Procesamiento multivariante de datos en R

Universidad Internacional de Valencia

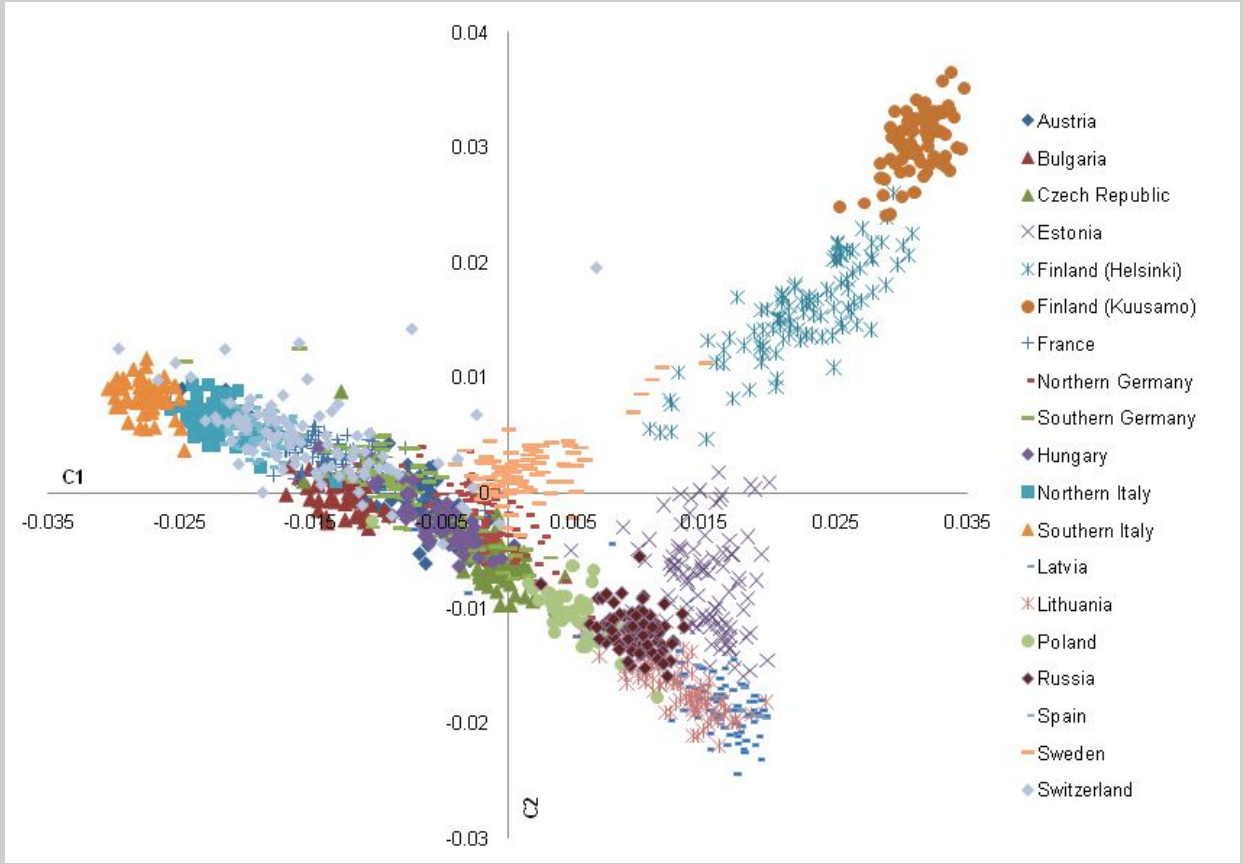
N. Sofía Huerta-Pacheco

19/02/2024

De:
Planeta Formación y Universidades

Agenda

1. Introducción
2. Métodos de interdependencia
    a. Análisis de Componentes Principales
    b. Análisis de Correspondencia
3. Métodos de dependencia
    a. Análisis de Correlación Canónica



¿Qué es?

Es el **conjunto de métodos estadísticos** cuya finalidad es **analizar simultáneamente conjuntos de datos**, esto quiere decir que hay varias variables medidas para cada individuo u objeto estudiado.

Que tiene como objetivos:

1. Proporcionar métodos cuya finalidad es el **estudio conjunto de datos multivariantes** que el **análisis estadístico uni y bidimensional es incapaz** de conseguir.
2. **Ayudar al analista o investigador** a tomar **decisiones óptimas en el contexto** en el que se encuentre, teniendo en cuenta la información disponible por el conjunto de datos analizados.

Tipos de variables - Escala de medición

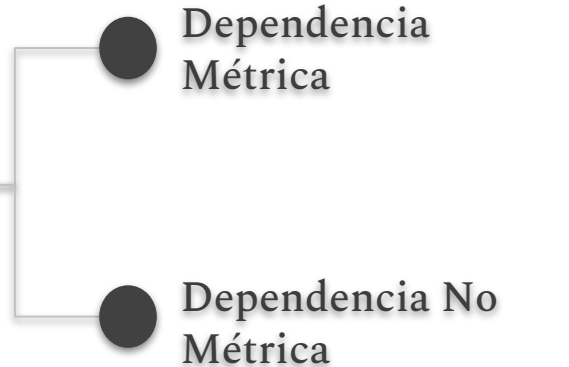
¿Con que trabaja?

Matriz de datos
=
Conjunto de datos ordenados de individuos y variables

Individuos (n) y variables (m) de una muestra que pueden ser representativos o no de una población.

Variables	1	2	...	...	m
Individuos o Caso	Sexo	Edad (Años)	Lugar de nac.	Dosis (mg /6h)	Temperatura (ºC)
1	F	18	Yucatán	600	38.4
2	M	22	Puebla	1300	40.3
...	...	...	...	...	...
...	...	...	...	...	...
n	F	19	Colima	800	36.1

Tipo de técnicas y/o métodos



- Análisis de regresión
- Análisis de Supervivencia
- MANOVA
- Correlación Canónica

- Análisis Discriminante
- Regresión Logística
- Análisis Conjoint

- A. Componentes Principales
- Análisis Factorial
- Escalas Multidimensionales
- Análisis Cluster

- Análisis de Correspondencia
- Modelos log-lineales
- Escalas Multidimensionales
- Análisis Cluster

Sosa Rojano, M (2006). Estudio experimental para la clasificación del grado de una quemadura

Universidad Internacional de Valencia

¿Cómo se interpretan?

- Conforme el tipo de información a trabajar y del análisis aplicado se pueden dar una **interpretación general del comportamiento** tanto de los individuos y de las variables, particularizando si se observan **patrones**, **tendencias**, **relaciones**, entre otros.

¿Cuando se aplican?

