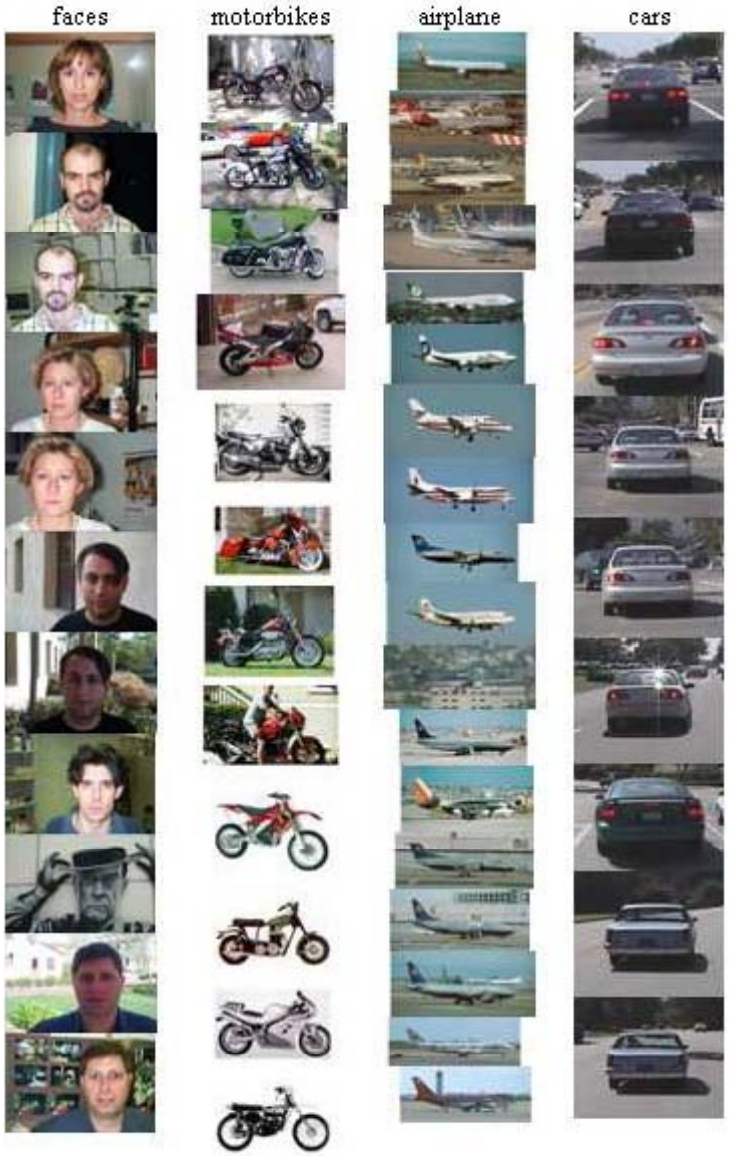


Image Analysis

Image data base



Coding the images into “visual words”

CA is able to “organize” the image data base

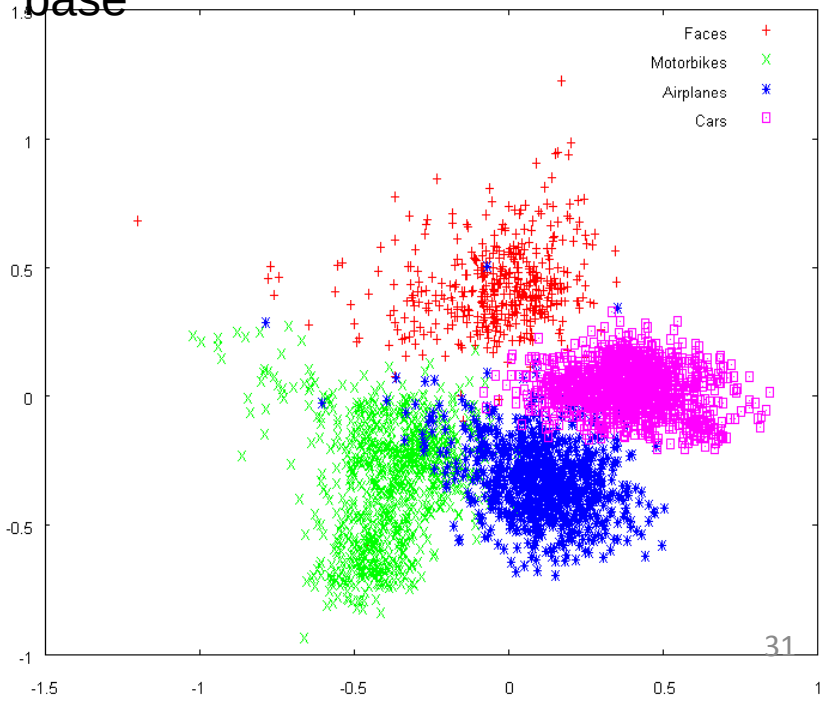


Figura 9.3.1.2. Resumen grupos en función de su nivel de redención.

MCA + clustering

SENSIBILIDAD AL PRECIO ALTA

SENSIBILIDAD AL PRECIO MEDIA

SENSIBILIDAD AL PRECIO BAJA

SEGMENTO 2 (27.02%)

-**Preferencia por productos baratos.**

-Segmento con mayor sensibilidad al cupón.
- Gasto medio-bajo.
-Número de visitas medio.
-Gasto por visita muy reducido (14,1€).

-Es el segmento de mayor edad destacando el grupo de mayores de 60 años.
-Segmento con mayor preferencia por productos de **marca blanca**.

SEGMENTO 3 (44.80%)

-**Compran por igual productos baratos, caros y productos de precio medio.**

-Gasto mensual alto. Los que más gastan junto con el grupo 4.
-Número de visitas medio-alto.
-Nivel de redención alto

-El grupo con mayor porcentaje de clientes que viven con menores de edad.

SEGMENTO 1 (18.34%)

-**Preferencia de productos de gama media.**

-Gasto medio-alto.
-Número de visitas medio
-Nivel de sensibilidad al cupón medio.

-Es el segundo segmento de mayor edad después del segmento 2. Predominio del grupo de mayores de 60 años.

SEGMENTO 4 (9.85%)

-**Preferencia por productos caros.**
-**Es el segmento menos sensible al cupón.**

-Los que más gastan junto con el segmento 3.
-Son los que más gastan por visita.
-Gasto mensual alto.
-Número de visitas medio.

- Es el grupo que menos productos de marca blanca adquiere.
- Predominio del grupo de 45-60 años.

116

Typology of the customers