- Mientras se cumplan los criterios básicos, donde;

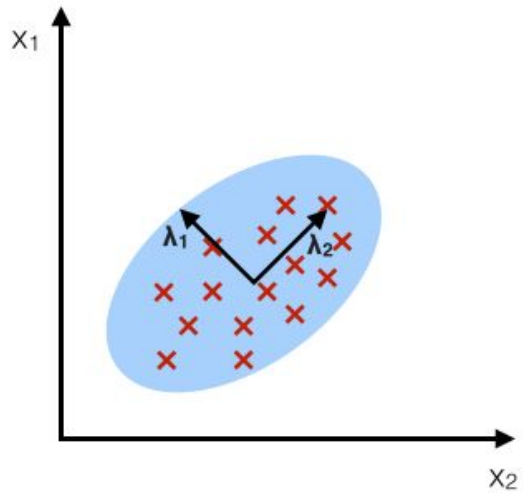
  #Individuos (n) presente variabilidad y sean > 20

  # Individuos (n) > # Variables (m) y las # Variables (m) ≥ 2

- Se requiera:
  - Reducir dimensionalidad
  - Agrupar o dividir de segmentos
  - Asociar individuos o variables o ambos
  - Identificar diferencias

Métodos de interdependencia

PCA:
component axes that maximize the variance



Métodos con el objetivo de identificar qué variables están relacionadas, cómo lo están y por qué, sin importar si son dependientes o independientes.

"Relaciones entre las variables que no pueden ser consideradas ni dependientes ni independientes desde un punto de vista teórico"

Meneses, J. (2019). Introducción al análisis multivariante.

Análisis de Componentes Principales (ACP o PCA)

Se utiliza para **analizar interrelaciones** entre un **número elevado de variables métricas explicando dichas interrelaciones** en términos de un **número menor de variables** denominadas **factores** (si son inobservables) o **componentes principales** (si son observables)

El fin esencial de la técnica es proveer al usuario **resúmenes bidimensionales de la matriz de datos**, con una **pérdida mínima** de información.

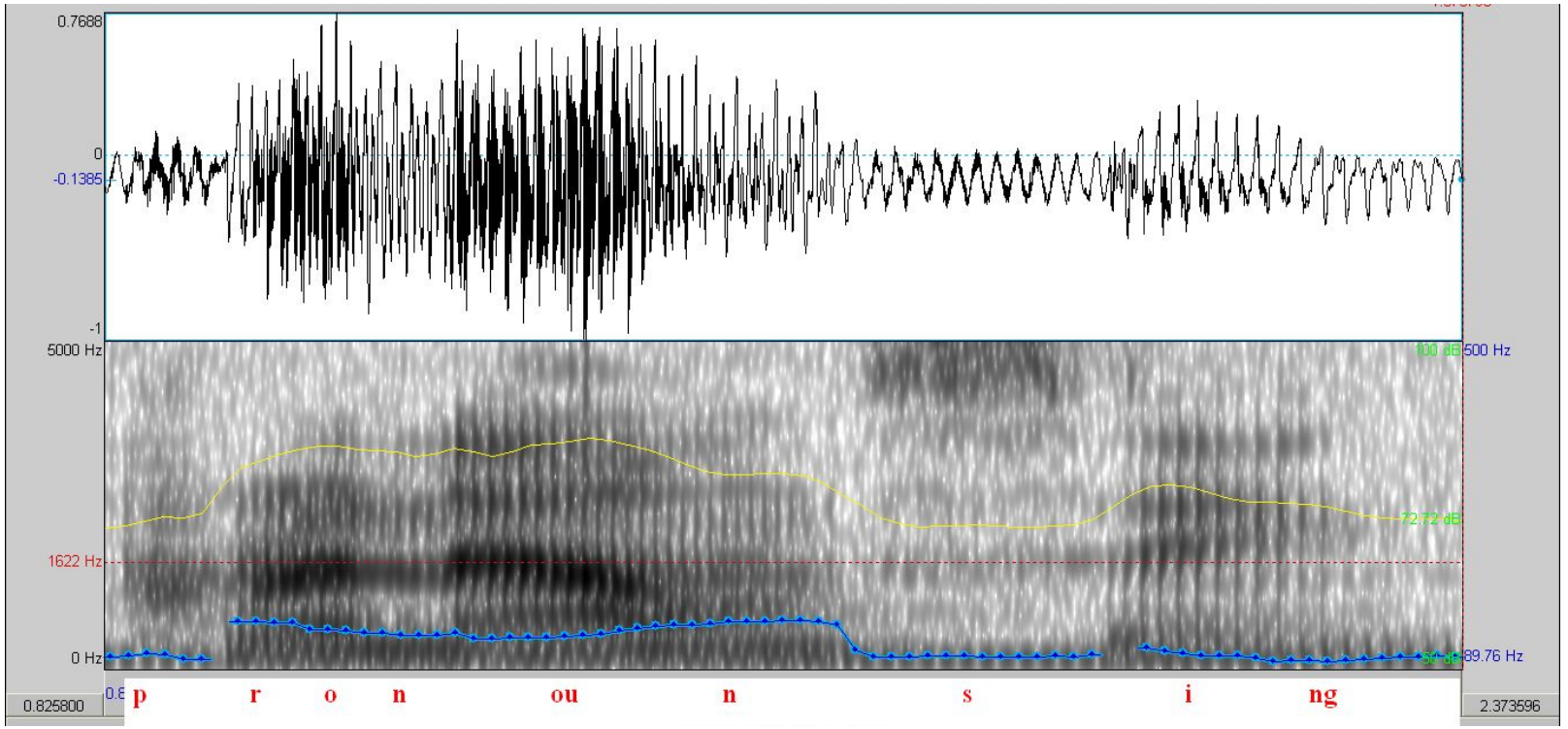
Análisis de Componentes Principales (ACP o PCA)

Por medio de p**rocedimientos matemáticos de optimización** tendientes a reproducir en el plano, lo mejor posible, **los individuos y las variables**, los cuales son vectores de espacios con dimensiones mayor que dos.

Es una de las herramientas más utilizadas de la estadística exploratoria multivariante

Datos a analizar en el PCA

Corpus de Lengua Oral del Español de México (CLOE)



Las variables del CLOE fueron extraídas de audios de entrevistas a personas del centro de México, cada variable presenta diferente escala de medición.

- ➔ ID usuario y etiquetas
- ➔ 5 formantes
- ➔ Duración

Identificar los parámetros acústicos que representan la variación inter-hablantes

Material de Consulta

Procedimiento

```
#Paquetes

library(FactoMineR)

library(factoextra)

#Selección de variables

ruta<-file.choose()

datos<-read.csv(ruta)

head(datos)

dim(datos)

d<-datos[,c("F0","F1","F2","F3","F4","Time")]

#Análisis de Componentes Principales

result<-PCA(X = d, graph = FALSE)

fviz_screeplot(result, addlabels = TRUE,
ylim = c(0, 100))

fviz_contrib(result, choice = "var", axes
= 1, top = 10)

fviz_pca_ind(result, geom.ind="point",
col.ind="gray", axes=c(1, 2),
pointsize=0.5)

fviz_pca_var(result, col.var = "cos2",
geom.var = c("arrow","text"), labelsize=2,
repel = FALSE,label = "all")

fviz_pca_biplot(result,label="var",
col.var ="cos2", alpha.ind="contrib",
pointsize = 0.5)
```

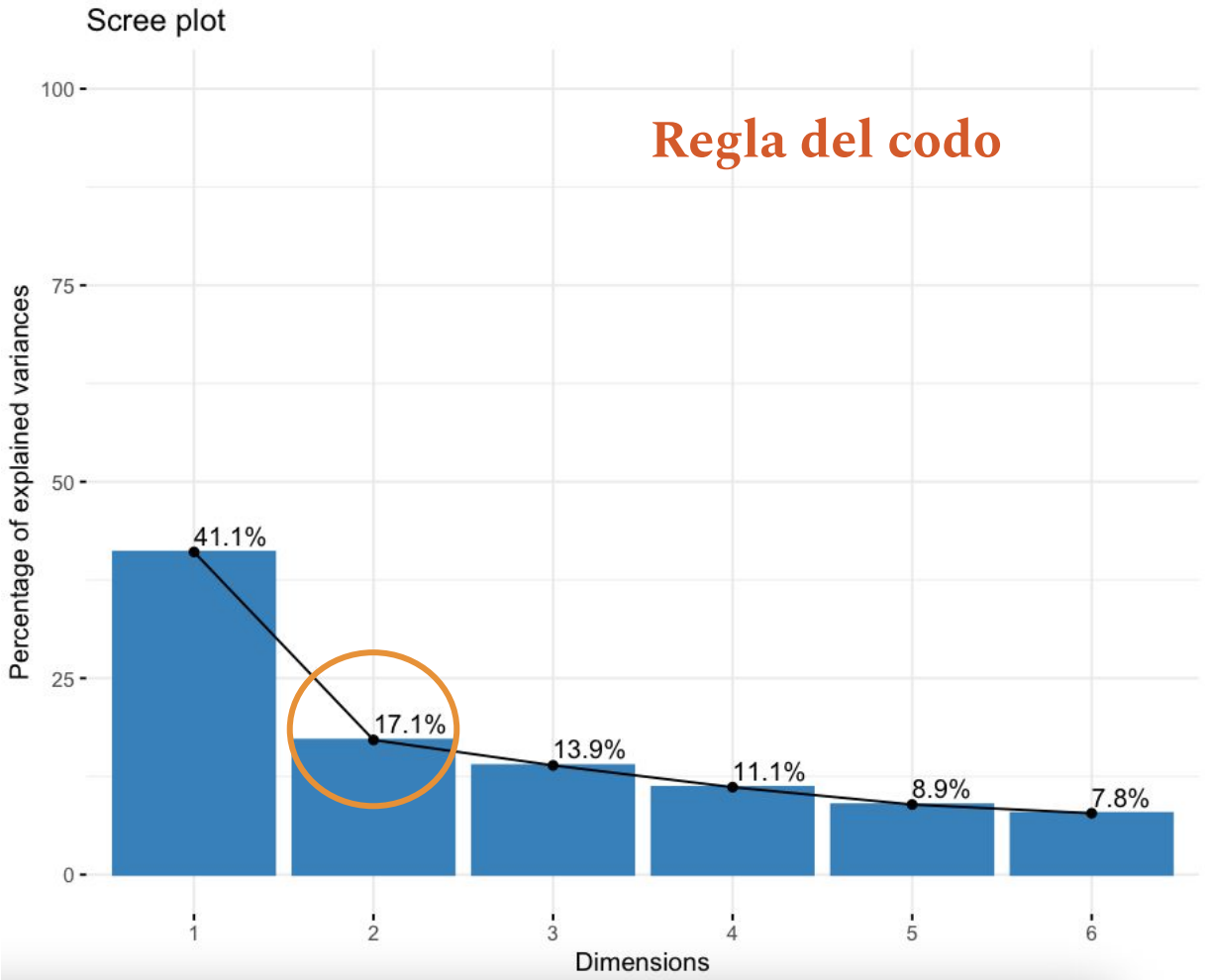
n = 7526 ; m = 6

Visualización e interpretación

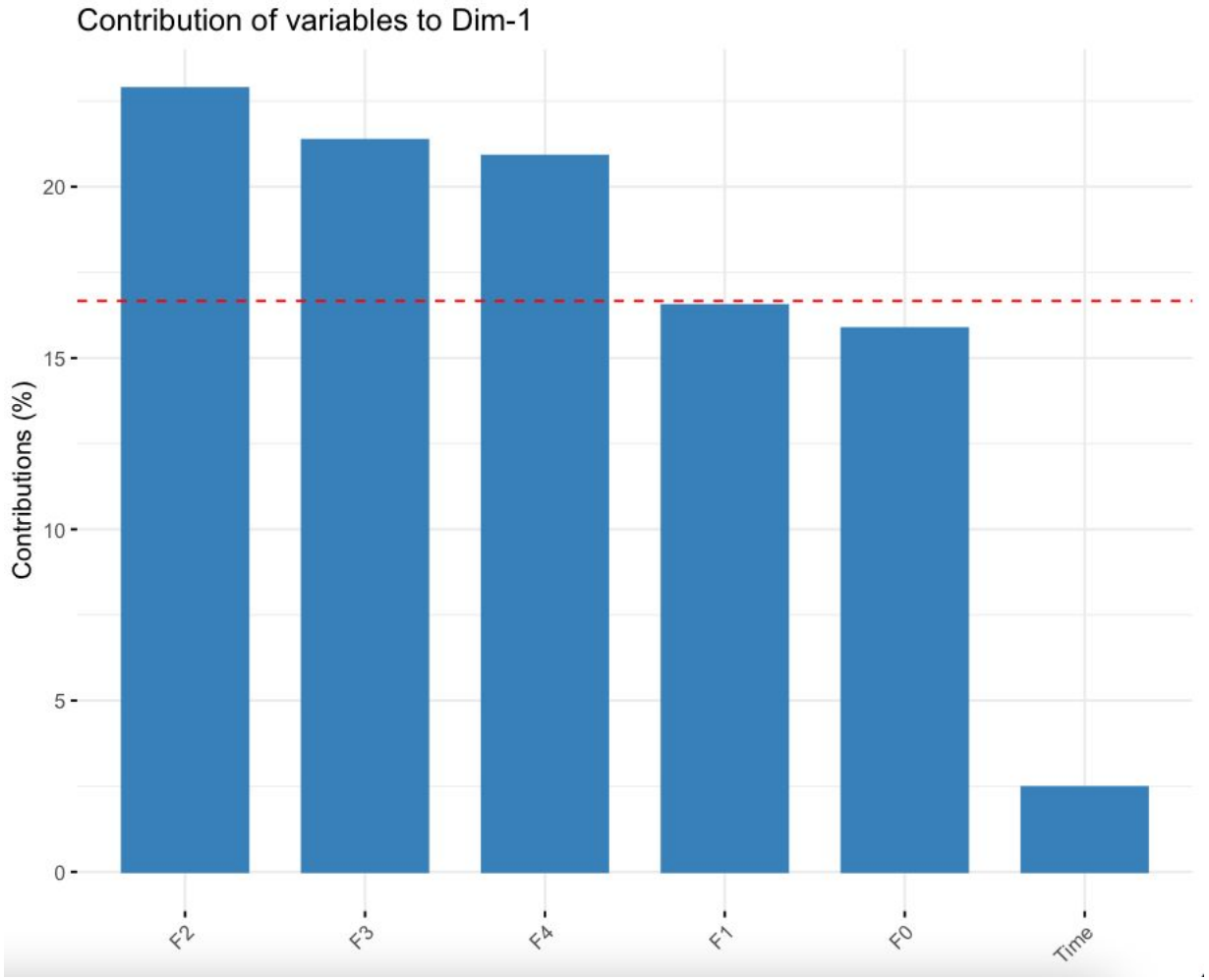
```
> head(datos)
                 ID    Labels       F0       F1       F2       F3       F4     Time
1 CDMN3E1SM007_LEC     CLa_T 152.7063 573.4208 1428.229 2353.437 3353.938 0.05603
2 CDMN3E1SM007_LEC   COSDa_A 147.6651 549.0151 1404.217 2210.491 3547.161 0.04310
3 CDMN3E1SM007_LEC    COBa_T 136.9980 564.0187 1323.520 2611.424 3519.893 0.06034
4 CDMN3E1SM007_LEC    CFAa_A 143.3070 595.4744 1254.156 2286.769 3362.817 0.04379
5 CDMN3E1SM007_LEC CLaCFA_T 128.0363 548.2760 1353.523 2791.432 3742.227 0.06034
6 CDMN3E1SM007_LEC    CFLa_A 137.5528 547.5389 1247.473 2236.449 3608.978 0.04741
```

```
> result$eig
       eigenvalue percentage of variance cumulative percentage of variance
comp 1  2.4639076              41.065127                           41.06513
comp 2  1.0284463              17.140771                           58.20590
comp 3  0.8344214              13.907024                           72.11292
comp 4  0.6680664              11.134440                           83.24736
comp 5  0.5357850               8.929750                           92.17711
comp 6  0.4693733               7.822888                          100.00000
```

Visualización e interpretación



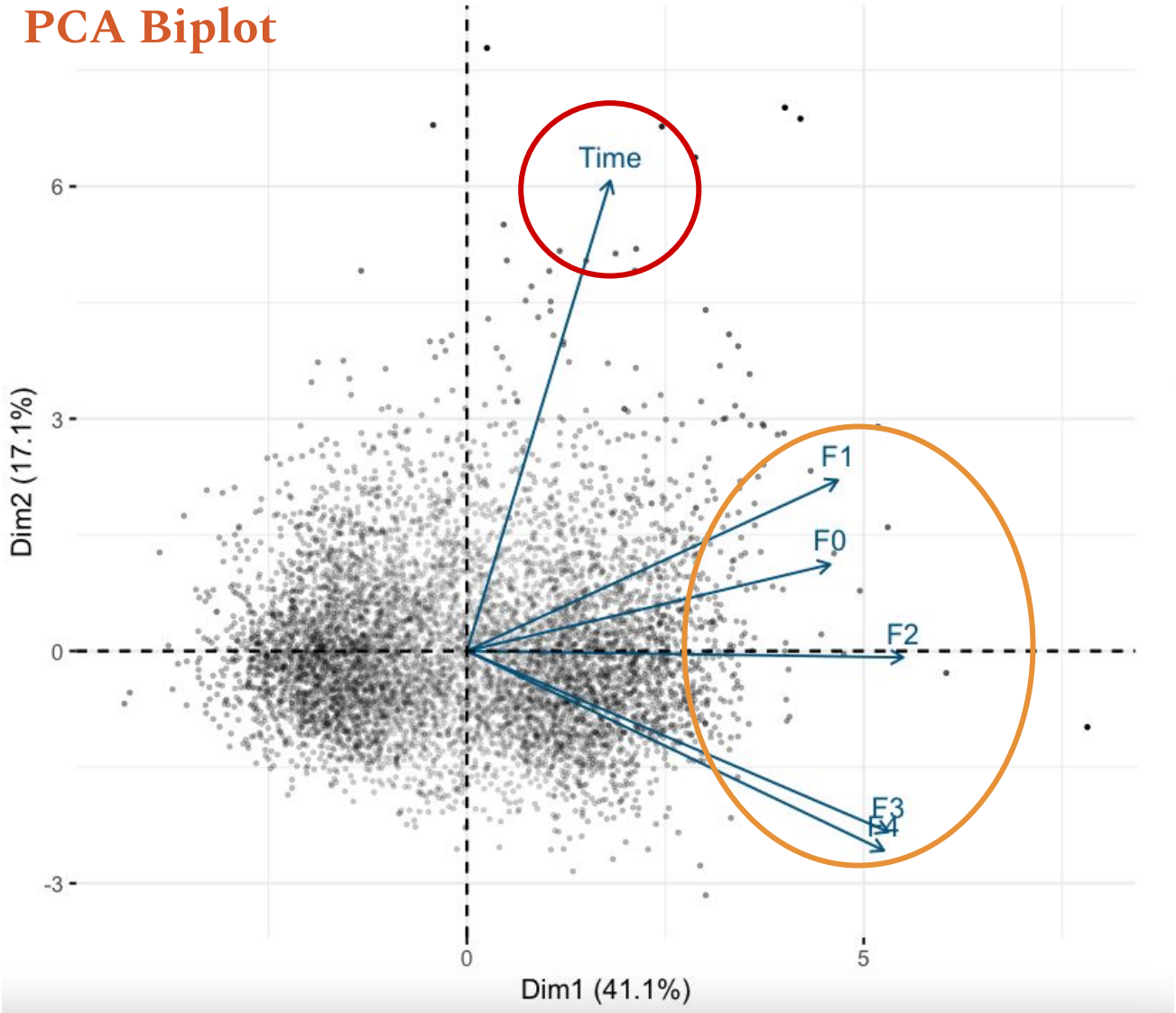
Regla del codo



Proporción de Variabilidad Acumulada > al 70 % (0.7)

Visualización e interpretación

PCA Biplot



Consideraciones

Los vectores (->) mientras más largos, mejor representado están en esos componentes.

Si el ángulo entre pares de vectores es pequeño (<90º), se puede decir que existe una asociación positiva.

Si el ángulo entre pares de vectores es recto (aproximado o = 90º), se puede decir que no existe relación.

Si el ángulo entre pares de vectores es grande (>90º) si puede decir que existe asociación negativa.

Análisis de Correspondencia (AC o CA)

Se aplica a **tablas de contingencia multidimensionales** y persigue un objetivo similar al de las escalas multidimensionales, el cual es **representar simultáneamente las filas y columnas de las tablas de contingencia.**

Este puede ser **simple** (representación de **dos variables no métricas**) o **múltiple** que considera **más de dos variables no métricas** (categóricas).

Datos a analizar en el CA

Desaparecidos de México 2006 al 2018
Centro Nacional de Planeación, Análisis e Información para el Combate a la Delincuencia, PGR

Variables recolectadas en los registros de personas desaparecidas en México durante los años 2006 al 2018.

- ➔ Información sociodemográfica
- ➔ Temporalidad de registros
- ➔ Características de media filiación
- ➔ Estatus de localización

Identificar si existe dependencia del estatus de localización con respecto al lugar de desaparición de las personas reportadas.

Material de Consulta

Procedimiento

```
#Paquetes

library(readxl)

library(FactoMineR)

library(factoextra)

#Selección de variables

ruta<-file.choose()

datos<-read_excel(ruta)

datos<-as.data.frame(datos)

head(datos)

dim(datos)

names(datos)

d1<-datos[,c("ESTADO","VIVO O MUERTO")]

#Prueba de Chi-cuadrada

td1<-table(d1)

chisq <- chisq.test(td1)

#Análisis de Correspondencia Simple

result<-CA(td1, graph = FALSE)

fviz_screeplot(result, addlabels = TRUE,
ylim = c(0, 100))

fviz_contrib(result, choice = "row", axes
= 1:2, top = 15)

fviz_ca_biplot(result, repel = TRUE, arrow
= c(FALSE, TRUE),pointsize = 0.5,labelsize
= 2,col.col = "#16A085",col.row = "black")
```

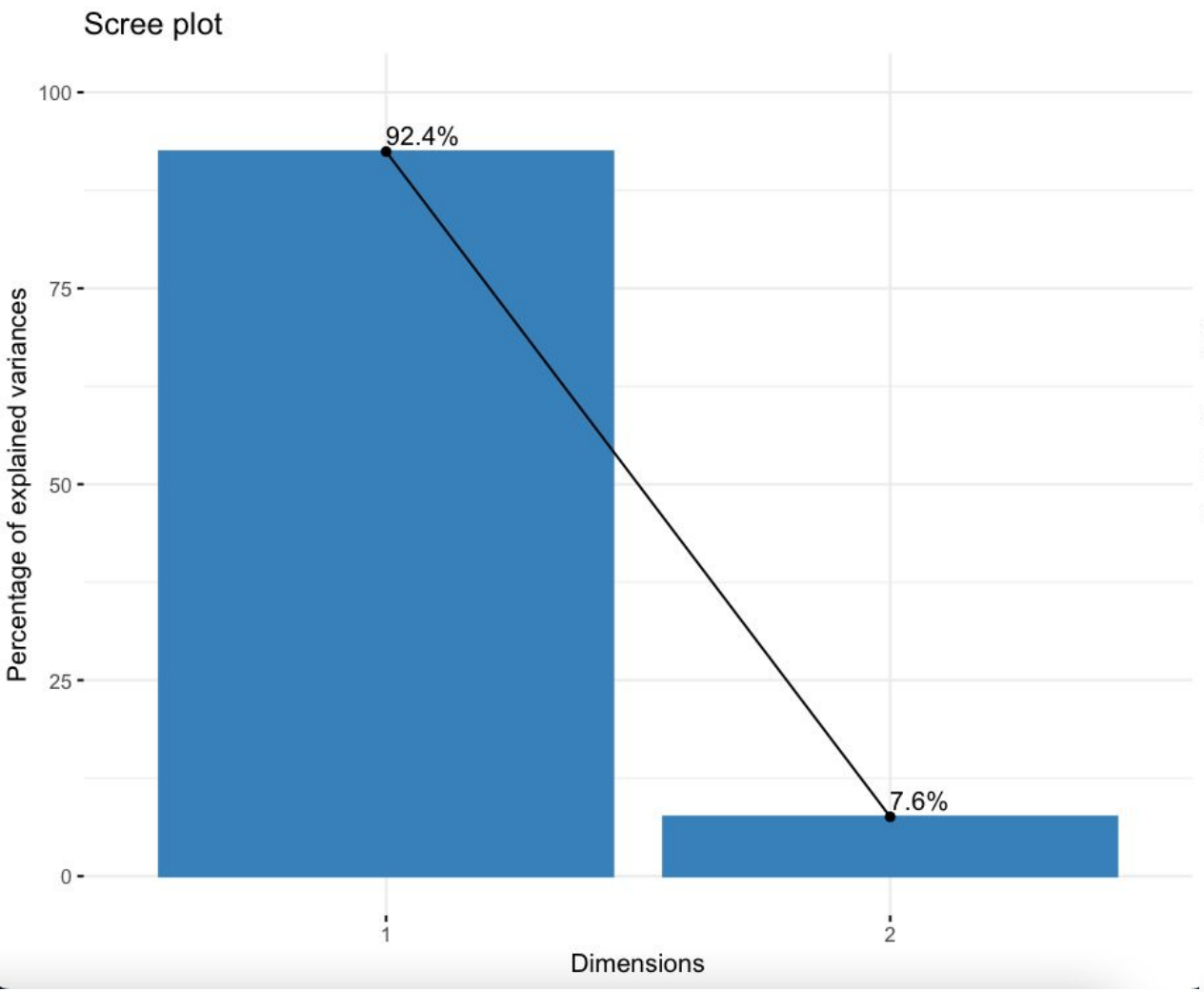
n = 99129; m = 2

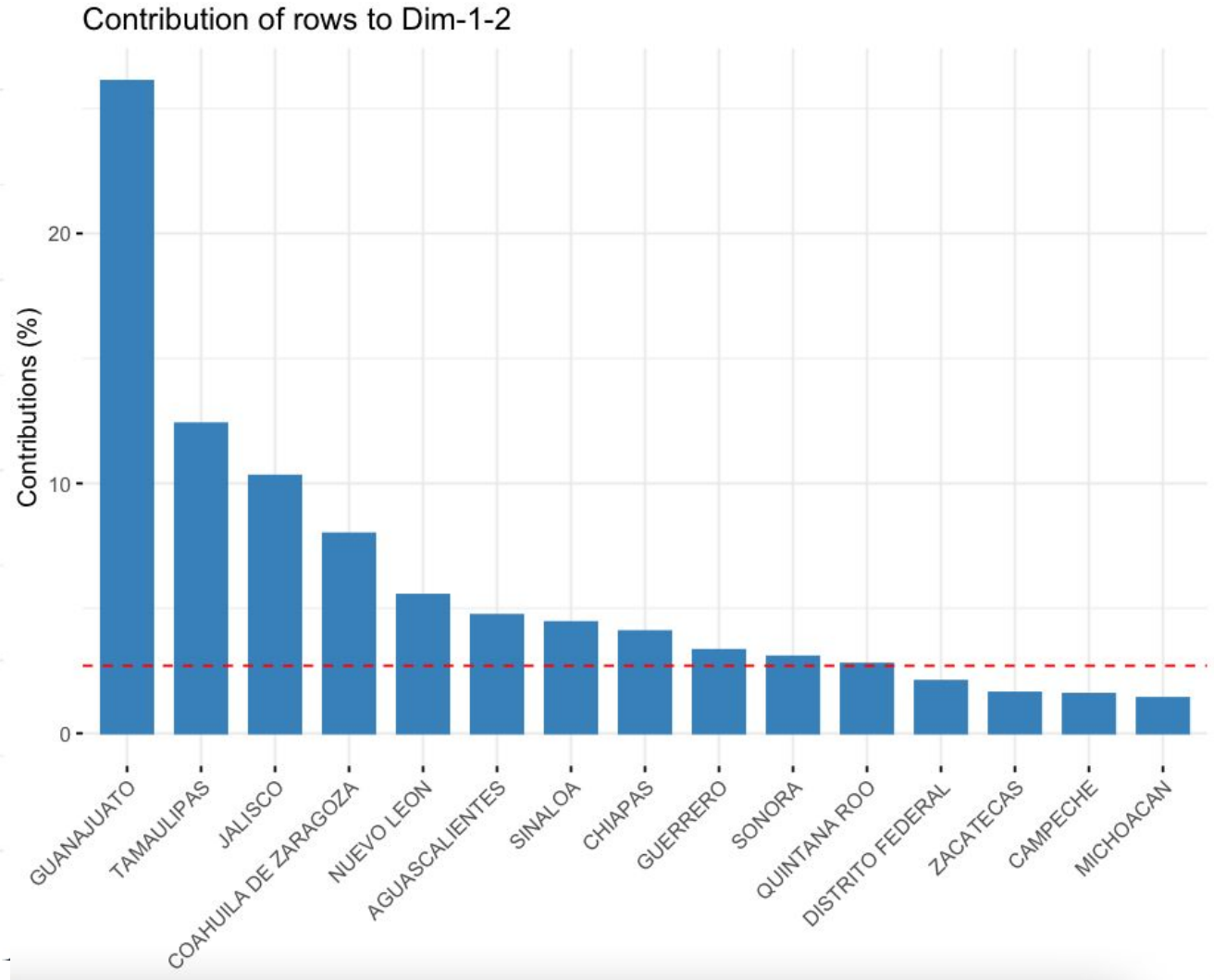
Visualización e interpretación


```
#Prueba de Chi-cuadrada

td1<-table(d1)

chisq <- chisq.test(td1)


Pearson's Chi-squared test

data:  td1

X-squared = 25953, df = 72,
p-value < 2.2e-16
```

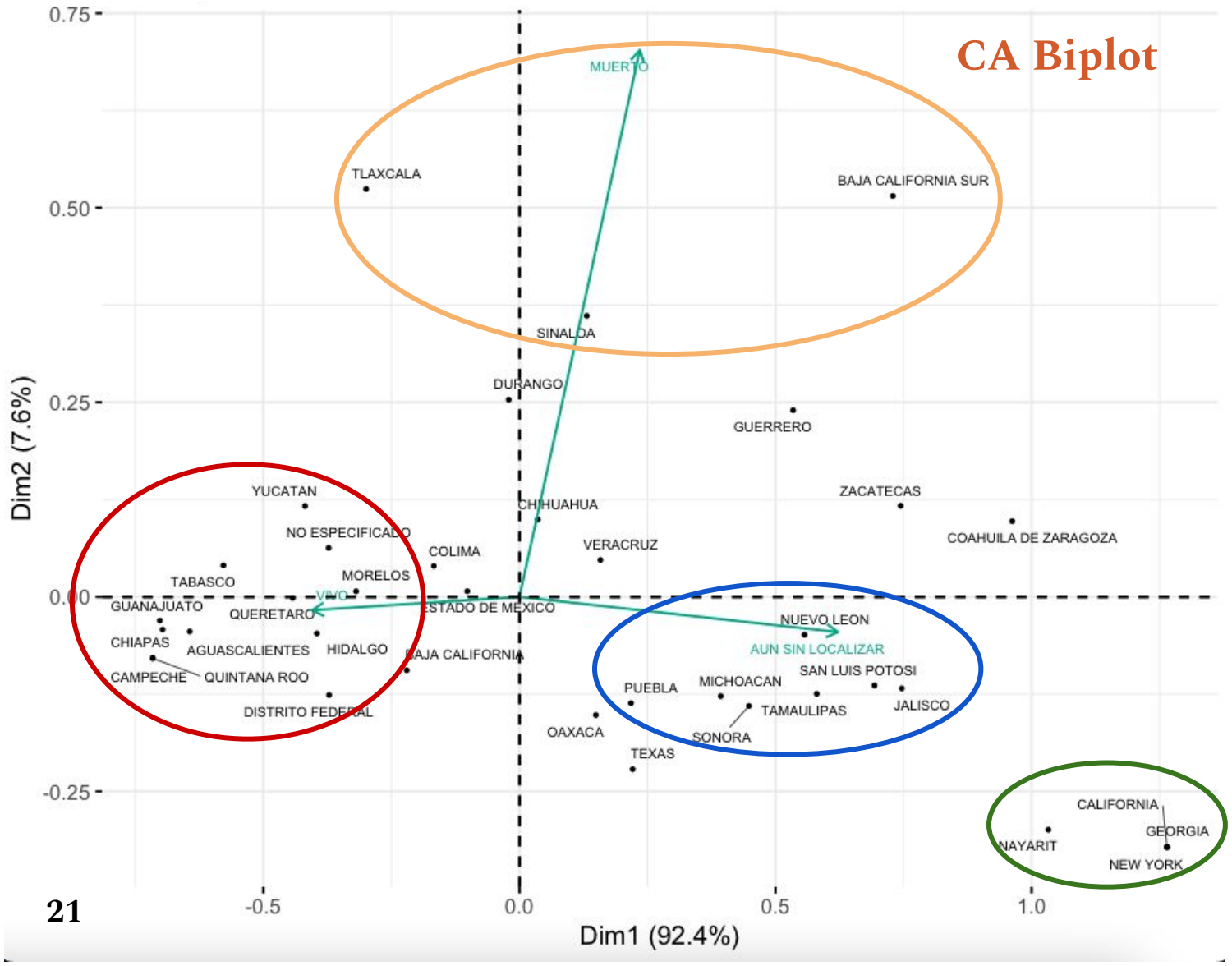
Prueba de Dependencia

Visualización e interpretación





Visualización e interpretación



Consideraciones

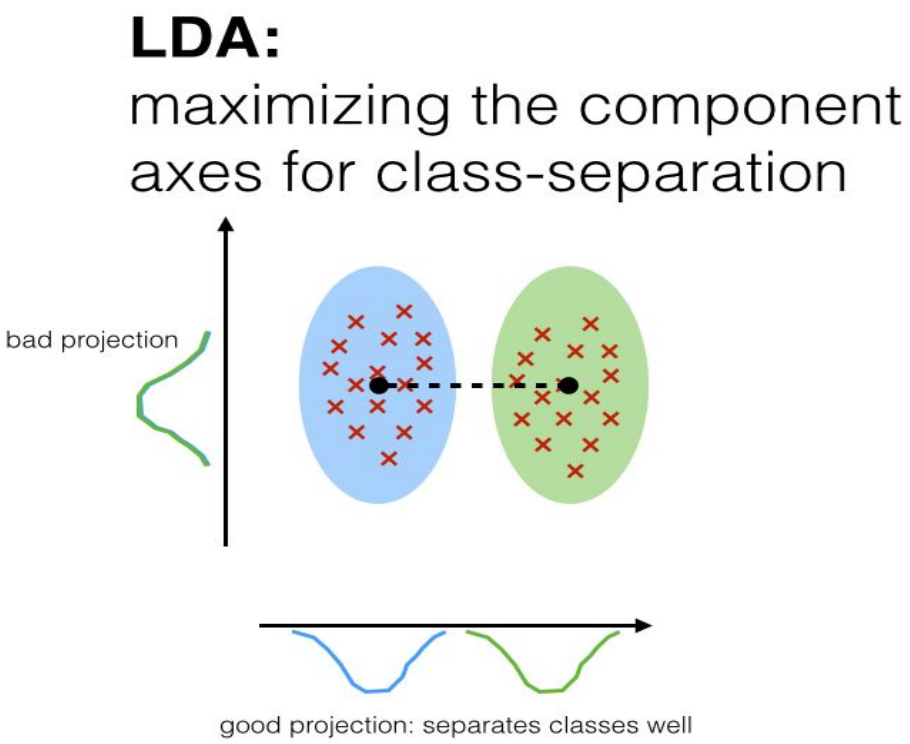
Las dimensiones (dos) representan la totalidad de variabilidad de la muestra.

Los individuos más cercanos a los vectores (->) muestran mayor asociación con ellos.

Si los vectores o individuos están muy cercanos o en el centro (0,0) no se encuentran bien representados.

Existen individuos con comportamientos distantes o no asociados a los vectores (->) evaluados.

Métodos de dependencia



Métodos que suponen que las variables analizadas están divididas en dos grupos: dependientes e independientes; con el objetivo de determinar si el conjunto de variables independientes afectan al conjunto de dependientes, además de ver la forma de cómo se afectan.

"Sirve para explicar o predecir a las variables dependientes a partir de las independientes"

Meneses, J. (2019). Introducción al análisis multivariante.

Análisis de la Correlación Canónica (ACC o CCA)

Tiene como objetivo **relacionar simultáneamente varias variables métricas dependientes e independientes** calculando **combinaciones lineales** de cada conjunto de variables que **maximicen la correlación existente** entre los dos conjuntos de variables.

Datos a analizar en el CCA

Desaparecidos de México 2006 al 2018
Centro Nacional de Planeación, Análisis e Información para el Combate a la Delincuencia, PGR

Variables recolectadas en los registros de personas desaparecidas en México durante los años 2006 al 2018.

- ➔ Información sociodemográfica
- ➔ Temporalidad de registros
- ➔ Características de media filiación
- ➔ Estatus de localización

Identificar la dependencia entre número de personas desaparecidas localizadas vivas y no localizadas por estado y temporalidad.

Material de Consulta

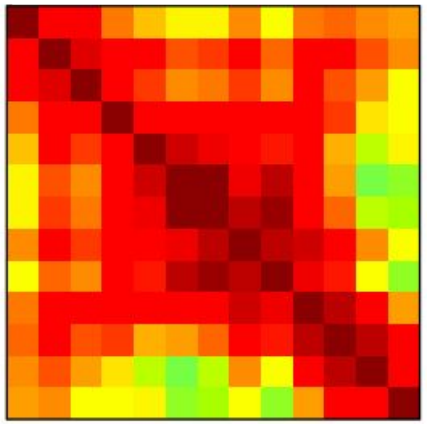
Procedimiento

```
#Paquetes
library("CCA")
#Selección de variables
ruta1<-file.choose()
datos1<-read.csv(ruta1)
ruta2<-file.choose()
datos2<-read.csv(ruta2)
head(datos1)
head(datos2)
dim(datos1)
ind <- match(datos1[,1], datos2[,1])
data<-data.frame(datos1[ind,],datos2[,2:14])
X <- data[,2:14]
Y <- data[,15:27]
#Análisis de Correlación Canónica
result <- matcor(X,Y)
img.matcor(result, type = 2)
cc <- cc(X,Y)
plt.cc(cc,ind.names = data[,1])
plt.cc(cc,d1=1,d2=2,type="v",var.label=TRUE)
```

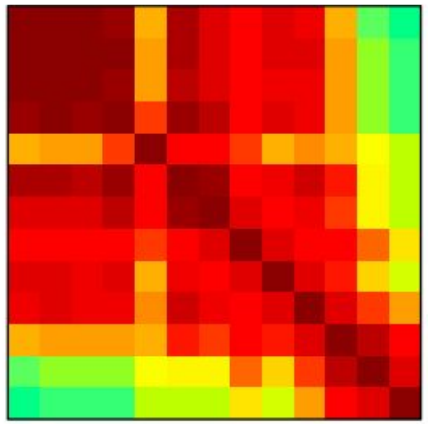
n = 32; m = 14

Visualización e interpretación

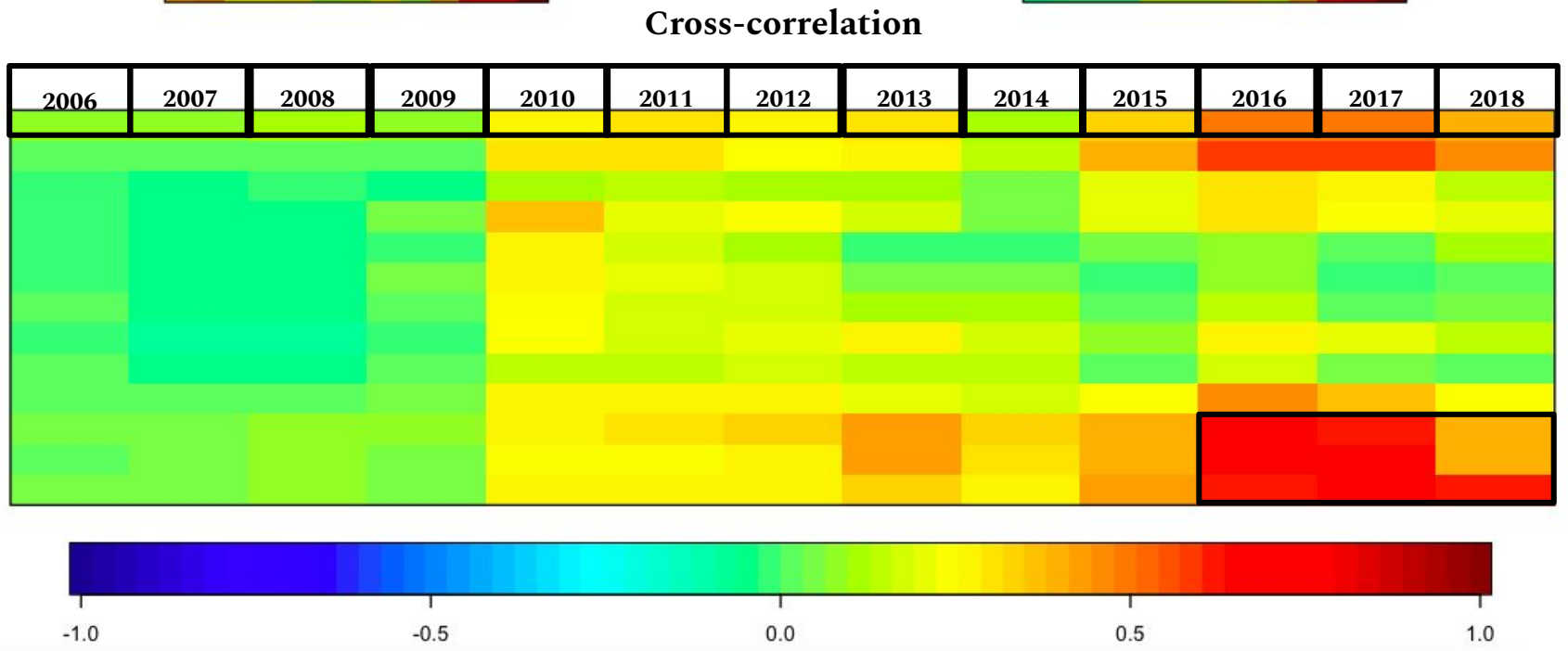
Matriz de conteos de personas localizadas con vida



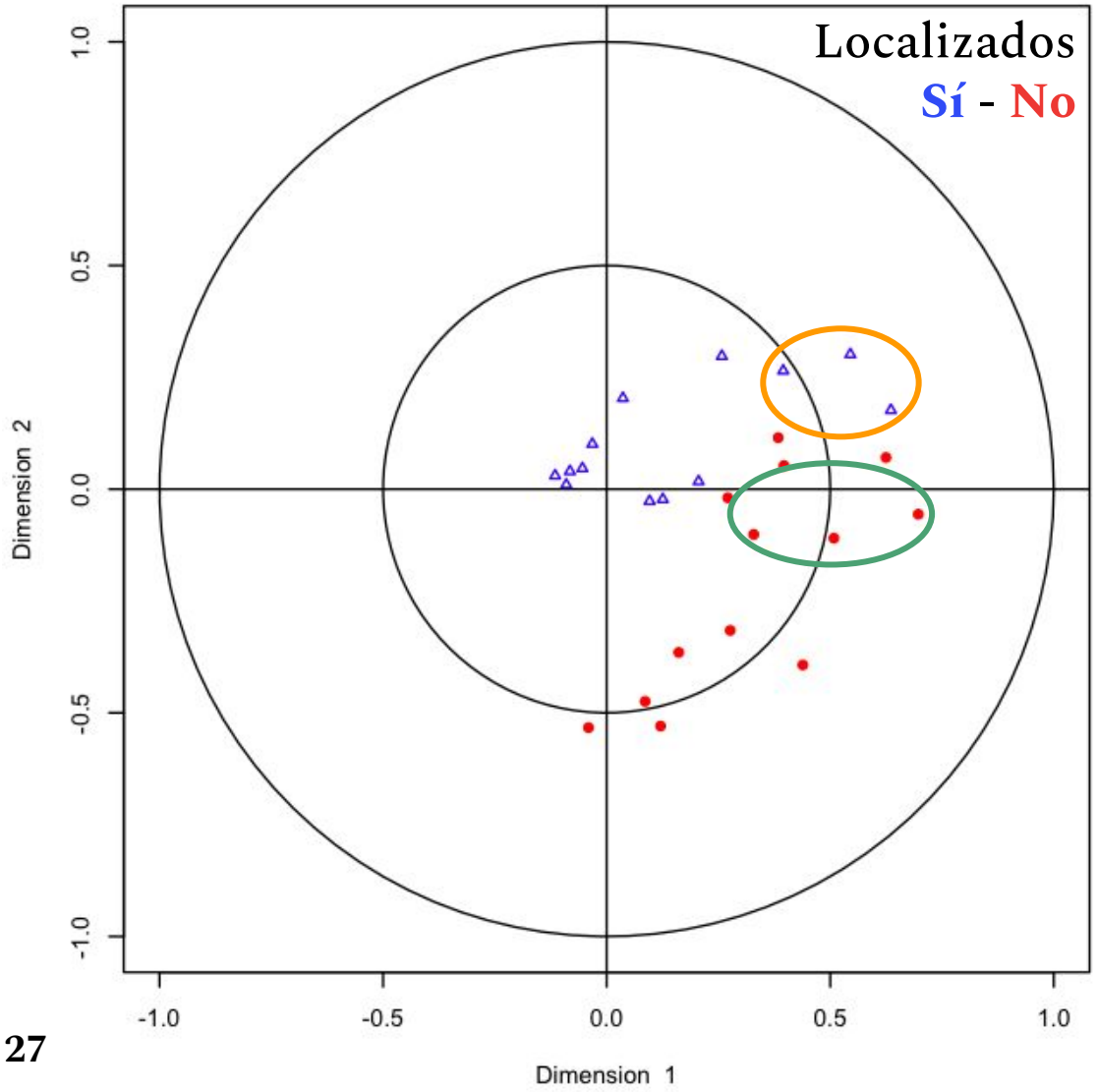
n = Estados
m = 2006,..., 2018

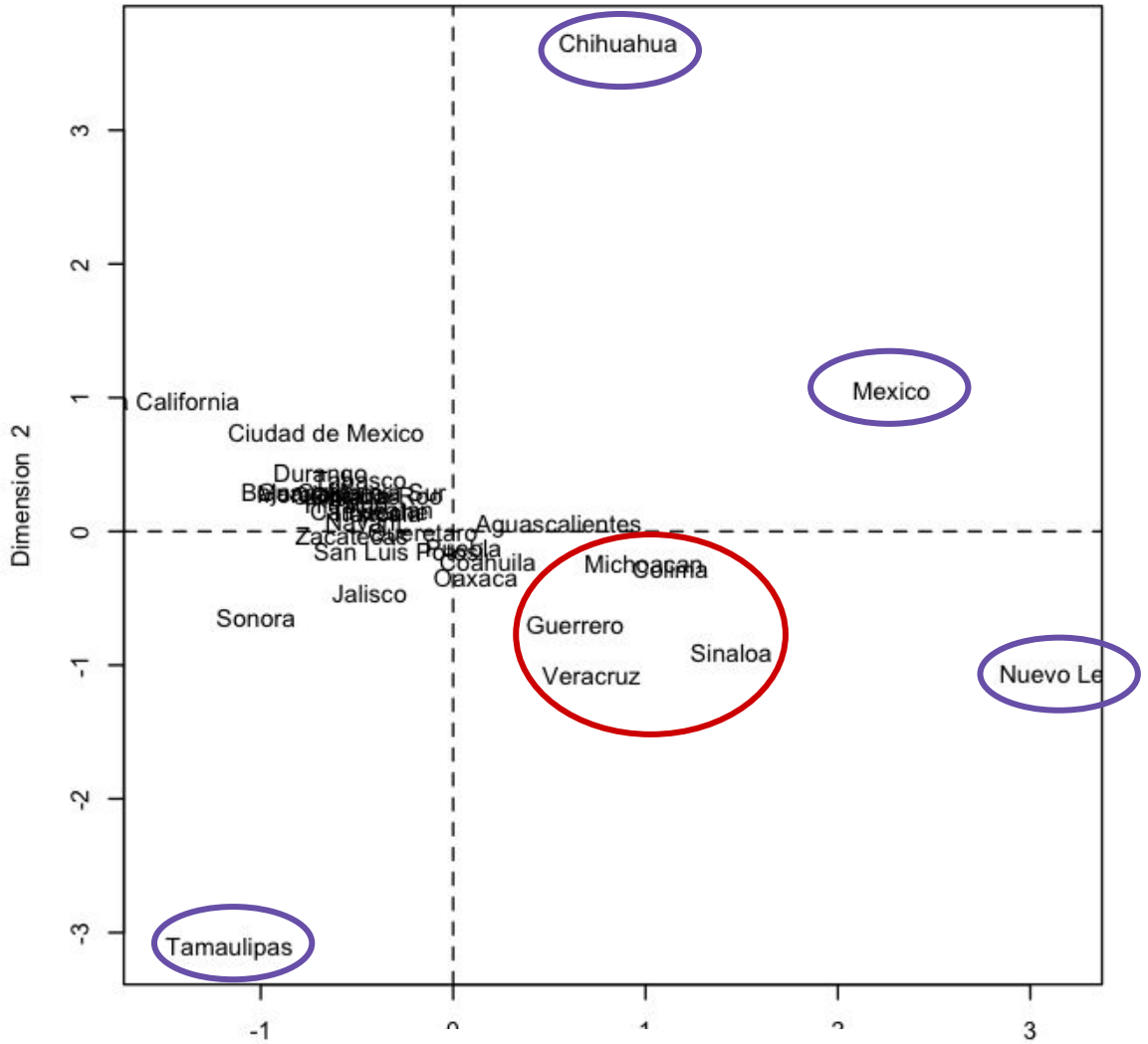


Matriz de conteos de personas no localizadas



Visualización e interpretación





Gracias su atención

Universidad Internacional de Valencia

¿Preguntas? o ¡Comentario!

N. Sofía Huerta-Pacheco
nshuerta@enacif.unam.mx

19/02/2024

De:
Planeta Formación y Universidades